

2014

A Study of Non-Linguistic Utterances for Social Human-Robot Interaction

Read, Robin

<http://hdl.handle.net/10026.1/3028>

<http://dx.doi.org/10.24382/1622>

Plymouth University

All content in PEARL is protected by copyright law. Author manuscripts are made available in accordance with publisher policies. Please cite only the published version using the details provided on the item record or document. In the absence of an open licence (e.g. Creative Commons), permissions for further reuse of content should be sought from the publisher or author.

This copy of the thesis has been supplied on condition that anyone who consults it is understood to recognise that its copyright rests with its author and that no quotation from the thesis and no information derived from it may be published without the author's prior consent.

**A STUDY OF NON-LINGUISTIC UTTERANCES FOR SOCIAL
HUMAN-ROBOT INTERACTION**

by

ROBIN READ

A thesis submitted to Plymouth University
in partial fulfilment for the degree of

DOCTOR OF PHILOSOPHY

School of Computing and Mathematics
Faculty of Science and Environment

17th April 2014

A Study of Non-Linguistic Utterances for Social Human-Robot Interaction

by
Robin Read

Abstract

The world of animation has painted an inspiring image of what the robots of the future could be. Taking the robots R2D2 and C3PO from the Star Wars films as representative examples, these robots are portrayed as being more than just machines, rather, they are presented as intelligent and capable social peers, exhibiting many of the traits that people have also. These robots have the ability to interact with people, understand us, and even relate to us in very personal ways through a wide repertoire of social cues.

As robotic technologies continue to make their way into society at large, there is a growing trend toward making *social* robots. The field of Human-Robot Interaction concerns itself with studying, developing and realising these socially capable machines, equipping them with a very rich variety of capabilities that allow them to interact with people in natural and intuitive ways, ranging from the use of natural language, body language and facial gestures, to more unique ways such as expression through colours and abstract sounds.

This thesis studies the use of abstract, expressive sounds, like those used iconically by the robot R2D2. These are termed *Non-Linguistic Utterances* (NLUs) and are a means of communication which has a rich history in film and animation. However, very little is understood about how such expressive sounds may be utilised by social robots, and how people respond to these.

This work presents a series of experiments aimed at understanding how NLUs can be utilised by a social robot in order to convey affective meaning to people both young and old, and what factors impact on the production and perception of NLUs. Firstly, it is shown that not all robots should use NLUs. The morphology of the robot matters. People perceive NLUs differently across different robots, and not always in a desired manner. Next it is shown that people readily project affective meaning onto NLUs though not in a coherent manner. Furthermore, people's affective inferences are not subtle, rather they are drawn to well established, basic affect prototypes. Moreover, it is shown that the valence of the situation in which an NLU is made, overrides the initial valence of the NLU itself: situational context biases how people perceive utterances made by a robot, and through this, coherence between people in their affective inferences is found to increase. Finally, it is uncovered that NLUs are best not used as a *replacement* to natural language (as they are by R2D2), rather, people show a preference for them being used *alongside* natural language where they can play a supportive role by providing essential social cues.

Contents

Abstract	i
Table of Contents	iii
List of Figures	ix
List of Tables	xv
Acknowledgements	xxiii
Author’s declaration	xxv
1 Introduction	1
1.1 Human-Robot Interaction	3
1.1.1 Emotion and Affective Robots	4
1.2 Non-Linguistic Utterances	5
1.3 Vocal Communication and Expression	7
1.3.1 The Human Voice	8
1.3.2 Speech Synthesis	8
1.4 The thesis	9
1.5 Contributions	11
1.6 Structure	12
2 Non-Linguistic Utterances	17
2.1 A working definition of Non-Linguistic Utterances	18
2.1.1 Examples of NLUs in Popular Culture	19
2.1.2 Examples of NLUs in Real Robots	20
2.2 Motivations and Applications in HRI	21
2.2.1 Utility as a tool in broader HRI research	24
2.3 Affective Expression through Sound	26
2.3.1 Affective Expression via the Human Voice	26
2.3.2 Affective Expression via Music	33
2.3.3 Sound Symbolism	34
2.4 Communicating Affect through NLUs and gibberish speech: re- viewing previous work	35
2.4.1 NLUs	36
2.4.2 Gibberish speech	39
2.4.3 Discussion	43
2.4.4 Review Summary	49
2.5 Are NLUs a language?	51
2.6 Summary	53

3	Methods	55
3.1	Creating NLUs	56
3.1.1	NLU Anatomy and Parameterisation	56
3.1.2	Utterance Generation	63
3.1.3	Remarks on the design of NLUs	63
3.2	The Nao robot	64
3.2.1	Programming Nao	66
3.2.2	How Nao has been used	67
3.3	Measuring Affect	67
3.3.1	Representations of Affect	68
3.3.2	Capturing Affect from People	71
3.3.3	The AffectButton	74
3.4	Summary	79
4	Alignment of NLUs with Agent Morphology	81
4.1	Experiment Setup	83
4.1.1	Utterance Stimuli	84
4.1.2	Visual Stimuli	85
4.1.3	Emotional States	87
4.1.4	Communicative Intents	87
4.2	Results	88
4.2.1	Affective Ratings	88
4.2.2	Intentional Ratings	91
4.2.3	Appropriateness Ratings	91
4.2.4	Summary of Results	97
4.3	Discussion	98
4.3.1	Methodological Remarks	99
4.4	Summary	100
5	Collecting Training Data for Machine Learning	103
5.1	Identifying an Affective Mapping	105
5.2	Experimental Setup	107
5.2.1	Utterance Specification	108
5.3	Results	114
5.3.1	Experiment #1	114
5.3.2	Experiment #2	115
5.3.3	Experiment #3	117
5.3.4	Experiment #4	119
5.3.5	Experiment #5	123
5.3.6	Summary of Results.	125
5.4	Discussion	127
5.4.1	Methodological Remarks	130
5.5	Summary	131
6	Categorical Perception of NLUs	133
6.1	Categorical Perception	134
6.2	Experimental Setup	136
6.2.1	Utterance Stimuli	138
6.2.2	Labelling Task	140
6.2.3	Discrimination Task	141
6.2.4	Identification Task	143

6.2.5	Experimental Procedure	144
6.3	Results	146
6.3.1	Labelling Task Results	146
6.3.2	Discrimination Task Results	151
6.3.3	Identification Task Results	158
6.3.4	Summary of Results	162
6.4	Discussion	165
6.4.1	Results Discussion	165
6.4.2	Methodological Remarks	167
6.4.3	Broader Discussion	170
6.5	Summary	171
7	Using Artificial Neural Networks to Automate NLU Production and Affective Charging	173
7.1	Overview of NLU/gibberish speech Generation.	175
7.2	Machine Learning	179
7.2.1	Approaches to Machine Learning	180
7.2.2	Supervised Learning	181
7.2.3	Formal ML Problem Statement	184
7.3	ANN Design and Implementation	185
7.3.1	Training Data	185
7.3.2	Using Feed Forward ANNs	188
7.3.3	Network Training	191
7.3.4	Affective Mapping Analysis	199
7.4	Subject Evaluation	205
7.4.1	Experimental Setup	205
7.4.2	Results	211
7.4.3	Summary of Results	219
7.5	Discussion	220
7.5.1	ANN Design and Implementation	220
7.5.2	Subject Evaluation	222
7.5.3	Methodological Remarks	224
7.6	Summary	229
8	The Influence of Situational Context upon the Interpretation of NLUs	231
8.1	Experimental Setup	233
8.1.1	Stimulus Production	236
8.1.2	Experimental Procedure	241
8.2	Results	242
8.2.1	NLU Videos	244
8.2.2	Action Only Videos	245
8.2.3	Action/NLU combination Videos	248
8.2.4	Summary of Results	251
8.3	Discussion	253
8.3.1	Main Effects	253
8.3.2	Interaction Effects	254
8.3.3	Methodological Remarks	255
8.3.4	Practical Use of NLUs	257
8.4	Summary	258

9	Combining NLUs with Natural Language	261
9.1	Experimental Setup	263
9.1.1	Stimulus Production	269
9.1.2	Experimental Procedure	271
9.2	Results	272
9.2.1	Appropriateness Ratings	274
9.2.2	Expressiveness Ratings	277
9.2.3	Preference Ratings	281
9.2.4	Naturalness Ratings	283
9.2.5	Like-ability Ratings	284
9.2.6	Summary of Results	286
9.3	Discussion	288
9.3.1	Methodological Remarks	291
9.3.2	Broader Remarks	293
9.4	Summary	295
10	Conclusions, contributions and future work	297
10.1	Summary	297
10.1.1	Summary of the main contributions	302
10.2	Discussion	303
10.2.1	The custom method for creating NLUs	304
10.2.2	Using a single robotic platform	305
10.2.3	The AffectButton measuring tool	306
10.2.4	Child and Adult Evaluations	307
10.2.5	Automating the generation of NLUs, their affective meaning and the role of situational context	308
10.3	Future Work	309
10.3.1	Robot embodiment and morphology	310
10.3.2	Exploration of different types of NLU synthesisers and repli- cation of results	311
10.3.3	Using NLUs alongside Natural Language	312
10.3.4	Long-Term Human-Robot Interaction	313
A	Methods	315
B	Alignment of NLUs with Agent Morphology	335
C	Collecting Training Data for Machine Learning	349
C.1	Experiment #1 - Tables of Results	351
C.2	Experiment #2 - Tables of Results	351
C.3	Experiment #3 - Tables of Results	352
C.4	Experiment #4 - Tables of Results	354
C.5	Experiment #5 - Tables of Results	356
D	Categorical Perception of NLUs	357
E	Using Artificial Neural Networks to Automate NLU Production and Affective Charging	367
F	The Influence of Situational Context upon the Interpretation of NLUs	373

G Combining NLUs with Natural Language	381
Bibliography	391

List of Figures

3.1	Utterance with 5 concatenated sound units, each with a different pitch contour and pause ratio, and a rhythm rhythm value that is less than 1.	57
3.2	Schematic of a common Attack Decay Sustain Release (ADSR) amplitude envelope with linear transitions.	58
3.3	Example of how the Tremolo parameter applies to the frequency envelope of a carrier signal in a sound unit.	59
3.4	Example of how the Node Ratio parameter applies to the frequency envelope of a carrier signal in a sound unit.	60
3.5	Example of how the Skew Ratio parameter applies to the frequency envelope of a carrier signal in a sound unit.	60
3.6	The Aldbaran Nao robotic platform used throughout this body of research.	65
3.7	The AffectButton prototype facial expressions with PAD values. From left, clockwise: Neutral (0,-1,0), Angry (-1,1,1), Excited (1,1,1), Scared (-1,1,-1), Surprised (1,1,-1), Annoyed (-0.5,-1,0.5), Happy (0.5,-1,0.5), Sad (-0.5,-1,-0.5), Content (0.5,-1,-0.5). Adapted from Broekens et al. (2010).	74
3.8	Example of how the Arousal value is calculated from the Pleasure and Dominance values in the AffectButton. Image adapted from Broekens and Brinkman (2013).	76
3.9	Front and Side plots of the possible PAD values in the AffectButton PAD space.	77
4.1	Graph of relationship between human likeness and perceived familiarity, as proposed by Mori (1970): familiarity increases with human likeness until a point is reached where subtle deviations from the human appearance evoke an adverse response. This is known as the <i>Uncanny Valley</i> . Figure adapted from MacDorman and Ishiguro (2006).	83
4.2	Images of the two robots used in the experiment.	86
4.3	Overall Percentage of Affective Ratings across both Robots and Utterance Categories (within bar makings indicate statistical significance that is above chance). * : $p < 0.05$, ** : $p < 0.025$, *** : $p < 0.01$, **** : $p < 0.005$. See table B.1.	90
4.4	Overall Percentage of Intention Ratings across both Robots and Utterance Categories (within bar makings indicate statistical significance that is above chance). * : $p < 0.005$. See table B.4. . . .	92
4.5	Percentage of Appropriate Responses for the Robot Types Across the Utterance Classes. * : $p < 0.05$, ** : $p < 0.001$	94

4.6	Percentage of “Yes” Responses for the male and female subjects across the robot types and utterance classes. Plot also shows whether results are due to chance (marked within each bar), and whether there are significant differences between the genders (marked between two bars). * : $p < 0.05$, ** : $p < 0.01$, *** : $p < 0.005$. . .	95
4.7	Percentage of “Yes” Responses for the pet owners and non-pet owners subjects across the robot types and utterance classes. Plot also shows whether results are due to chance (marked within each bar), and whether there are significant differences between the subjects groups (marked between two bars). * : $p < 0.025$, ** : $p < 0.005$.	96
5.1	Image of the experimental setup in the classroom with two children and the experimenter.	108
5.2	Parallel Plots of the the NLU specifications in the Utterance Parameter space.	113
5.3	Plot of all the Affective Ratings for all the NLUs.	114
5.4	Box plot of the Dominance ratings for the utterances in Experiment #1with a tremolo value of 0, grouped by the the first contour shape in the utterance (see table C.2 for a summery of the descriptive statistics).	116
5.5	Box plot of the Arousal ratings for the all the utterances in Experiment #2 (regardless of the sound uint count), across the three different Rhythm values (see table C.4 for a summary of the descriptive statistics of the box plot).	118
5.6	Box Plots showing the difference in ratings grouped by Frequency Range for NLUs with 3 Sound Units, along the Dominance dimension in Experiment #3 (see table C.7 for descriptive statistics for this figure).	120
5.7	Box Plots showing the difference in ratings grouped by Frequency Range for NLUs with 5 Sound Units, along the Pleasure dimension in Experiment #3 (see table C.8 for descriptive statistics for this figure).	120
5.8	Box Plots showing the difference in ratings between the Sound Unit Count values, for each of the Affect Space dimensions in Experiment #3 (see table C.9 for descriptive statistics for this figure).	121
5.9	Box Plots showing the difference in ratings between the Sound Unit Count values, for each of the Affect Space dimensions in Experiment #4 (see table C.13 for descriptive statistics for this figure).	122
5.10	Box Plots showing the difference in ratings between Parameter Configurations 1 and 2 in Experiment #5, regardless of the Sound Unit Count, across each of the three affective dimensions (see table C.14 for the descriptive statistics for this figure).	125
6.1	Example of the dynamics of class membership associated with Categorical Perception.	135
6.2	Spectrograms of the 12 NLUs used as the stimuli.	140
6.3	Images of the subjects’ laptop screen during each of the three tasks in the experiment.	141

6.4	Plot of the Pleasure/Dominance values as a function of the horizontal onscreen cursor position and the resulting AffectButton prototype expressions associated with the PAD values (from left to right): <i>neutral</i> (0, -1, 0), <i>sad</i> (-0.5, -1, -0.5), <i>neutral</i> (0, -1, 0), <i>excited</i> (1, 1, 1), <i>neutral</i> (0, -1, 0).	144
6.5	Image of the experimental setups for both the adult and children subjects.	145
6.6	Plots of the Mean values and Standard Deviations of the ratings for each affective label in the Labelling Task. These values are summarised in table D.1. Figures D.1 and D.2 show the AffectButton facial gestures for the mean values for the adults and children respectively.	148
6.7	Bar graphs showing the percentage of “different” ratings given by the adults for the neighbouring utterance AX pairs for both Stimulus Sets. Bars marked with a star are ratings found to be significantly above chance at the 0.05 level. The ratings shown in this figure are summarised in table 6.7.	155
6.8	Bar graphs showing the percentage of “different” ratings given by the children for the neighbouring utterance AX pairs for both Stimulus Sets. Bars marked with a star are ratings found to be significantly above chance at the 0.05 level. The ratings shown in this figure are summarised in table 6.9.	157
6.9	Plots showing the mean values and 95% confidence intervals of the adult ratings of each Utterance Parameter specification across the Stimulus Sets (figure 6.9a), and subject genders (figure 6.9b). The descriptive statistics for these figures are summarised in tables D.2 and D.3.	161
6.10	Plots showing the mean values and 95% confidence intervals of the child ratings of each Utterance Parameter specification across the Stimulus Sets (figure 6.10a), and subject genders (figure 6.10b). The descriptive statistics for these figures are summarised in tables D.4 and D.5.	163
7.1	Illustrative examples of a Classification problem and a Regression problem which may both be solved through Supervised Learning.	183
7.2	Plots of the Input and Output training data.	186
7.3	Multi Layer Perceptron topologies explored for mapping affective input values with output NLU algorithm parameters.	190
7.4	Example of a hidden layer with a sigmoid transfer function, and an output layer with a linear transfer function.	191
7.5	Plots of the mean errors (based upon 50 ANNs) in the Training data and Validation data sets over the ANN training cycles (epochs), for each NLU parameter.	195
7.6	Plots of the ANN output values (of 50 ANNs) vs the target output values for each of the NLU parameters. See table 7.4 for the Correlation Coefficients.	198
7.7	Plots of average mappings learnt by the independent ANNs for each of the 8 parameters.	203

7.8	AffectButton PAD values used as inputs to the trained ANNs used to obtain the output utterance parameter values used to generate the experimental NLUs.	207
7.9	Normalised ANN output values for the utterance parameter obtained from the PAD inputs.	208
7.10	Spectrograms of NLUs with the two difference pitch contour specifications.	208
7.11	Images of the subjects' laptop for each of the three tasks performed during the experiment.	210
7.12	Image of the experimental setup.	211
7.13	Plot of the PAD ratings for the 26 NLUs presented to subjects during the Identification Task.	212
7.14	Mean and Standard Deviations of the ratings for each NLU specification. Blue data shows the ratings for CS1 and red points for CS2. The green point shows the original PAD input value used to generate the NLU parameters. The descriptive statistics for these plots are summarised in table E.1.	216
7.15	Plots of the mean and standard deviations of the ratings for each affective label in the Labelling Task (see table 7.16 for a summary of the descriptive statistics). Figure E.2 shows the AffectButton facial gestures for the mean ratings for each of the affective labels.	217
8.1	Example of how the video conditions C_{Action} , C_{NLU}^P and C_{Action}^P may hypothetically be rated, for each of the two Hypotheses.	235
8.2	Setup of the professional audio equipment used to capture the audio. The video recorder is located at the right hand side of the image (just out of sight).	236
8.3	Bar graph showing the mean and standard deviations for the ratings of the Action Videos in the pilot study. These ratings are summarised in table F.3.	239
8.4	Bar graph showing the mean and standard deviations for the ratings of the NLU videos in the pilot study. These ratings are summarised in table F.4.	240
8.5	Images from the videos showing the five different action scenarios used in the final study, selected via the pilot study.	241
8.6	Bar graph showing the Mean values and 95% Confidence Interval for the valence ratings for each of the 5 NLU videos. These ratings are summarised in table F.5.	244
8.7	Bar graph showing the Mean values and 95% Confidence Interval for the valence ratings for each of the 5 Action videos. These ratings are summarised in table F.6.	245
8.8	Results of the 3-way repeated measures 5x2x2 ANOVA showing the Mean ratings for the videos as well as the 95% Confidence Interval for each of the 5 video conditions. across the 5 action scenarios. Primary statistically significant differences are shown, and the other significant differences may be inferred from those already displayed. These ratings are summarized in table F.7	246

8.9	Plots of the interaction effects identified through the 3-way ANOVAs. Each plot shows the Mean values and 95% CI for the ratings across each of the five video conditions, with each line either representing the subject gender or robot familiarity factors.	247
9.1	Frames from the videos depicting each of the four components to the cups and balls game.	265
9.2	Spectrograms of the 12 NLUs used as stimulus.	270
9.3	Images of the apparatus used in the stimulus recording.	271
9.4	Plots showing the mean and 95% confidence intervals for the appropriateness ratings and interaction between male/female subjects and video conditions, with the data split across the robot familiarity factor.	276
9.5	Plots showing the mean and 95% confidence intervals for the appropriateness ratings and interaction between subjects familiar and unfamiliar with the robot and the video conditions, with the data split across the subject gender factor.	278
9.6	Plots showing the mean and 95% confidence intervals for the expressiveness ratings and interaction between male/female subjects and video conditions, with the data split across the robot familiarity factor.	280
9.7	Plots showing the mean and 95% confidence intervals for the expressiveness ratings and interaction between the subjects familiar and unfamiliar with the robot and video conditions, with the data split across the subject gender factor.	282
9.8	Plot of the mean ratings and 95% Confidence Intervals for the Preference ratings for all the subjects combined, and the subjects whom were familiar/unfamiliar with the robot.	283
9.9	Plot of the mean ratings and 95% Confidence Intervals for the Naturalness ratings across the four video conditions, for all the subjects collectively and subject genders.	284
9.10	Plot of the mean ratings and 95% Confidence Intervals of the Likeability across the four video condition for all the subjects collectively and the genders, and the interaction effect between the subject gender and robot familiarity.	287
B.1	Overall percentage of the Affective Ratings across both the two robot and subject gender. This is shown for each of the three utterance categories.	339
B.2	Overall percentage of the Affective Ratings across both the two robot and pet/non-pet owners. This is shown for each of the three utterance categories.	340
B.3	Overall percentage of the Intention Ratings across both the two robot and subject gender. This is shown for each of the three utterance categories.	344
B.4	Overall percentage of the Intention Ratings across both the two robot and pet/non-pet owners. This is shown for each of the three utterance categories.	347
D.1	AffectButton facial gestures of the mean ratings given by the adult subjects for each of the affective labels in the Labelling Task. . . .	359

D.2	AffectButton facial gestures of the mean ratings given by the children subjects for each of the affective labels in the Labelling Task.	359
D.3	Mean AffectButton facial gestures for both the Stimulus Set ratings provided by the adult subjects during the Identification Task.	360
D.4	Mean AffectButton facial gestures for the Stimulus Set 1 ratings provided by the adult subjects during the Identification Task.	361
D.5	Mean AffectButton facial gestures for the Stimulus Set 2 ratings provided by the adult subjects during the Identification Task.	361
D.6	Mean AffectButton facial gestures for both the Stimulus Set ratings provided by the child subjects during the Identification Task.	362
D.7	Mean AffectButton facial gestures for the Stimulus Set 1 ratings provided by the child subjects during the Identification Task.	363
D.8	Mean AffectButton facial gestures for the Stimulus Set 2 ratings provided by the child subjects during the Identification Task.	364
E.1	Images of the AffectButton facial gestures used during the Matching Task in the ANN Evaluation experiment presented in chapter 7.	369
E.2	Images of the AffectButton facial gestures for the mean ratings obtained during the Labelling Task in the ANN Evaluation experiment presented in chapter 7.	371
F.1	Spectrograms of the NLUs outlined in table F.1	376

List of Tables

2.1	Description of the Acoustic Cues in Vocal Expression. Table adapted from Juslin and Laukka (2003).	31
2.2	Summary of the acoustic patterning of the human voice for the basic emotions. Table adapted from Scherer (2003).	33
2.3	Acoustic parameters used to generate utterances in previous work, the emotions that were portrayed and the means of affective measurement.	46
2.4	Number of subjects and subject age ranges in reviewed previous work.	48
3.1	Values of the five nodes for the amplitude envelope located in the bi-normalised space for each carrier signal in each sound unit (see figure 3.2). These values have been held constant throughout the work described in this thesis.	58
3.2	Summary of the parameters for characterising NLUs in this work.	62
4.1	Pairing of Emotion/State Classifications.	87
4.2	Krippendorff's α values showing the agreement between subjects in their affective (Happiness, Sadness, Relaxation, Anger, Affection, Fear, Interest, Boredom and Disgust) ratings of the different classes of utterances.	89
4.3	Krippendorff's α values showing the agreement between subjects in their interpretation ratings (Approval, Attention, Prohibition, Comfort and Neutral) of the different classes of utterances.	93
4.4	Krippendorff's α values showing the agreement between subjects in their judgement of the appropriateness of different classes of utterances with the robot image.	97
5.1	Overview of the nine different Pitch Contour profiles for the utterances in experiment #1. Sound unit pitch contours are encoded as follows: F = flat contour, U = rising contour and D = falling contour.	111
5.2	Specifications of the Utterance Parameter Configurations (PC) for the five mini experiments.	111
5.3	Pitch Contour specifications for Experiment #2 through to Experiment #5, across the different sound unit counts. Contours are encoded as follows: F = Flat, U = Rising, D = Falling, Ud = Rising-Falling, Du = Falling-Rising.	112
5.4	The two Utterance Parameter Configurations for mini experiment #5	112

5.5	Results of Kruskal-Wallis tests for Experiment #1, comparing the Pitch Contour and Tremolo parameter specifications against the affective ratings. Pitch Contour has three possible groupings: by the whole pitch contour combination, by the pitch contour of only the first sound unit, or by the contour of only the last sound unit.	116
5.6	Results of the Kruskal-Wallis tests in Experiment #2, testing the influence of the Rhythm parameter, with the data grouped by the sound unit count.	117
5.7	Results of the Kruskal-Wallis tests in Experiment #3, collapsing the Base Frequency and testing only for the influence of the Frequency Range across the different sound unit counts and affective dimensions.	119
5.8	Results of the Kruskal-Wallis tests in Experiment #3, comparing the difference in ratings across the two difference sound unit counts.	119
5.9	Results of the Kruskal-Wallis tests, checking for the differences in ratings due to the different Sound Unit Counts (3 and 5) of the utterances, along each affective dimension in Experiment #4.	122
5.10	Results of the Kruskal-Wallis tests checking for differences in ratings due to the two different Utterance Parameter configurations (PC1 and PC2) in Experiment #5, with data grouped by the utterance Sound Unit Count (3, 5 and both). The table shows the degrees of freedom, χ^2 values and p values of the tests that were performed for each affective dimension individually.	124
5.11	Results of the Kruskal-Wallis tests checking for significant differences between the five different pitch contour specifications in Experiment #5, with data grouped by Parameter Configuration (PC1 and PC2) and the Sound Unit Count (3 and 5) of an utterance. Tests show the degrees of freedom, χ^2 values and p values of the tests that were performed for each affective dimension individually.	124
6.1	Utterance Parameter configurations for each utterance in both Stimulus Sets 1 and 2 (see figure 6.2).	139
6.2	Pitch Contour specifications for the utterances in Stimulus Set 1 and Set 2 (see figure 6.2).	139
6.3	Affective co-ordinates in the AffectButton affect space of the labels (and associated prototypical facial gestures) used during the Labelling Task. Note that Calm and Relaxed are no prototypes used in the AffectButton.	142
6.4	Overview of the neighbouring utterance pair comparisons in the Discrimination Task (note that A and X were randomly ordered).	143
6.5	Results of the post-hoc Friedman pairwise comparisons for the adults's affective ratings for the affective labels in the Labelling Task. The table show the $\chi^2(1)$ results and indicate the associated p -value for each dimension of the AffectButton affect space independently.	150
6.6	Results of the post-hoc Friedman pairwise comparisons for the children's affective ratings for the affective labels in the Labelling Task. The table show the $\chi^2(1)$ results and indicate the associated p -value for each dimension of the AffectButton affect space independently.	152

6.7	χ^2 Goodness of fit tests for the adult subjects' comparison of neighbouring utterances in each of the Stimulus Sets.	154
6.8	Results of the two way independent samples χ^2 tests checking for significant differences between the genders in their ratings of each utterance pair. The table shows the χ^2 statistic and p value for both the adult and child subjects across the two Stimulus Sets. . .	156
6.9	χ^2 Goodness of fit tests for the child subjects' comparison of neighbouring utterances in each of the Stimulus Sets.	158
6.10	Mean values, Standard Errors and 95% Confidence Intervals of the adult ratings for each of the Utterance Parameter configurations. .	160
6.11	Mean values, Standard Errors and 95% Confidence Intervals of the child ratings for each of the Utterance Parameter configurations. .	162
7.1	Break down of training data set with respect to the original experiments within which the data was collected.	187
7.2	Schema for encoding the Contour shape of a given Sound Unit in the training data.	187
7.3	Performance values for the eight trained ANNs. The table shows the number of Epochs performed during training, and the performance on the Training, Validation and Test data sets after training was completed, as well as the final gradient values.	196
7.4	Correlation Coefficients (ρ) between the target output values and the ANN output values (based on 50 ANNs), for each NLU parameter.	197
7.5	Mean, Standard Deviations, Minimum and Maximum values for the ANN Mappings for each of the utterance parameters.	199
7.6	Partial Linear Correlation Coefficients (ρ) between the affect space dimensions and the generation parameter values output from the ANN.	199
7.7	General description of utterance characteristics in different regions of the AffectButton PAD affect space.	201
7.8	Spearman's ρ Correlation Coefficients for the correlation between the different ANN output values of the utterance parameters. . . .	204
7.9	Obtained Mappings between the PAD values input to the ANNs and the output parameter values, scaled to fit the working range of the NLU generation algorithm.	209
7.10	Cronbach's α values indicating the degree of agreement between subjects in their ratings of the NLUs. α values are shown for ratings grouped by gender, and for each of the pitch contour specifications.	213
7.11	Pearson Correlation Coefficients (ρ) between the input and observed PAD values. None of the ρ values are statistically significant at the 0.05 level.	213
7.12	Partial Linear Correlation Coefficients (ρ) between the parameters of the experimental NLUs and their PAD affective ratings.	214
7.13	Results of the Friedman tests, testing for significant differences in the affective ratings for each of the NLUs. Ratings are grouped by gender as well as the contour specification. The table shows the degrees of freedom for each test, the χ^2 values and the associated p values.	215

7.14	Results of the Friedman tests, testing for significant differences in the ratings across the two different Pitch Contour Specifications with ratings grouped by subject gender. The table shows the degrees of freedom for each test, the χ^2 values and the associated p values.	215
7.15	Cronbach's α values calculated for the results obtained during the Labelling Task, across the subjects genders.	217
7.16	Mean and standard deviations of the ratings for each of the affective labels in the Labelling Task, along each of the affect space dimensions.	218
7.17	Results of the pairwise Friedman test comparing the ratings between of the affective labels in the Labelling Task. The table shows the $\chi^2(1)$ values with an indication of the degree of statistical significance along each affect space dimension.	219
9.1	Breakdown of Language and NLU's used across each of the four video conditions. Note the inverses use of NLU's and language in conditions 3 and 4.	267
9.2	The text input to the TTS engine to produce the language samples used during the videos.	267
9.3	Specification of the Generation Parameters used to generate each NLU.	268
9.4	Cronbach's α ratings for each of the rating scales, across all of the language/NLU conditions.	273
B.1	Ratings and χ^2 values for overall Affective Ratings across robot and sound categories. The columns for each utterance class are the overall percentage of ratings for a given affective label, the results of a χ^2 test checking that that rating is above chance, and the results of a Stewart-Maxwell test indicating whether there was an overall difference in the distribution of ratings due to the robot image that was also presented. Please see Figures 4.3a, 4.3b and 4.3c	337
B.2	Gender differences for Affective Ratings across the Utterance Classes and Robot Morphologies. The table shows the percentages of each affective ratings for each robot across the two genders and the $\chi^2(8)$ values indicating whether this rating is above chance (compared to a flat uniform distribution).	338
B.3	Pet Ownership differences for Affective Ratings across the Utterance Classes and Robot Morphologies. The table shows the percentages of each affective ratings for each robot across the two genders and the $\chi^2(8)$ values indicating whether this rating is above chance (compared to a flat uniform distribution).	341
B.4	Ratings and χ^2 values for Intentional Ratings across both robot types and sound categories. The columns for each utterance class are the overall percentage of ratings for a given intentional label, the results of a χ^2 test checking that that rating is above chance, and the results of a Stewart-Maxwell test indicating whether there was an overall difference in the distribution of ratings due to the robot image that was also presented	342

B.5	Gender differences for the Intention ratings. The table shows the percentages of each affective ratings for each robot across the two genders and the $\chi^2(8)$ values indicating whether this rating is above chance (compared to a flat uniform distribution).	343
B.6	Pet Ownership differences for the Intention ratings. The table shows the percentages of each affective ratings for each robot across the two genders and the $\chi^2(8)$ values indicating whether this rating is above chance (compared to a flat uniform distribution).	345
B.7	Appropriateness ratings for utterances and associated χ^2 values for the one-way χ^2 test comparing the ratings a flat uniform distribution, and the χ^2 values of the McNemar tests checking whether the difference in the rating distributions across the two robots are significantly different.	346
B.8	Appropriateness ratings for utterances and associated χ^2 values for the one-way χ^2 test comparing the ratings a flat uniform distribution, and the χ^2 values of the two-way tests checking whether the difference in the rating distributions between the genders are significantly different.	346
B.9	Appropriateness ratings for utterances and associated χ^2 values for the one-way χ^2 test comparing the ratings a flat uniform distribution, and the χ^2 values of the two-way tests checking whether the difference in the rating distributions between pet/non-pet owners are significantly different.	346
C.1	Results of the Kruskal-Wallice tests testing the difference in affective ratings due to the Tremolo values in Experiment #1.	351
C.2	Descriptive statistics for the Box Plots shown in figure 5.4, showing the ratings for the Utterances with a Tremolo value of 0, grouped by the first pitch contour shape. The table shows the Median values, 1st and 3rd Quartiles, Inter-Quartile Range, and the Lower and Upper estimated 95% Median Confidence Intervals.	351
C.3	Results of the Kruskal-Wallice tests checking for differences in the affective ratings across the two different sound unit counts in Experiment #2.	351
C.4	Descriptive statistics for the Box Plots shown in figure 5.5, showing the ratings for the Utterances with a Rhythm value of 0, 0.5 and 1. The table shows the Median values, 1st and 3rd Quartiles, Inter-Quartile Range and the Lower and Upper estimated 95% Median Confidence Intervals.	351
C.5	Results of the Kruskal-Wallice tests, collapsing the Frequency Range and testing only for the influence of the Base Frequency across the different sound unit counts, along each affective dimensions.	352
C.6	Results of the Kruskal-Wallice tests with interleaved Base Frequency and Frequency Range values, across the different sound unit counts, along each affective dimensions.	352
C.7	Descriptive statistics for the Box plots in figure 5.6, showing the ratings for all Utterances with 3 Sound Units, grouped by the Frequency Range values, along the Dominance dimension. Table shows the Median values, 1st and 3rd Quartiles, Inter-Quartile Range, and the lower and upper estimated Median Confidence Intervals.	352

C.8	Descriptive statistics for the Box plots in figure 5.7, showing the ratings for all Utterances with 5 Sound Units, grouped by the Frequency Range values, along the Pleasure dimension. Table shows the Median values, 1st and 3rd Quartiles, Inter-Quartile Range, and the lower and upper estimated Median Confidence Intervals.	353
C.9	Descriptive Statistics for the Box Plots in figure 5.8, showing the ratings for utterances across two two difference Sound Unit Counts, for each affective dimension. Table shows the Median values, 1st and 3rd Quartiles, Inter-Quartile Range, and the lower and upper estimated Median Confidence Intervals.	353
C.10	Results of the Kruskal-Wallice tests, collapsing the Speech Rate and testing only for the influence of the Pause Ratio across the difference sound unit counts, along each affective dimension.	354
C.11	Results of the Kruskal-Wallice tests, collapsing the Pause Ratio and testing only for the influence of the Speech Rate across the difference sound unit counts, along each affective dimension.	354
C.12	Results of the Kruskal-Wallice tests, interleaving the Pause Ratio and Speech Rate, across the different sound unit counts, along each affective dimension.	354
C.13	Descriptive Statistics for the Box Plots in figure 5.9. Table shows the Median values, 1st and 3rd Quartiles, Inter-Quartile Range, and the lower and upper estimated Median Confidence Intervals.	355
C.14	Descriptive Statistics for the Box Plots in figure 5.10. Table shows the Median values, 1st and 3rd Quartiles, Inter-Quartile Range, and the lower and upper estimated Median Confidence Intervals.	356
C.15	Results of the Kruskal-Wallice tests checking for differences in ratings, along each affective dimension, between PC2 and PC3 in Experiment #5.	356
D.1	Mean values and Standard Deviations of the ratings for each of the affective labels presented in the Labelling Task. The table shows the values for the adults and children, along each affective dimension of the AffectButton.	360
D.2	Mean values and 95% confidence intervals for the overall adult ratings of stimuli in the Identification Task. The table shows the values for the overall different Utterance Parameter configurations, as well as those for Stimulus Set 1 and Stimulus Set 2.	362
D.3	Mean values and 95% confidence intervals for the overall adults ratings of stimuli in the Identification Task. The table shows the values for the overall different Utterance Parameter configurations, as well as those for the two genders.	363
D.4	Mean values and 95% confidence intervals for the overall child ratings of stimuli in the Identification Task. The table shows the values for the overall different Utterance Parameter configurations, as well as those for Stimulus Set 1 and Stimulus Set 2.	364
D.5	Mean values and 95% confidence intervals for the overall child ratings of stimuli in the Identification Task. The table shows the values for the overall different Utterance Parameter configurations, as well as those for the two genders.	365

E.1	Mean and Standard Deviations for the Affective PAD ratings for each NLU parameter specifications, across the two pitch contour specifications.	370
F.1	Parameter Configurations of the NLUs used in the Pilot and Main Survey presented in chapter 8. Spectrograms of these are shown in figure F.1 and outline the Pitch Contour	375
F.2	Question orders for the online CrowdFlower surveys. Questions with a (0) are action videos with no NLU (C_{Action}). Questions with a (1) are action videos combined with NLU #8 (C_{Action}^P). Questions with a (2) are action videos combined with NLU #7 (C_{Action}^N). . .	377
F.3	Mean and Standard Deviations of the Ratings for the Action Videos in the Pilot Study (refer to section 8.1.1.3 and figure 8.3).	377
F.4	Mean and Standard Deviations of the Ratings for the NLU Videos in the Pilot Study (refer to section 8.1.1.3 and figure 8.4).	378
F.5	Mean, Standard Error and 95% Confidence Intervals of the NLU Video Ratings obtained from the CrowdFlower Study (refer to section 8.2.1 and figure 8.6).	378
F.6	Mean, Standard Error and 95% Confidence Intervals of the Action Video Ratings obtained from the CrowdFlower Study (refer to section 8.2.2 and figure 8.7).	378
F.7	Mean, Standard Error and 95% Confidence Intervals for the overall ratings for the Video Conditions, across each action scenario (refer to section 8.2.3 and figure 8.8).	379
F.8	Mean, Standard Error and 95% Confidence Intervals for the ratings of the interaction effect between Subject Gender and the Video Condition for the Flicking Action (refer to section 8.2.3.3 and figure 8.9a).	379
F.9	Mean, Standard Error and 95% Confidence Intervals for the ratings of the interaction effect between Subject Gender and the Video Condition for the Stroking Action (refer to section 8.2.3.4 and figure 8.9b).	380
F.10	Mean, Standard Error and 95% Confidence Intervals for the ratings of the interaction effect between Subject Gender and the Video Condition for the Eye Covering Action (refer to section 8.2.3.5 and figure 8.9c).	380
F.11	Mean, Standard Error and 95% Confidence Intervals for the ratings of the interaction effect between Robot Familiarity and the Video Condition for the Eye Covering Action (refer to section 8.2.3.5 and figure 8.9d).	380
G.1	Results of the 2-way ANOVAs for Appropriateness ratings with the subjects split across the robot familiarity factor (refer to section 9.2.1.1). The Tables shows the Mean Values, Standard Errors and 95% Confidence Intervals. See figure 9.4.	383
G.2	Results of the 2-way ANOVAs for Appropriateness ratings with the subjects split across the subject gender factor (refer to section 9.2.1.2). The Tables shows the Mean Values, Standard Errors and 95% Confidence Intervals. See figure 9.5.	384

G.3	Results of the 2-way ANOVAs for Expressiveness ratings with the subjects split across the robot familiarity factor (refer to section 9.2.2.1). The Tables shows the Mean Values, Standard Errors and 95% Confidence Intervals. See figure 9.6.	385
G.4	Results of the 2-way ANOVAs for Expressiveness ratings with the subjects split across the subject gender factor (refer to section 9.2.2.2). The Tables shows the Mean Values, Standard Errors and 95% Confidence Intervals. See figure 9.7.	386
G.5	Results of the Univariate tests for the Preference ratings (refer to section 9.2.3). The table shows the Mean values, Standard Errors and 95% Confidence Intervals for the Main effects due to the video condition and robot familiarity factors. See figure 9.8.	387
G.6	Results of the Univariate tests for the Naturalness Ratings (refer to section 9.2.4). The table shows the Mean values, Standard Errors and 95% Confidence Intervals for the Interaction effect found between the video condition and subject gender factors. See figure 9.9.	388
G.7	Results of the Univariate tests for the Like-ability Ratings (refer to section 9.2.5 and figure 9.10a) The table shows the Mean values, Standard Errors and 95% Confidence Intervals for the Interaction effect between the video condition and subject gender.	389
G.8	Results of the Univariate tests for the Like-ability, Subject Gender and Robot Familiarity Interaction (refer to section 9.2.5 and figure 9.10b). The table shows the Mean values, Standard Errors and 95% Confidence Intervals for the Interaction effect found between the subject gender and robot familiarity.	389

Acknowledgements

As with any large body of work, this thesis would not have been what it is without input and influence in a variety of different ways from a large number of people. I want to take the time to thank many of you here.

First and foremost, I must thank my Director of Studies, Prof. Tony Belpaeme. Tony, to be frank, without you, this thesis would not exist. You've been an invaluable part of this research, and have been there at my side throughout my scientific exploration. You have taught me the scientific process as it applies the field of social HRI, and much, much more. When times got a bit tough, you were always quick to spot it and provide the support and guidance that I needed and have always been able to give me that extra boost of confidence in what I was doing, particularly in those times when I was skeptical and unsure. You've also given me countless opportunities throughout this PhD and done many things to make my life easier; from introducing me to the ALIZ-E project, to providing the Nao robot as hardware, funding opportunities, conference opportunities, replacing that Windows machine with a shiny new Mac Pro (that did actually make a big difference for me), and handling all kinds of odd bits of paperwork that were frankly beyond me. The list goes on, and on. I know that these examples only touch the surface, and I fear that this bit of text will never truly express the full extent of the gratitude that I have for everything that you've done for me over the last four years. I can only repeat myself, thank you so very much!

I would like to thank my examiners, Prof. Angelo Cangelosi and Dr. Kai Arras, and my chair, Dr. Davide Marocco, for all your time and efforts. You made the whole experience very enjoyable and rewarding, and you provided me with very informative and valuable feedback as well as many inspirational ideas.

To Paul Baxter, Rachel Wood, Joachim De Greeff, James Kennedy, and when you've popped in, Tony Belpaeme and Joanne Clements. I've thoroughly enjoyed sharing an office space with you throughout my PhD and the ALIZ-E project. The enthusiastic, warm, friendly, stimulating and welcoming environment that you have all helped create and shape has been an important corner stone for me in which to develop. You have also given me my first experience of a large scale research project, and everything that I have learnt is sure to go a long, long way. The discussions that I've had with you have been inspirational, productive and of course, at times, hilariously funny and entertaining. I very much hope that we can work together in the future and that we do keep in touch when the time does eventually come for us to part ways. Thanks also to all the (other) people who have been, or still are members of the ALIZ-E consortium. There are a very many of you, and I have relished the times that we have spent together when we have met.

I am indebted to my proofreaders, Paul, Joachim, James and Tony. You have been incredibly efficient in your turn around of some rather long and densely populated bits of text, providing very important feedback that has significantly

improved and helped reshape this thesis into what it is now, as well as spotting all those little silly mistakes that I missed when I couldn't see the trees from the wood anymore.

It cannot go without saying that my surrounding, professional working environment has had a considerable impact on shaping my ideas and research, and so I owe a great deal of thanks to the people who have been involved in the ABC Lab and the Centre for Robotics and Neural Systems at various times throughout my PhD. In no particular order, I'd like to thank Tony Belpaeme, Paul Baxter, Rachel Wood, Joachim De Greeff, James Kennedy, Joanne Clements, Angelo Cangelosi, Davide Marocco, Guido Bugmann, Phil Culverhouse, Tony Morse, Salomon Ramirez-Contla, Alex Smith, Martin Peniak, Frédéric Delaunay, Sam Adams, Frank Broz, Alessandro Di Nuovo, Beata Grzyb, Michael Klein, Anna-Lisa Vollmer, Federico Da Rold, Marek Rucinski, Francesca Stramandinoli, Giuseppe Filippone, Nicholas Hemion, Victor Gonzales, Hande Ceilikkanat, Martin Stoelen, Fernando Alonso, Joris Bleys, Hiroyuki Iizuka, Naveen Kuppuswamy, Elena Dell'Aquila, Georgios Pierris, Fabio Ruini, Zoran Macura and Christopher Ford.

Many thanks to both Lucy Cheetham and Carole Watson. While we have had limited interactions over the last four years, you have always made things easier than they could have been. As people say, it's the little things that count too, and you have been two figures who have played a background, but very supportive role that has been much appreciated.

I would also like to thank the other members of what I consider to be my PhD "class": James Humble, Tim Rumble and Salomon Ramirez-Contla. Though we have all finished at different times over the last year or so, I like to think of us as a little group of people who each decided to embark on a very strange journey at the same time. That has created a small but special bond for me, and am very glad that we've all made it out the other end in one piece! I wish you all the best of luck in your future careers as Doctors!

While here in Plymouth I have made some very close friends, which I hope are for life. Tom Roc, Linda Baumane, Salomon Ramirez-Contla, Paola Garcia Meneses, Joachim De Greeff, Marieke Verhagen, Rob and Jennie Schindler, Dries Trippas and Laura König, we have shared some amazing times together that have been vital distractions for me and I look forward to many more to come!

Here in Plymouth I have also lived with a number of exceptional people: Tom Roc, Linda Baumane, Chris Rees, Stuart McGeachie, Clarissa Ribeiro, Joao Gomes. I have very much enjoyed the times that I shared with you while we were living together and I wish you all the very best in your future endeavours.

To Claudia, the last 6 months of my writing up was a roller coaster ride of emotions and stress. You helped see me through it and provided me with a vital distraction and support. Thank you for your love.

Finally, to my family. Mum, Dad and Mark. You have all been vital factors in determining who I have become, and in getting me to where I am today. Close or far, you have always been there for me and given me endless support in whatever I have set out to achieve in life. A further note to Dad also. We have spent many, many hours on the phone discussing the work that I have been doing, and through this you have managed to establish a very firm grasp of the nature of my research and the intricacies that it has, well done! I have always come away from our chats with a greater sense of clarity.

Author's declaration

At no time during the registration for the degree of Doctor of Philosophy has the author been registered to any other University award without prior agreement of the Graduate Committee.

This work has been carried out by Robin Read under the supervision of Prof. Dr. Tony Belpaeme.

Publications

Read, R. and Belpaeme, T. (2010). Interpreting Non-Linguistic Utterances by Robots : Studying the Influence of Physical Appearance. In *Proceedings of the 3rd International Workshop on Affective Interaction in Natural Environments (AFFINE 2010) at ACM Multimedia 2010*, pages 65–70, Firenze, Italy. ACM

Read, R. and Belpaeme, T. (2012). How to Use Non-Linguistic Utterances to Convey Emotion in Child-Robot Interaction. In *Proceedings of the 7th International Conference on Human-Robot Interaction (HRI'12)*, pages 219–220, Boston, MA, U.S.A. ACM/IEEE

Read, R. and Belpaeme, T. (2013a). People Interpret Robotic Non-Linguistic Utterances Categorically. In *Proceedings of the 8th International Conference on Human-Robot Interaction (HRI'13)*, pages 209–210, Tokyo, Japan. ACM/IEEE

Read, R. and Belpaeme, T. (2013b). Using the AffectButton to Measure Affect in Child and Adult-Robot Interaction. In *Proceedings of the 8th International Conference on Human-Robot Interaction (HRI'13)*, pages 211–212, Tokyo, Japan. ACM/IEEE

Read, R. and Belpaeme, T. (2014b). Situational Context Directs How People Affectively Interpret Robotic Non-Linguistic Utterances. In *Proceedings of the 9th International Conference on Human-Robot Interaction (HRI'14)*, Bielefeld, Germany. ACM/IEEE

Read, R. and Belpaeme, T. (2014a). Non-Linguistic Utterances Should be Used Alongside Language, Rather than on their Own or as a Replacement. In *Proceedings of the 9th International Conference on Human-Robot Interaction (HRI'14)*, Bielefeld, Germany. ACM/IEEE

Word count for the main body of this thesis: 78,092.

Signed: _____

Date: _____

Chapter 1

Introduction

Social interaction with others is a capability of humans that comes effortlessly. From birth until the end of their life, humans are constantly exposed to and engaged in social behaviour and interactions with their parents, peers and offspring. The ability to interact with groups of other human beings in a seamless and coherent manner is arguably deeply intertwined with our general development, and as a result, it has been suggested that social interaction and collaboration has also helped shaped *how* the human intelligence has evolved and developed over the centuries. This is known as the *Social Intelligence Hypothesis* (Holekamp, 2007).

It is through our physical embodiment, and the affordances that it provides, that social interaction has been facilitated. Over time, these modalities have become more specialised in their function and capability as has our use of these. This is particularly the case of the human vocal tract which provides the means to articulate a range of complex acoustic signals, as well as having the ability to encode efficiently complex and meaningful information via this single modality

As the technologies around us have developed and become more advanced, so too as their integration into our daily lives and society in general, through a large number of different application domains. As a result, and in order to further harness the impact that they have, modern technologies have begun to undergo a transformation where they allow the layman to gain access to, and use these technologies in a manner that is as natural and intuitive as possible, where there is minimal requirement for having technology specific skills and no need for

the user to adapt to the technology in order gain harness the benefits it has to offer. In essence, the user interfaces are becoming more aligned with how people naturally interact with objects and other people that share the same environment as them - these interfaces are becoming more *socially capable*. The aim is to provide technology with an interface that allows seamless bridging between how people interact with each other, and how they interact and use the technologies around them, such that these technologies fall into the background and become *invisible*. Perhaps the most prominent ways in which this is currently being done is throughout exploitation of advances in computer speech, where technology is being furnished with natural language capabilities that allow people and technology to exchange a rich and complex array of information with each other in a fast and efficient manner.

A prime example of a technology undergoing such a transformation is robotic technology, where robots are now being designed so that they are able to function in and negotiate the physical world around them, and are able to engage in a natural manner with the people that they share this physical world with through multi-modal interaction laden with social and affective cues. It is this facet of robotic technology that the thesis presented here is concerned with.

Setting the stage

The nature of the work in this thesis is interdisciplinary, drawing insight and inspiration from related fields such as psychology and speech synthesis. As such, themes in these related fields will be brought to the foreground. The remainder of this chapter serves to do this, and sets the general backdrop to which the related fields and this thesis itself can be viewed. We shall briefly introduce the field of Human-Robot Interaction and briefly outline why robots are becoming more social and emotional, explain what the main topic of the thesis is - Non-Linguistic Utterances, and outline how related fields such as psychology and speech synthesis feed into the thesis at hand, that is studying how a social robot is able to use Non-Linguistic Utterances during social Human-Robot Interaction.

1.1 Human-Robot Interaction

While there are many robots already in operation in the world around us, the majority of the general population have not seen, or come into contact with them. These robots can primarily be found in facilities such as car factories or storage warehouses, but also include the depths of the seas, war zones, nuclear power plants, and outer space. A large part of this segregation between humans and robots is simply due to the physical, mechanical design of these robots. They have not been designed to co-inhabit an environment with humans, rather they have been designed to function in environments that are deemed unsuitable for people to inhabit safely or with ease. A second, but more potent problem however, is that robots have not been designed (with respect to both mechanical and software design) to be sensitive to humans, which means that they lack the capabilities to read and comprehend the social cues that people use, nor can they express their internal state to people using these same modalities and cues. The problem is simple, people do not understand what a robot *thinking*, robots do not understand what a person is *thinking*, and neither has the ability to *communicate* with the other through the same set of *social* channels. This leads to problems regarding the safety of both humans and robots, and the best solution has been to keep them segregated.

However, in the last decade or so, the fundamental attitudes towards the types of potential applications that robots can have has broadened considerably and now includes applications that require close contact and interaction with people. Examples of this are the use of robots in care homes for the elderly (Torta et al., 2012; Tapus et al., 2007; Mccoll and Nejat, 2013), or as assistive tools for people with Autism Spectrum Disorder (Robins et al., 2004; Kozima et al., 2009; Feil-seifer and Skinner, 2007). Through such potential applications, the field of Human-Robot Interaction (HRI) has emerged, and specifically, *social* HRI. Social HRI is concerned with the understanding how humans and robots are able to interact naturally and fluidly, and how robots can be better designed to do this. The field draws heavily from the related field of Human-Computer Interaction

(HCI), where established insights have provided to be very useful when built upon and tailored toward robots (Breazeal, 2004b). For example, the observation that people tend to naturally treat inanimate objects, such as computers, in a social manner and as socially competent (Reeves and Nass, 1996) has opened up many new possibilities for interactions and applications (Duffy, 2003). Through such tendencies, it is observed that a robot is more than just a computer in a mobile shell: it has something special which allows it to elicit natural, multi-modal social behaviour towards it from people who have never seen or interacted with it before, and when both the software and hardware geared toward exploiting this, a robot is able to utilise social channels to engage in multi-modal social interaction with people, and where people see the robot as a social partner (Breazeal, 2002, 2004a; Breazeal and Brooks, 2005). Through such social interaction, it is possible for both humans and people to express their inner states to others, which in the case of humans inherently includes *emotion*.

1.1.1 Emotion and Affective Robots

All complex, intelligent animals (including humans) have emotions and display them to some varying degree, and humans in particular are the most complex, expressive and sophisticated in this manner, with emotions biasing a great how we behave and function both as single entities and as a social group also (Darwin, 2009). From an evolutionary perspective, in order to function and survive in a complex, unpredictable and unstructured environment, people and animals are faced with a complex problem of how to allocate their limited resources in order to meet their multiple overall goals in an efficient and flexible manner (Breazeal and Brooks, 2005), and in the case of social species, this involves social communication with the rest of the group. Prominent theories state that emotions are used by the creature not only to evaluate events that occur within the environment and assess their impact and overall value with respect to the creature, but that they also influence the cognitive process in humans (Damasio, 1994; Ortony and Turner, 1990) as well as serving as a regulatory tool regarding the use of resources (e.g.

energy levels, muscles, limbs, perceptual systems, etc.).

In this light, as pointed out by Breazeal and Brooks (2005), it can be argued that robots, particularly those that are social, face many of the same problems: they have limited resources (in the form of motors, sensors, computational power, batteries, etc.), the world around them is also unstructured and unpredictable and they need to interact with other social creatures that inhabit the same physical space as them. Furthermore, given the prominent tendency for humans to treat machines as social and apply human-social models to these in order to understand their behaviour, and the important role that emotions play in living creatures, the modelling of affect at both a computational/cognitive levels, as well as at a *behavioural* level has been deemed to be vital for establishing effective and engaging HCI and HRI (Picard, 1997; Breazeal, 2004b). The work in this thesis is concerned with the latter, with a particular focus on how a robot is able to meaningfully express its inner affective to others through the medium of sound. One way of doing this is through the use of natural language and speech synthesis, however, as we shall see later, this approach has a number of shortcomings that can hinder HRI in general. However, taking inspiration from the world of animation, another seemingly fruitful approach is through the use of expressive, abstract sounds, which in this thesis are referred to as *Non-Linguistic Utterances*.

1.2 Non-Linguistic Utterances

Non-Linguistic Utterances (NLUs) are utterances that consist of beeps, squeaks and whirrs, rather than natural language. They have been used almost exclusively, to great effect, in the world of Animation and modern culture (R2D2 and Wall-E provide vivid examples of this). To take the infamous example of R2D2 from the Star Wars films, this robot does not speak a single word of any real natural language through any of the six films that have been made so far. It has in essence, three basic modalities (a single multi-coloured light, motion through the rotation of its head and body, and through sound) through which it expresses a rich and vivid variety of different emotions and social cues that bring the char-

acter to life. Arguably, the acoustic modality is the most expressive out of the three, as sound is not limited by the physical morphology of the robot and has all the affordances that the human-voice has also: the potential to dual encode both semantic information and affective information through a single signal that does not require line of sight in order to be communicated, as well as increasing the distance through which communication between two agents can be established and take place. What is fascinating about this character is that throughout the films, not only did it not make a single utterance in a real natural language, the audience did not appear to *care* about this, they readily attributed very human-like states to the robot even though the main mode of communication had very little resemblance to how people communicate through sound. This is anthropomorphism and suspended disbelief in action.

Clearly there is also something special about NLUs as there is with the human voice and natural language, however, as we shall see in this thesis, there has been very little scientific and documented investigation into this different mode of expression in general, and the vast amount of potential that it has to offer the world of real social HRI. For example, little is known whether people indeed perceive NLUs made by a real robot has having any affective meaning. Furthermore, in the Star Wars films everything that R2D2 *said* is heavily scaffolded by the other events occurring around the robot. Is this a trick that fundamentally underpins the meaning the people project onto these abstract sounds it made, or is there something about the utterances that naturally elicits particular affective inferences in people? These are the kinds of questions that this thesis seeks to address.

Now we have identified the types of utterances that this thesis is concerned with, we shall take a step back and consider what vocal expression and communication affords a social agent more generally.

1.3 Vocal Communication and Expression

Expression through sound is a capability shared by many different species and has a number of affordances that sets it aside from other modes of communication and expression, many of which have been exploited by intelligent animals, particularly those that group together (though ants are a notable exception to this). At the group level, vocalisations are commonly used as a means for a member of the group to identify, gain the attention of, and communicate with other members with regard to the state of the environment, such as warnings to announce the presence of a predatory threat, the availability of food and group herding. At the individual member level, vocalisations allow an animal to express its inner state, such as its sexual availability, its dominance over another animal during confrontations, and its general state of health (whether it is in pain, distress or is ill). However, not all vocalisations have a communicative role. There are species that exploit the properties of acoustics to gain an understanding of their surrounding environment, such as Bats and Dolphins which use sound as a means to perform echolocation to locate and map the physical surrounding environment as well as prey.

What sets vocalisations apart from other modes of communication is that while they are relatively energetically costly, acoustic signals do not require line of sight which allows communication to occur over large distances and through a variety of different mediums such as fluids (in the case of marine mammals), solid materials (in the case of elephants) and through the atmosphere (birds, chimpanzees and humans are vivid examples). Furthermore, as we shall see with the human voice, sound provides a rich medium through which both simple and very complex meaning can be encoded and communicated, and as such has been subject to many efforts to recreate this in synthetic systems through the fields of Computer Speech and Speech Synthesis.

1.3.1 The Human Voice

Arguably, out of all the species that use vocalisations, humans have exploited the affordances of sound the most through the emergence of natural language that is in part facilitated via the high degree of control that humans have over their vocal system. In this regard, humans have developed a sophisticated communication system in which both a reflection of a persons internal (affective) state and language can be encoded (and decoded by others) into the same acoustic signal. This results in vocalisations that hold a very rich amount of information and social cues which can be communicated with ease and at speed, making for a highly effective means of communication and expression (Scherer, 1995, 2003).

Specifically with respect to affective communication, there has been a considerable amount of research over the last five decades that has tried to disentangle language from emotion in the voice in order to understand the nature of emotional expression in human vocalisations (see Scherer (1986); Banse and Scherer (1996); Scherer (2003); Juslin and Laukka (2003) for extensive reviews on this). With respect to creating NLUs that allow a robot to make vocal displays of affect, this is useful work to take insights and inspiration from.

1.3.2 Speech Synthesis

An area heavily related to HRI is speech synthesis, which has tasked itself with the challenge of using technological developments to understand, and artificially re-synthesise the human voice and create artificial agents that can converse naturally with people. Such technology has a broad range of applications both in the medial domain, but also in the field of psychology as a tool to aid the scientific exploration of the human voice and the acoustic correlates of emotional speech, but also to the field of computer technology and robotics in general. Like human voice research, speech synthesis too has a long and rich history and the two fields have arguably developed in unison in recent years given their shared underlying goal of fundamentally understanding the human voice. As with the field of psychology and the study of the human voice, this is a useful field to drawn knowledge from

when developing methods that allow a robot to communicate vocally with others.

1.4 The thesis

Within this thesis, the fields of study outlined above come together. Affective displays are important roles during social interaction and the ability to make affective displays is an important capability required in order to make socially competent robots. The world of animation has shown us that robots do not need to speak in a real natural language, rather it has shown us that the use of abstract sounds is a rich and fruitful means of animating a robot and bringing it to life through vibrant, expressive displays. Furthermore, while natural language is identified as a key ingredient required to make social robots interacting and engaging, the current state of the art in Natural Language Processing (NLP) technology has some hindering limitations, not only with respect to an actual interaction, but also on other research within HRI that has a certain degree of reliance on vocal interaction and expression. Non-Linguistic Utterances have a number of properties that make them appealing to HRI and can potentially mitigate some of the problems accosted with NLP, as well as providing a rich source of expressive vocal displays. However, it is currently unclear as to whether the rich affordances that are provided by NLUs in fictitious robotic characters can be achieved and applied to real world social robots in the same way, and what may influence this.

Now that the backdrop has been presented, and the stage is set, the following general questions may be formulated:

- *Do people perceive robotic NLUs as having different affective meanings, and if so, are people coherent in meaning that they perceive?*
- *Can NLUs be generated, and affectively charged in an automated manner to evoke a desired affective interpretation?*
- *What factors impact how a robot uses NLUs and how people interpret them?*

These questions are addressed as follows through this thesis. Firstly, the term *Non-Linguistic Utterances* is given a more formal definition, and a review of the

previous work in the literature is presented. Also, related work regarding the expression of emotion through the human voice and through musical pieces is reviewed as this likely has some useful insights regarding how different acoustic signals may be characterised and *charged* to convey different affective meanings (chapter 2).

Next, a review of the experimental tools that have been used in this research are presented and detailed. Firstly, using the insights of the human voice and music, a custom NLU generator is described in detail as this forms the backbone scientific tool that is used throughout this thesis to systematically study NLUs (chapter 3). This parameterised approach to NLU generation is vital to the goal of understating how NLUs should be designed and how different acoustic features relate to different affective interpretations in people. This is followed by a detailed description of the Aldebaran Nao humanoid robot and the manner in which it has been used as the sole platform through which NLUs have been embodied in a multi-modal social agent. Finally, after a review of issues surrounding the two main approaches to representing affect (affective categories and affective dimensions) theoretically and in synthetic systems, different tools designed for explicitly capturing affective ratings from humans, the affective measuring tool of choice - *The AffectButton* - is detailed. These are all issues that hold great relevance when it comes to the study of affect in general, and specifically here, through sounds and vocalisations.

As the Nao is the sole platform used in this research, it is important to probe whether it is an appropriate platform (in the eyes of people) through which NLUs may be studied, whether the use of a different robot may impact how people perceive NLUs, and in turn how this may potentially impact and limit the conclusions and contributions of this thesis. This is the focus of chapter 4, and the results provide strong support for embodying NLUs in the Nao platform.

Having developed a means to create and systematically manipulate the various features of an NLU, it is necessary to perform an exploration into how the different parameters impact how people perceive the affective meaning of an utterance, and

into understanding the the nature and dynamics of peoples' perception of affect in NLUs. This is done in chapters 5 and 6. These explorations provide data that reveal the underlying relationships between the different parameters of an utterance and how these relate to different affective interpretations and provide the basis for a mapping between these to be determined. This data the then used to train Artificial Neural Networks to learn these mappings, where a desired affective meaning is input to the networks and a specification of the utterance parameters is output (chapter 7). These mappings are compared with the general findings relating to the human voice and music to check for similarities, followed by a subject evaluation of the mappings that tests whether people are indeed sensitive to these mappings and associate utterances with different acoustic features with different affective meanings.

Having established the impact that a robot's physical design has upon how people perceive NLUs, and how the different acoustic features of an utterance influence the affective meaning that is conveyed, the next major factor to investigate is how the particular situation and context in which NLUs are used by the robot influences how the NLUs themselves are perceived and interpreted by people and whether general coherence between people is increased (chapter 8). Chapter 9 then extends this notion of situational context further, beyond physical interactions with a robot, to contexts that are set through verbal interactions that include natural language. In this chapter the notion of a robot using NLUs alongside natural language is explored as natural language is a rich source of defining the context and mood of an interaction, and presents a large range of potential research directions.

1.5 Contributions

In light of the description of relevant topics described and thesis above, what the general research questions are, and how they have been addressed, the original contributions that have emerged from this are summarised as follows:

- The design and development of a new and novel means of creating and

characterising NLUs that can be used to generate and systematically explore the acoustic characteristics of NLUs beyond single tones with either a rising or falling pitch envelope (a.k.a. “earcons”).

- Not all robotic platforms are compatible with NLUs. There is an alignment that must be made between the physical design of a robot and the acoustic behaviour it exhibits for people to deem this combination as acceptable.
- People show little coherence in their perception of affective meaning of utterances when they are presented in a context-free manner.
- NLUs that have similar acoustic features as the human voice and music do when expressing an emotion do not evoke the same affective interpretation. The acoustic correlates of the human voice and music in emotional expression do not have the same effect when applied to NLUs.
- People exhibit Categorical Perception when affectively interpreting NLUs as their affective interpretations are drawn to particular, basic, emotional states.
- Coherence in affective interpretations of NLUs emerges from, and is directed by the situational context which the utterance is used within.
- During an interaction NLUs, people show preference for NLUs being used along-side language rather than on their own.

1.6 Structure

The structure of this thesis is outlined below, giving a brief description of the theme and context for each chapter. Also, to accommodate for the hasty reader, chapters 3 through 9 begin with a list of the main key points and findings.

- In this introductory, *chapter 1*, the main relevant themes to this thesis have been introduced and their relation to each other highlighted, and contributions and structure of the thesis outlined.

- *Chapter 2* provides a deeper and more extensive background regarding Non-Linguistic Utterances. The similarities and differences between NLUs and natural language are discussed with the boundary between the two shown to be vague at best. Previous work in both NLUs and the related area of Gibberish Speech is covered in detail, as is the related literature on affective expression through the human voice and music is also reviewed as this potentially holds useful insights with respect to how NLUs can convey affective meaning.
- The methods that have been used in this thesis are detailed in *chapter 3*. It begins with a detailed description of a new algorithm designed and developed to facilitate real time generation and synthesis of NLUs as well as a high degree of precision in specifying and manipulating the acoustic properties. The algorithm described in this chapter serves as the means of producing, characterising and controlling the acoustic properties and features of the utterances used in all the experiments presented in the subsequent chapters of this Thesis (with the exception of chapter 4). A description of the Nao humanoid robot and the manner in which it has been used in this work follows this. Finally a discussion regarding the measurement of affect in humans is presented, and the affective measuring tool of choice and its use is detailed.
- Throughout the work described in this thesis, the Nao robotic platform has been used as the platform through which the NLUs have been embodied in a physical, social agent. *Chapter 4* presents an experiment that seeks to provide an experimental justification for this, and highlights the importance of the relationship between the physical appearance of an agent, and the (audible) behaviour that it exhibits, and how this can impact the holistic perception that people have of the agent, particularly in the case of NLUs.
- In *chapter 5*, the parameters of the NLU generation algorithm are explored in a systematic manner through a series of small experiments in order to test

their impact upon the affective meaning of an utterance. These experiments are designed in such a way as to also accommodate the collection of training data to be used in chapter 7.

- *Chapter 6* presents two experiments centred around investigating whether both adults and children exhibit categorical perception when affectively interpreting NLUs. It is well established that people categorise a wide variety of sensory stimulus, such as colours, facial expressions and emotional speech, and chapter 5 presents evidence that suggests that the same may be true for NLUs also. Using the methodologies that have been refined and well matured in the domain of psychology, this chapter seeks to uncover whether it is indeed the case that subjects affective interpretations of NLUs are also subject to a perceptual magnet effect and drawn to particular prototypes.
- Using the data collected from the experiments in chapters 5 and 6, *chapter 7* details how this data has been used to train a collection of Artificial Neural Networks to learn a mapping between an dimensional representation of affect and the parameters of the generation algorithm outlined in chapter 3. These networks (and the learnt mappings) are then evaluated with young children.
- In *chapter 8*, the interaction between the situational context which NLUs are used within and the subsequent affective interpretation of these utterances is investigated. More specifically, this chapter queries whether the context has a biasing effect whereby the nature of the context within which utterances are used directs how they are subsequently interpreted, or conversely, whether the use of NLUs can bias how the context is interpreted.
- Taking the findings of chapter 8 - that situational context biases affective interpretation of NLUs - into consideration, *chapter 9* explores the potential use of NLUs along side natural spoken language rather than being used as an alternative through an online experiment as natural language is another rich source of situational context.
- *Chapter 10* provides a summary overview of the work that has been pre-

sented in this thesis, and reflects upon the aspects that are related to the limitations of the thesis, as well as in the broader sense and ends with a discussion of a collection of topics that are considered as potentially fruitful future research.

Chapter 2

Non-Linguistic Utterances

This chapter serves to sketch a theoretical and practical background of Non-Linguistic Utterances (NLUs). It begins with a brief definition and formalisation of what NLUs are, and are not, and what distinguishes this, particularly with respect to natural language, as well as their relation to a similar strand of research surrounding the use of gibberish speech in social agents. This is followed with some examples of how NLUs have been used in real robots, as these help provide more tangible and concrete examples of the type of utterances that are the focus of this research. Following this, the general motivations and potential applications of NLUs (and gibberish speech) to social HRI are then outlined.

A review of research on emotional expression through the human voice and music is then presented, drawing particularly from the fields of psychology and musicology, as facets of these fields have had great influence upon the the previous work in NLUs and gibberish. Furthermore, in this review, certain links between methods developed to facilitate the study of emotional expression through sounds and the methods used to create NLUs and gibberish speech are highlighted, as many of these have been overlooked in the previous works.

Following this, a review of the previous work on NLUs and gibberish speech is presented in tandem, charting the developments that have already been made. This work is then discussed and important gaps in the research are highlighted, as these have influenced the manner in which the work informing this thesis has been conducted.

Finally, a note on the properties of NLUs that ultimately distinguish them from language is presented, as this justifies why NLUs and gibberish are not considered to be an artificial language with respect to this thesis.

2.1 A working definition of Non-Linguistic Utterances

Non-Linguistic Utterances (NLUs) are sounds comprised of beeps, squeaks and whirrs rather than resembling a real spoken language. They are utterances that are specifically designed not to resemble the complex acoustic signals that can be made by the human vocal system, and are not designed to resemble any *real* natural language, and thus are inherently unable to convey complex, linguistic semantic information to humans who do speak real natural languages. However, NLUs are still theoretically able to convey affective information as this is not directly dependant upon a shared linguistic and semantic vocabulary between two people or agents, but rather can be *encoded* and *decoded* through more general features and characteristics in both simple and complex acoustic signals. This highlights an important distinguishing feature between natural language and NLUs: while the human voice affords the dual encoding of both affect and linguistic information via the same acoustic channel (Picard, 1997; Scherer, 2003), NLUs do not as there is not (intended to be) a defined linguistic vocabulary or encoding/decoding protocol (at this point in time).

At this stage, a similar, related method of expression should also be introduced, *gibberish speech*. As we shall see later, gibberish speech has the same underlying motivations as NLUs, as well as the same utilities with respect to their application to HRI. However, there is one fundamental difference, and that is that unlike NLUs, gibberish speech is indeed designed to resemble the timbre and voice quality of human speech, without containing any linguistic or semantic content. The reason for mentioning this seeming different modality is that NLUs and gibberish speech represent “two sides of the same coin”, so to speak. And as such, it is

useful to draw upon work relating to both in order to outline the shared underlying qualities and applications, as well as the motivations for these.

Finally, a note on why this thesis is about *affective* vocal expression through NLUs. As we shall see later, NLUs currently do not constitute a language, and as such, their use in robotic systems is not to try and communicate high level, complex meaning and information, as this is fundamentally not possible at this stage in time due the lack of established cultural norms regarding the use of NLUs during social interaction. However, NLUs do have the capacity to provide rich paralinguistic cues and convey affect. Setting language aside, affective expression is well established as being a fundamental ingredient required for facilitating and regulating engaging and quality social interaction (Breazeal, 2001b,a, 2002, 2003b,a; Belpaeme et al., 2012), hence this is why the body of research presented in this thesis focuses upon affective expression through the use of NLUs.

2.1.1 Examples of NLUs in Popular Culture

NLUs have been used almost exclusively to great effect within the world of Animation as means of bringing inanimate objects, particularly robots, to life and allowing them to be portrayed as social agents/individuals who can interact with social peers with ease, and without the need to use a *real* spoken language¹. Rather, in this respect, NLUs have been portrayed as a fictional language, where other characters within the films are able to understand what the robots are saying, while the audience in reality do not. Their understanding of what has been said by the robot is highly scaffolded by the events that occur within the rest of the scene, and the script of the other characters. As such, it can be viewed that the actual sounds themselves have little meaning to the audience, when used on their own, and the meaning is deduced by the other salient cues provided (with many of these cues being specifically tailored toward helping decode the utterances made by the robot). However, this is something that needs clarification.

As a result of the success of the Star Wars franchise in particular, NLUs have

¹Fictional robots such as R2D2 from the Star Wars films, and WALL-E and Eve from the Pixar film WALL-E provided vivid examples of this.

gained a certain *iconic* status in popular culture and media, in that they are now synonymous with the fictional robot R2D2 (and similar robots within the films), and how it expressed a rich variety of socially relevant cues (such as affect, humour, and logic) through a variety of beeps, squeaks and whirrs, which did not resemble vocalisations made by the human voice.

Given this status alone, using NLUs in real robotic systems can be seen as an appealing alternative to having a robot that speaks with a real natural language, as the popularity and iconic status of robots such as R2D2 (at least in the developed world) means that people who see a real robot using NLUs are likely to perceive the utterances as expressive displays. By using the association between NLUs and social capable robots such as R2D2, a roboticist can increase the likelihood that a real robot making similar sounds will be perceived as socially capable also, and that the utterances indeed have a social meaning and utility, or so the theory goes.

2.1.2 Examples of NLUs in Real Robots

While Animation has been the main beneficiary of NLUs overall, there is also a growing number of examples of both research and commercial systems that employ(ed) NLUs as a means of expressive displays. For example, Keepon (Kozima et al., 2009), and the commercial sister robot, My Keepon, have both used small database of simple sounds that are used to provide expressive vocalisations both in response to sensory input, and as a means of attracting attention. Similarly, WowWee’s RoboQuad (and various other robot toys in their robot product line) also used a small collection NLUs for reactive behaviours to sensory input, as well as commands input by the user through an Infra-Red remote control.

There are some stark differences, however, between the use of NLUs in Animation and in real world robotic systems. Primarily, robots such as R2D2 are not subject to the same limitations that real robots are. For example, while R2D2 is a *robot*, it does not have a real computer with limited memory and processing resources, or a similar target production line cost for that matter. As such, R2D2

is not limited in the variety of different utterances that it can make, while robots such as Keepon are, given the computational resources within which a functional system had to be created. Furthermore, utterances have to be carefully designed by hand by a sound engineer, which is a time consuming process in itself. This is why real robots have tended to have a limited repertoire of utterances, something that is easily spotted by the consumer when interacting with the robot.

2.2 Motivations and Applications in HRI

While it is well established that Natural Language Interaction (NLI) plays a vital role in modulating and enhancing the quality of social interactions between robots and people (Belpaeme et al., 2012), the current state-of-the-art of Natural Language Processing (NLP) still suffers from notable short-comings that can significantly impact HRI in adverse manners (Mubin et al., 2009). For example, while Automatic Speech Recognition (ASR) has gradually become a reasonably robust and common place technology², facets of NLP such as Natural Language Understanding, Dialogue Management and Natural Language Generation still remain challenging tasks. Furthermore, given the serial “pipeline” nature of NLP, there is very little room for error, and when errors do occur they quickly propagate and often lead to breakdowns in NLI, such as incorrect, or worse, no responses from the system, both of which are uncomfortable for users (Shiwa et al., 2009; Lee et al., 2010). This makes facilitating natural language in current robots a challenging and cumbersome task.

Strategies stemming from NLP for coping in situations where NLP might fail include constraining and scripting interactions and dialogues, narrowing the scope of user responses (e.g. Lohse et al. (2008b)), or employing a set of general purpose responses to try and catch the failing interaction (e.g. Lison and Kruiff (2009)). These strategies do have their limitations and inherent risks however, as incorrect

²Robust and common place, in this case, refers to the fact that ASR technology is now a common feature on most smart phones and tablets, and in some cases, games consoles. However, noisy environments, multiple speakers and speaker variation such as accents and age (to name a few), still pose problems.

or repetitive linguistic responses are quickly identified by users, often revealing the limitations of the system (Ros Espinoza et al., 2011). Such revelations tend to hamper the development of long-term, open-ended HRI, which is a long-term goal of the field (Belpaeme et al., 2012). In such situations it may be appealing to be able to *disguise* the limitations of the system from the user in some way to mitigate the overall negative effect, and if the problems extend to such a degree that NLI is no longer possible, perhaps to revert to a replacement modality, such that interaction can continue, albeit in a limited capacity. NLUs can potentially provide a solution in both cases. However, as we shall see in chapter 9, this potential use is built upon some fundamental assumptions regarding the use of NLUs with natural language, and the validity of these assumptions needs to be tested and confirmed.

While the short-comings in comparison to natural language are obvious, NLUs do have qualities that hold promise for HRI however. For example, utterances are not bound to a particular spoken dialect, thus their use in multi-lingual and cultural settings may be advantageous. Secondly, given that NLUs hold/communicate little semantic content, there is generally a lower need to process semantic information from user speech, thus settings that pose challenges for sensory equipment and technologies such as microphones and NLP can be considered less problematic³. Furthermore, less parsing and processing of semantic content results in fewer delays in agent response times, helping bring the interaction closer to real time and aiding the fluidity of the vocal exchanges which has been shown to be crucial for HRI (Shiwa et al., 2009). Finally, as NLUs are generally considered to hold less semantic content (with less need for a robotic system to consider semantic content), the burden of interpretation lies with the user, the *intelligent other*, with their inherent understanding of situational context and natural tendency to anthropomorphise inanimate objects such as robots (Duffy, 2003), and treat them as socially competent (Reeves and Nass, 1996). Given this, the presence of an intelligent other may also be exploited to allow utterances to be used in far

³Such settings tend to be in *real world* environments that are far from the protected and “safe” laboratory environments.

less restricted scenarios, widening the range of potential application areas, where the person may project meaning onto the abstract sounds based upon how the interaction is unfolding. This notion is explored in chapter 8.

NLUs also have another, subtle, but powerful potential affordance - the ability to allow robotic designers to subtly manage user expectations. It is a common observation in HRI that as the sophistication of a robotic system increases, so too does the user's expectations of the system, and thus the greater the risk that they discover the system's limitations and disengage from the interaction (Ros Espinoza et al., 2011). This however can be circumvented through *expectation setting* where both information about a robot's capabilities (e.g. vision, tactile sensing, speech recognition, etc.) and observable behaviour (e.g. reactive behaviour to input stimulus, and expressive displays, etc.) can be used as a tool to set user expectations (Paepcke and Takayama, 2010). In theory, by employing NLUs rather than Natural Language, the robot designer is able to help keep the "bar" of expectation low by producing a robot that does not risk engaging in open-ended NLI but can remain responsive to external stimuli and make expressive displays and engage in open-ended HRI. Again, gibberish speech provides a good tangible example of this through Kismet (Breazeal, 2002), where people were observed to readily engage in stimulating multi-modal interactions with the robot without the need to rely on natural language interaction. It is also worth noting that in such design philosophies, the ability for naive subjects to suspend disbelief (Duffy and Zawieska, 2012) is also used as a powerful tool, and is a aspect that could make the use of NLUs particularly useful during Child-Robot Interaction as children are observed to readily suspend disbelief and are very willing to engage in social interactions with robots (Robins et al., 2004; Belpaeme et al., 2012, 2013).

There are already a number of robotic systems, both fictional and non-fictional, that demonstrate how a variety of the qualities of NLUs can be applied to social robots. Moreover, there are also examples that stem from HRI research showing that not only are NLUs and gibberish a useful means of facilitating expressive vocal displays, but they also hold potential as a useful tool that can help advance

and support research into other areas of HRI in general. These are outlined here.

2.2.1 Utility as a tool in broader HRI research

As is shown later in this chapter, both NLUs and gibberish speech have the potential to be used beyond a tool for creating and animating expressive robots, but also as a tool for studying affective expression through sound and speech more generally. However, this section serves to point out that NLUs and gibberish speech have properties that make them very appealing as tools to be used in other areas of HRI also. The Kismet robot (Breazeal, 2002) is prominent example of how gibberish speech can be used as a means of vocal expression in a robot, but at the same time is used as a tool to help facilitate research into other areas of HRI simultaneously, such as evaluating the influence that affective models of the robot’s internal states can have on the observable behaviour of the robot.

For example, Chao and Thomaz (2013) have used gibberish speech in a similar manner with their robot, Simon. In this work, the focus of the research was on evaluating their computational model turn taking during multi-modal HRI. Their evaluation required subjects to interact with Simon, in a natural manner, and so they told subjects to *teach* the robot about a variety of different objects⁴. As the focus of the work was on turn taking, the robot was required to engage in the interaction and make both visual gestures and audible vocalisations. In order to avoid having to implement an NLP system, which if it failed could have had adverse consequences on interactions, they implemented a gibberish speech system in the robot, in the same way as (Breazeal, 2002) and for the same reasons - to elicit natural behaviour and turn taking from the human, without the need to cater for increased complexity and risks that come with NLP and the use of natural language.

In research focused upon the physical, anthropomorphic design of robots and how this impacts the perception people have of a robot, Walters et al. (2007) used NLUs to facilitate vocal animation of a robot that was deemed to be “machine-

⁴In reality, the robot did not do any learning. By asking subjects to teach the robot about objects, they were subconsciously encouraged to behave and interact in a natural manner.

like”, as opposed to having a more anthropomorphic design. Again, in this example, NLUs have been used as a tool to facilitate vocal animation of a robot, in order to be able to study aspects of HRI that fall far beyond affective expression via sound. Another example of this use of NLUs is research in with the robot Keepon (Kozima et al., 2009), which is a robotic tool designed to be used with young autistic children, many of which are pre-verbal. In this cases, not only does the robot’s morphology not lead itself to the use of gibberish speech, but the use of natural language with pre-verbal infants serves little purpose and runs the risk of over complicating interactions.

In these examples, the benefits of NLUs and gibberish speech shines through clearly. The use of natural language in robots is currently cumbersome due to the limitations that the technology has, and if the research does not strictly require natural language, but does require some form of vocal expression, NLUs and gibberish can be seen as attractive options that can be implemented with considerable ease in comparison to NLP. Furthermore, the examples above are only a select few which have actually not used natural language for vocal expression when they have not needed to. There are many examples of research experiments that have adopted the *Wizard of Oz* (WoZ) experimental method (Kelley, 1984; Riek, 2012), where there is (unknown to the subject) a human controlling aspects of the robot. Moreover, in the majority, the reason why WoZ has been used is to facilitate vocal expression and natural language⁵, which highlights two points: firstly that NLP technology is not in a mature enough state where it can be implemented into robotic systems for state-of-the-art research, and secondly, that there is a growing body of HRI research that is becoming contingent upon vocal communication and natural language in order to progress, and WoZ is used as a means to circumvent this contingency. The particular problem with the latter point is that the robot systems that are ultimately used in this research are not *autonomous* systems, but rather are mock-ups. This in itself can be a limiting factor in the general progress toward creating fully autonomous social robots.

⁵A recent review by Riek (2012) found that approximately 70% of WoZ studies used this technique to facilitate vocal and natural language.

It is in this light that the use of NLUs and gibberish speech in HRI research can be considered as highly fruitful as it provides a means of creating vocally expressive robots that are autonomous and thus can be programmed to operate in a consistent manner, making their use in experimentation particularly useful, and they remove any bias that natural language may have. Something that is useful when exploring other modalities in a robot.

2.3 Affective Expression through Sound

As both NLUs and gibberish speech are intended to resemble the function of human vocalisations during social interaction, in the sense that they can be used for *proto-dialogues* and affective displays (while not necessarily resembling human speech with respect to voice quality or timbre), it makes sense to draw upon the insights that have been gained regarding emotional expression through sound, and in particular through the human voice. This section serves to do this and presents an overview of affective expression through both human speech and through music, as both have relation to how affective meaning can be projected through NLUs and gibberish speech. Furthermore, this section also seeks to highlight the relationships between the methods used to create stimuli to investigate the acoustic correlates of human speech, and the methods used to create NLUs and gibberish speech, and many of these relations and similarities go overlooked.

Finally, the section presents a brief overview of the psychological phenomenon of *sound symbolism*, which theoretically may also hold potentially useful insights with respect to the design and synthesis of NLUs and gibberish speech. However, much of this particular notion is speculative and thus is not dwelled upon.

2.3.1 Affective Expression via the Human Voice

As mentioned in the introduction, research into the acoustic correlates of the human voice with different emotional states has received a great deal of attention for a number of years, and has been spurred on by technological advances such as the telephones, audio recording technology (Scherer, 2003), and more recently,

advances in computer speech/speech synthesis (Schröder et al., 2010). Through these efforts, a vast body of work has accumulated, and as a result a number of review articles have been published in order to chart and summarise the findings (e.g. Scherer (1986), Banse and Scherer (1996), and Scherer (2003)). In light of this, this section will not provide a detailed coverage of the whole field of human emotional speech research (this is covered in adequate detail in the review articles), rather it outlines the main important aspects of the field that apply to the study of NLUs and gibberish speech.

Firstly, it must be pointed out that in comparison to NLUs consisting of beeps and squeaks, human vocal expressions are very complex acoustic signals, particularly as they exploit the affordances of sound in order to dual encode both affect and natural language - where *what* is said and *how* it is said are transmitted via the same channel at the same time, in the same signal (Picard, 1997; Scherer, 1986, 2003). This makes the study of the human voice and specifically emotional speech particularly challenging. The first problem is to try and isolate these two components of the signal such that their underlying acoustic characteristics can be studied. In order to achieve this a number of novel techniques have been developed in order to address this, varying from methods of masking linguistic context, to artificially creating signals that have no linguistic context at all. Here, it is worth pointing out that this sounds familiar: both NLUs and gibberish speech have a similar underlying goal, and as we shall see later in this chapter, previous literature has used very similar techniques to create both NLUs and gibberish speech, while perhaps not being completely aware of this. Using these techniques, it has been possible to explore the acoustic correlates of emotional expressive via the human voice and via music. This is presented in section 2.3.1.2.

2.3.1.1 Removing linguistic context from the human voice.

There have generally been two approaches to this problem of removing the distortion of natural language from expressive speech: *cue masking*, and *cue manipulation* by re-synthesis (Banse and Scherer, 1996; Scherer, 2003).

In cue masking approaches, the verbal cues are masked, distorted/corrupted or removed from expressive vocalisations that have been captured from humans (either via eliciting natural emotional expressions in people, or by recording actor portrayals) to study the influence of the ensuing acoustic features on peoples' inferred emotional meaning and content (Scherer, 1986). This particular approach has been used early on in the field, using techniques such as *low pass filtering* in order to remove the higher frequency components of a voice sample in order to suppress the intelligibility of phonemes (e.g. Knoll et al. (2009)), and *randomised splicing*, where voice recordings are split up into small segments and reordered in such a manner that the prosodic features of the utterance are generally retained, while the verbal cues are distorted and the verbal content corrupted (e.g, Scherer (1971) and Scherer et al. (1972)). Remez et al. (1981) used *Sine Wave Synthesis* to investigate the nature of speech perception. This technique involves analysing the voice recordings and generating time-varying sinusoidal wave patters that match the time-varying patters of the vocal formants of the voice.

While the benefits of the cue masking approach are that they are using actual expressive human speech, which ensures as high degree of voice quality and accuracy, there are problems surrounding the methods through which the voice recordings have been captured. In particular when voice actors are employed, there is a risk that when they are asked to portray an emotion, they exaggerate this and thus the voice recording does not necessarily comes and accurate reflection of genuine emotional speech (Scherer, 2003). Also, it has been reported that people do still exhibit an ability to recognise and understand to a degree the verbal context of the speech, which demonstrates the degree to which affective and verbal content are intertwined in the voice (Remez et al., 1981; Scherer et al., 1972) .

Cue manipulation via re-synthesis is a more modern approach has been proven to be a remarkably useful tool (Cowie and Cornelius, 2003), particularly given the developments in general speech synthesis technology. Through this technology, the human voice can be explicitly parameterised which allows for systematic manip-

ulation of the vocal patters and parameters and how peoples' affective inferences change as a result (Scherer, 2003). An early example of this, before the large scale developments of speech synthesisers, comes from Scherer and Oshinsky (1977), who used a MOOG synthesiser to create concatenated tones of sounds that were designed to resemble both sentence-like utterances as well as musical melodies, by specifically manipulating the pitch, rhythm, contour, timbre and tempo of tones. More recently, the use of speech synthesisers has become popular as reflected by the large number of publications on the subject (e.g. Cahn (1990), Murray and Arnott (1993), Murray and Arnott (1996), Burkhardt and Sendlmeier (2000), Laukka (2005), Schröder (2001), Schröder (2003a) and Schröder et al. (2010)), partly due to the direct application that findings have for speech technology applications, of which there are many.

The purpose of highlighting these two methods of creating stimulus with affective content for psychological studies is that there are a very many number of parallels between both NLUs and gibberish speech, with respect to the underlying goals, but also the techniques that are used to actually produce utterances and stimuli. This is something that the related literature in both NLUs and gibberish speech has failed to observe⁶. Furthermore, this emphasises the strong relationship and relevance between the human voice, speech synthesis and NLUs/gibberish speech, and highlights that the use of NLUs and gibberish speech does not only need to be geared toward the application in social HRI, but the methods used to create utterances can also have utility as scientific tools that can be used to help further address research questions regarding emotional expression in the human voice.

2.3.1.2 Acoustic Correlates of Emotional Speech

Work investigating the acoustic correlates of emotional speech have tended to focus on a relatively small number of vocal cues given the complexity of the

⁶It can be argued that the true roots of NLUs and gibberish speech lay in psychology, and the only the area of application, HRI, is now different and new. It may be due to the difference in age of the respective fields of psychology (old) and HRI (young) that authors have not observed the strong links between the two fields with respect to the methods used to create stimuli/utterances.

human voice as an acoustic signal, and who they change across different basic emotional categories (Juslin and Scherer, 2005). Moreover, as there is a very large body of research that addresses this, much of which reports different findings that sometimes conflict, it is difficult to consolidate the results of these studies into a coherent overview of how these different parameters vary across the different emotional states. This is where invaluable review efforts come into their own (e.g. Scherer (1986), Banse and Scherer (1996), Scherer (2003), Juslin and Laukka (2003)). Drawing upon these review articles, this section serves to provide an overview of the different vocal cues that have been studied, and how they vary. These parameters are taken into consideration in the next chapter which outlines a custom method for characterising sentence-like NLUs.

Table 2.1 lists the main vocal cues that have been studied in the human voice, broadly speaking, providing a brief description of each. It can be seen that these different parameters are all commonly associated with different general properties of the voice, namely the *pitch*, *intensity*, *temporal aspects* and *voice quality*. It is the changes in both the the properties of pitch and the temporal aspects that translate to changes in prosody which is an general umbrella term for referring to the dynamics of the acoustic signal over time. With respect to NLUs and gibberish speech, all of these parameters hold relevance as they provide high level ways of characterising utterances, and as we shall see later in this chapter, many of these vocal cues are been used when creating affectively charged utterances.

Table 2.1: Description of the Acoustic Cues in Vocal Expression. Table adapted from Juslin and Laukka (2003).

Acoustic Cues			Perceived	Corre-	Description
			late		
Pitch	Fundamental	Fre-	Pitch		F0 represents the rate at which the vocal chords oscillate. Acoustically, the F0 is the lowest periodic cycle component of the waveform
	quency (F0)				
	F0 Contour		Intonation contour		The F0 contour is the sequence of F0 values across an utterance over time. Besides changes in pitch, the F0 contour also contains temporal information, and as such is difficult to operationalise.
	Jitter		Pitch Perturbations		Jitter is the small scale perturbations in the F0 related to random vibrations of the vocal chords.
Intensity	Intensity		Loudness of speech		Intensity is the measure of acoustic energy in the acoustic signal, and reflects the amount of effort required to produce an utterance. It is usually measured as the amplitude of the acoustic signal.
	Attack		Rapidity of voice onsets		The attack of a signal refers to the rate of the rise in the amplitude of the voiced segments of an utterance.
Temporal Aspects	Speech Rate		Velocity of speech		The rate can be measured as the overall duration of an utterance, or as units per duration. It can either include only the voiced segments of speech, or the entire utterance as a whole.
	Pauses		Amount of silence in speech		Pauses are usually measured as the number or duration of silences in the acoustic waveform.
Voice Quality	High Frequency	Energy	Voice quality		High frequency energy refers to the relative proportion of total acoustic energy above a certain threshold. As the energy in the spectrum increases, the voice sounds more sharp and less soft.
	Formant	Frequen-	Voice quality		These are the frequency regions in which the amplitude of acoustic energy is high, reflecting the natural resonances in the vocal tract. The first two or three formants largely determine the quality of vowel pronunciation, while higher formants are usually speaker dependent.
	cies				

With respect to how these voice cues change across the expression of different emotions, Scherer (2003) has attempted to provide a rough characterisation for the main vocal cues across the basic emotions as based upon the general findings reported in the literature. These are shown in table 2.2. It can be seen from the table that not all the voice cues have a characterisation for the different emotions. This is because not all studies focus on the same emotions, and many studies report contradictory and conflicting results (Scherer, 2003). Generally, it can be seen that high arousal states such as angry, fear, and joy are commonly associated with an increase in the F0 frequency, as well as the variability and range of this. In these states it is also commonly found that the speech and articulation rate are higher than lower arousal states such as sadness and boredom, as is the high frequency energy.

While this table is generally rather vague, it does serve as a good basic guideline for how different acoustic signals might be designed to convey different affective states, and particularly how the features of the signals covary. What the main drawback is that these are very general, while when it comes to implementing such insights into a system for creating synthetic utterances, many of the parameters characterising and utterance are system specific and so transfer of these broad characteristics of vocal cues to system specific parameters can be limited, particularly in the case of NLUs, which are designed to be abstract sounds rather than resembling human speech. Also, these characteristics identified by Scherer (2003) do not provide exact specifications for each of the voice cues with respect to their measured values. The reason for this is that each human voice is different, and so the exact parameter values differ greatly from person to person, and so also from experiment to experiment in the literature. However there appears to be more consistency in the dynamics of the human voice across the different emotional states than there is in raw parameter values, so this serves as a good initial start point by which to design and compare the dynamics of NLUs with, but it makes gauging the initial cue values (such as speech rate, pauses, F0 range, etc) difficult.

Table 2.2: Summary of the acoustic patterning of the human voice for the basic emotions. Table adapted from Scherer (2003).

Voice Property	Basic Emotion					
	Stress	Anger/rage	Fear/panic	Sadness	Joy/elation	Boredom
Intensity	↗	↗	↗	↘	↗	
F0 floor/mean	↗	↗	↗	↘	↗	
F0 variability		↗		↘	↗	
F0 range		↗	↗ (↘)	↘	↗	↘
Sentence contours		↘		↘		
High frequency energy		↗	↗	↘	↗	
Speech and articulation rate		↗	↗	↘	↗	↘

2.3.2 Affective Expression via Music

It has been theorised that affective expression in the human voice and through music share a common origin from an evolutionary perspective (Scherer, 1995), mainly with respect to the use of the voice (i.e. singing), but this can also be extend to the use of musical instruments (Juslin and Laukka, 2003). As such, it makes sense to touch upon the expression of affect through music also. As with expressive human speech, musical expression has also been explored for a number of years and as a result a large body of research has also accumulated. This work will also not be reviewed in detail as it draws away from the focus of this chapter, but readers are pointed to the extensive review by Juslin and Laukka (2003) which covers in detail affective displays in human speech and music the the similarities/differences that have been identified.

However, to provide a brief overview of their findings, after reviewing a substantial volume of studies on both the human voice and music they found that there are indeed a great number of similarities between the acoustic cues in music and in the human voice when it comes to conveying a particular affective state. For example, when conveying anger, characteristics of a musical piece are found to have a fast rate, have a high intensity with a great deal of variability in this intensity, a high overall pitch with a high variability and fast onsets of notes. Similar characteristics are also found for the expression of happiness, while sadness was associated with a slow overall tempo, a lower pitch with less variability and less overall intensity in the acoustic signal and with less aggressive onset of notes. This is generally consistent with the findings in the human voice also.

With regard to the notion of whether the human voice and music share a

common origin with respect to emotional expression, they conclude with the view that expression of affect through music is likely based round the manner through which this is done in the human voice. As such, when considering the potential use of insights gain from both the fields as the application to creating NLUs and gibberish, it seems more rational to focus on the insights regarding the human voice, as the human voice is more similar to the type of utterances (i.e. sentence-like) that NLUs and gibberish are aspiring to replicate, though NLUs are not intended to have the same overall voice quality to timbre.

Perhaps a final comment that should be made does not actually regard how musical pieces should be affectively charged, but rather, how the pieces themselves are synthesised. More specifically, the technologies that have been developed in the domain of computer music as these can potentially provide a broad range of tools through which NLUs, rather than gibberish speech, can be created and synthesised.

2.3.3 Sound Symbolism

While affective expression through the voice is a well established phenomenon, there is also another curious aspect of the human voice that may hold a relevance to NLUs and gibberish speech, and so shall be briefly discussed. This is the controversial phenomenon of *Sound Symbolism* (Nuckolls, 1999). In essence, the basic notion is that people exhibit an almost universal tendency to make certain associations between sounds and some form of meaning, where the meaning is not culturally defined. One of the most classic cases is the association between a verbal label and a visual referent, which in language is assumed to be arbitrary. However this is not always the case as is shown by the “*bouba/kiki*” *phenomonon* (Maurer et al., 2006).

In this phenomena, it have been extensively found that people, both adults and children (Ozturk et al., 2013), from different cultures all appear to exhibit a strong association between the verbal label “bouba” and a variety of shapes that have the common feature of smooth edges, while the label “kiki” is associated

with shapes that have sharp, jagged edges. The notion being suggested is that there are some naturally biased mappings between objects and sounds, and that these mappings are not culturally determined, but rather may have an actual innateness in the brain.

This phenomena is unusual and still is the case of debate and continued empirical study in the field of psychology as it draws out certain questions regarding the innateness of mappings between visual and auditory perception (Nuckolls, 1999). What makes this phenomenon related to NLUs and gibberish speech, though only very loosely, is the notion that certain types of acoustic features of an utterance may have a naturally biased mapping to having certain interpretations. For example, it may be that a NLU created using a saw-tooth carrier signal has a natural bias toward a particular affective interpretation in comparison to an NLU with a sine wave carrier signal. If such biased mappings do exist, than it is potentially very exploitable when it comes to trying to identify whether there are any NLUs or gibberish speech utterances that could evoke a particular affective interpretation in a universal manner. While this sounds appealing, it should be noted that this notion is somewhat unfounded as there does not seem to be any empirical evidence stemming from the field of psychology that supports the notion of sound symbolism and affect.

2.4 Communicating Affect through NLUs and gibberish speech: reviewing previous work

Previous work on NLUs as we shall now see, has been sparse, with only a few groups of authors seeking to explore and leverage the potential utility of this modality as applied to HCI and HRI. Work on gibberish speech, on the other hand, has received marginally more attention but ask remains a niche area of research also. This section serves to provide an overview of the research that has been conducted on both NLUs and gibberish speech and seeks to highlight the overall undeveloped state and knowledge gaps of this general area of research.

This section begins with the limited amount of previous work on NLUs and then addresses the work in gibberish speech. The different approaches that have been used and developed to create utterances are described in from a high level, as is the general focus and direction of the experiments that have been conducted prior to the research presented in this thesis.

2.4.1 NLUs

Given the abstract nature of robotic NLUs as communicative sounds, there have both been approaches to creating and affectively charging NLUs taking inspiration from research into the human voice, as well as more musically inspired approaches. For example, Jee et al. (2007, 2009) have adopted an approach to generating and affectively charging NLUs informed by the findings on affective expression through music (e.g. Juslin and Laukka (2003)). They used musical notation, theory and synthesisers to hand create a small collection of utterances that were designed to have a particular affective charge (Happy, Sad, Fear and Dislike), by varying the acoustic features of the tempo, key, pitch, melody, harmony, rhythm and volume. Subjects were then asked to perform three tasks. Firstly, listen to each of the utterances and rate how intensely the intended emotion was conveyed. Second, subjects were shown a cartoon face with an expression matching each of the labels, and again were asked to rate how well the face conveyed the desired emotion. Finally, Subjects were presented with both the face and utterance for a given label and were asked to rate the intensity of the conveyed emotion. Their results showed that both the utterances and facial expressions alone produced affective labels recognition rates between 60 - 70%, while when combined together, the recognition rates increased to approximately 85%. This shows that by combining the two modalities subjects see a more intense emotion than when the face and utterances were presented individually.

Jee et al. (2010) furthered this work by hand creating five sounds that were designed to convey particular intensions (Affirmation, Denial, Encouragement, Introduction, Question), and three emotions (Happy, Sad, Shyness), again using

musical theory and synthesisers to change the intonation, pitch range and timbre of the utterances. Subjects were then presented with each of the utterances and again asked to rate how intensely the utterances conveyed the desired emotion. Their results showed that generally, subjects did agree that the utterances said indeed convey the intended intension/emotion. While these efforts have employed a novel way of viewing NLUs, as musical *pieces*, the evaluations that have been performed provide limited insights and generalisation as only a small number of utterances were hand crafted based upon very specific observations NLUs made by robots in animation (namely, R2D2). Furthermore, only a single utterance was created for portraying each emotion and so the insights gained specifically apply to their unique NLUs alone. That said, what these examples do demonstrate is that using this music inspected approach also has potential for future research.

In a rare example of NLUs being studied and used outside of HRI to provide social cues, Tuuri et al. (2011) explored the use of NLUs in a sports watch as a means of providing feedback (telling them to slow down, urge them on, tell them that their performance was on par, and to provide a rewarding sound for good progress) to the wearer on their current performance. NLUs were created by taking human recordings of people making vocalisations to convey these four types of feedback and extracting the fundamental frequency of the voice, and mapping the prosody of these to MIDI notes in a music synthesiser. The results of their evaluations show that again people were able to accurately decode the meanings of the utterances presented to them. However, the shortcomings of this work are the same as that of Jee et al. (2010) in that only four NLUs were created, and so generalisation is limited. Rather, this serves as another example of how music technology does indeed have potential with regard to creating expressive NLUs for interactive agents and objects.

The largest overall body of work into NLUs with robotic systems has been conducted by Komatsu (2005); Komatsu and Yamada (2007, 2008); Komatsu et al. (2010, 2011); Komatsu and Kobayashi (2012). This body of work has had multiple, specific focuses but in general has sought to investigate how NLUs are

able to influence how people perceive and attribute states to the agents the make the utterances, and whether the agents using their utterances are able to change the way in which a person behaves when performing a task. In their body of work, utterances consisted of single, short and simple sine waves with either a rising, flat, or falling frequency modulation. Initial work focused on identifying how utterances should be designed in order to convey a notion of a *positive* or *negative* attitude, and agreement and disagreement (Komatsu, 2005). It was found that utterances with rising frequency modulations were commonly rated as positive or expressing agreement, while utterances with a falling frequency modulation were conveyed a negative attitude. These can be considered as very *iconic* sounds (e.g. *earcons*) as similar types of sound are commonly used in everyday technologies such as mobile phones, computer programs and even computer games, as a means to provide feedback on whether something positive or negative has happened. The drawback with these kinds of sounds is that they are very short a brief, something that could potentially limit their use during an interaction, where richer utterances with more variety may have more overall application and utility.

Komatsu and Yamada (2007, 2008); Komatsu et al. (2011) then investigated how different agent embodiments would impact how the same utterances were interpreted, recognising that embodiment and morphology may have a influence over how people infer agents' attitudes through NLUs. Utterances were embodied in a PC, an Aibo robot, and a mobile robot made of Lego. Subjects were presented with each of the three robots, and asked to rate how positive or negative they thought the utterances were. Their results showed that the when the utterances were made by the PC, people showed a high agree of accuracy in interpreting the utterances, while this was not the case when the utterances were made by the Aibo and Lego robots. More specifically, they found that subjects struggled to correctly identify the positive utterances as positive, while their identification of the negative utterances remained high.

In later work, Komatsu et al. (2010) then investigated whether these same utterances would be used by a robot to bias how a person performed a task. More

specifically, the setup involved having a subject play a treasure hunting game on a computer. The game showed a strait road, with hills appearing along the way. Under one of the hills a golden coin was hidden, and subjects had to guess under which one. Sitting next to the subjects was a Lego robot that was told subjects which hill the coin was under, and then made an utterance with either a flat pitch contour or a falling pitch contour as a means of indicating how confident the robot was in its predication. Their results show that when the robot's predication was accompanied by a utterances with a falling pitch contour, they rejected the predication significantly more than when an utterance with a flat pitch contour was used. In essence, the pitch modulation had a direct impact over the perception of how confidence the robot was about the information that it gave.

Extending this work into communicating the level of confidence that an agent has about information that it presents to people, Komatsu and Kobayashi (2012) conducted a further experiment to see whether the use of NLUs can mitigate the potential adverse effects that the presentation of incorrect information may have. Comparing NLUs and natural language, their results show that when the computer provided completely correct information, natural language was preferred over the use of NLUs. However, in situations where the agent's confidence in the information that is was providing was mis-judged, and thus the agent was shown to be confidence about information that was ultimately incorrect, NLUs were the preferred method of expression regarding the agents confidence. Their argument in this work is that currently computers and robots are not perfect - they make mistakes, and that when agents use natural language to communicate, this sets a high expectation level, and when this expectation is violated, this evokes an adverse reaction in people. This is a tangible example of how NLUs may be used to manage the expectations that people have of robots.

2.4.2 Gibberish speech

In comparison to NLUs, the volume of research that has been conducted is marginally larger, however while this is the case, both areas of research are rather

niche when compared to a field of research such as Speech Synthesis. This section provides an overview of the research that has been conducted into gibberish speech and the findings that have been obtained. As outlined earlier in this chapter, gibberish speech is a complementary approach to NLU in general, but is also a reflection regarding the general pre-occupation that the field of robots has with robots that speak like humans.

Perhaps the most famous and earliest examples of gibberish speech in a robot comes from the work by Breazeal (2001b, 2002) for the expressive vocal system on the robot Kismet. The work extended that of Cahn (1990), who had investigated how to produce expressive and affect laden speech using the commercial DECtalk synthesiser. This system worked by creating a pseudo-random text string using a custom algorithm and using this as the input to the DECtalk system. A mapping between the acoustic correlates of emotion in human speech (particularly drawing on the work of Fernald (1989)) and the synthesiser parameters was also developed, and thus when used together, the robot was able to make expressive gibberish utterances and as a result engage in proto-dialogues. Human listener evaluations were performed where they listened to eighteen different utterances made by Kismet, and designed to cover six expressive qualities (anger, fear, disgust, happiness, surprise and sorrow) and asked to select an emotion from the list of six labels. Her results found that while subjects tended to perform well, and were able to correctly interpret the utterances with respect to their affective colouring, subjects did confuse some labels, namely fear was confused with surprise, and disgust was confused with anger and surprise also. While all the work on the Kismet presents a mile-stone with respect to the field of HRI, there is one primary problem with the work on the vocal system. This is that DECtalk was a *closed* system, and thus much of the system remains undocumented making it difficult to reproduce in other systems.

Oudeyer (2003) has developed a similar system to that of Breazeal, however using the freely available MBROLA synthesiser, seeking to remedy the problem of replication. The general aim of this was to produce, at low computational

cost, expressive and emotional, cartoon like speech for both robotic and virtual agents⁷. As with Kismet, an algorithm was developed to generate pseudo-random text strings of phonemes that were then fed into the TTS engine along with a specification of the synthesiser settings in order to provide an affective colouring. Five different emotional states were modelled using this method: Happiness, Sadness, Anger, Comfort and Calmness. Subject evaluations consisted of presenting subjects with 30 different utterances, each representing one of the five emotional states but with different input text strings, and subjects asked to say which of the five labels the utterances corresponded to. These evaluations revealed that with this system, subjects had overall high recognition accuracies ($> 65\%$), though were found to confuse utterances conveying Calm and Comfort.

Following this, the work on gibberish speech lay dormant for a few years, until efforts were again undertaken by Yilmazyildiz et al. (2006), who were keen to pick up the field and investigate the possibilities that gibberish speech may have during Child-Robot Interaction with their robot, Probo. In this work they also moved away from the approach of feeding gibberish input into TTS engines. Rather, they developed a concatenative speech synthesis engine that used a database of both natural and expressive speech from a voice actor. Gibberish expressions were created by selecting a random expressive speech sample and using this as a prosodic template, and then concatenating syllabic samples from the natural speech database to create a gibberish sentence, and copying the pitch and timing structure of the template to the carrier signal. Unfortunately while utterances could be created to represent four basic emotions (anger, joy, sadness and fear), no human decoding evaluations were performed, only the method of utterance creation was detailed.

Revisiting the approach of creating gibberish test string that are fed into a TTS engine, Yilmazyildiz et al. (2010) presented a novel method for creating the text input string. They present a simple algorithm which takes a string of real natural language text as an input, and replaces all of the vowels with another (based

⁷This particular research effort had a longer term goal in that it was intended to be used as a base expressive vocal system through which language acquisition in a real robot could be studied.

upon the frequency of use of a given vowel in the language of the input string). What is novel about this approach is that rather than creating an utterance that has no meaning, their algorithm corrupts the meaning. This results in gibberish speech that resembles a real language also, though linguistic content remained unintelligible. Through human evaluations, they found that it is important to match the language of the original input text string with the language of the TTS engine as this impacts the voice quality of the synthesis. Furthermore, they found that subjects were better able to discriminate between *positive* and *negative* utterances when the original input text contained uncorrupted semantic context that matched the valence of the para-linguistic cues of the utterances.

The work described by Yilmazyildiz et al. (2011) charts the initial steps toward merging their two previous approaches, in that they used their novel method of creating gibberish text strings, applying this to various pieces of text, and asking a voice actor to make expressive recordings of these for a small number of emotions (neutral, anger, disgust, fear, happiness, sadness and surprise). A human subject evaluation using a decoding paradigm found impressively high recognition rates for the different portrayed emotions, however one must be wary of the issues surrounding a limited number of emotional categories when performing these evaluations. This particular work has a great deal of potential however: with this alone they have a database of very high quality expressive, gibberish speech which they can take samples of and use for affective expression for their Probo robot. Furthermore, they are able to apply their method for concatenative speech synthesis in order to extend the repertoire of utterances recorded from the voice actor, and produce an infinite number of utterances that have a certain affective charge, this is reported by Yilmazyildiz et al. (2013). Furthermore in this work, they performed an evaluation where their gibberish speech samples (conveying *Angry*, *Disgust*, *Fear*, *Joy*, *Sadness* and *Surprise*) were either played alone, or combined with facial expressions in the Probo robot. It was found that recognition rates varied for the auditory only conditions (29 % for *Joy* to 100% for *Sad*), and that when combined with facial gestures, there was a significant

increase in the overall recognition of the emotions expressed by the robot. This is the same general finding as Jee et al. (2007), where mixing modalities increased overall decoding accuracy, and is a useful, yet intuitive insight to have.

2.4.3 Discussion

As can be seen from the review of work directly related to HRI, research focusing NLUs and gibberish speech as been very limited indeed. While on one hand this is a concern as there is little directly relevant work to relate to, the positive aspect is that it leaves many open questions that need to be addressed. Some of these are discussed here, as they have influenced the direction that the work in this thesis has taken and the methodology that has been adopted. These are the overall methodology used for evaluation affective meaning in NLUs, the evaluation settings and subjects used, and the need to go beyond only evaluating different methods for creating utterances, but rather explore how real world HRI influences how NLUs should be used and what impacts this.

2.4.3.1 Evaluation Methodology

Generally, throughout both the work on NLUs and gibberish speech there has been a tendency toward studies using a *decoding* paradigm (Banse and Scherer, 1996; Scherer, 2003), where a limited number utterances have been created to represent a limited number of affective categories/labels (e.g. *anger*, *happiness*, ect), and subjects are to *decode* each utterance and assign one of the labels to the utterance. Table 2.3 seeks to illustrate this and charts the various studio that have been referred to in this review, outlying the method of utterance generation, the acoustic parameters that were varied, and the difference affective states that were portrayed by the utterances during evaluations and the means of measurement of subject interpretations. As we can see, most of the studies have employed discrete affective categories as the measurement method.

These studies often report confusion matrices in their results (e.g. (Breazeal, 2002; Oudeyer, 2003; Tuuri et al., 2011)), indicating which states are confused

with others. While through this it is possible to identify confusion patterns between states, leading to potential insights regarding how close different concepts of affective states are, such information provides limited insights into what features of each sound were involved in the users *decoding* process, and to what degree these utterance features influence affective charging.

This method of evaluation also brings up a larger debate surrounding the use of categorical measurement as a whole (particularly in the speech synthesis domain): how many different categories should be used? Banse and Scherer (1996) and Scherer (2003) argue that with fewer forced choice categories presented to subjects, evaluation becomes more a task of *recognition* rather than *discrimination*. They take the view that simple, basic emotion categories (Ekman, 1992; Ortony and Turner, 1990; Plutchik, 1994) are better broken down into pairs, such a hot and cold anger, providing higher granularity in affective measurement and increases the number of labels presented. This is an ongoing debate that has not been resolved as of yet, and is likely to be very dependent upon the subjects that are being tested (e.g. adults v.s. children). While such decoding studies are interesting insights from the perspective of probing emotional representations, a general drawback is that this approach is less fruitful when one is concerned with understanding the mechanisms through which utterance features are exploited to charge an utterance. These are vital insights when one is concerned with understanding how utterance features are best exploited to express particular states, and if one is to introduce variability to this.

There is also a debate surrounding whether categories are the most appropriate representation of affect to present to subjects. There are suggestions that measurement tools based upon dimensional representations of affect are a more suitable foundation to employ since continuous dimensions are more resolute to subtle changes in affective states (Cowie and Cornelius, 2003; Schröder, 2004; Laukka, 2005), changes that are likely to occur during social interactions. Furthermore, if one were to use dimensional approaches for measurement, one becomes well poised to gain insights about how subjects actually perceive the stimuli. Do

they exhibit coarse differences in affective inference, or are the inferences subtler? Unfortunately, there is also the question as to the number of dimensions there should be in such an affect space and what each dimension represents (Fontaine et al., 2007). This is something that shall be addressed in more detail in the next chapter.

There appear to have been no *inference* studies thus far. Rather than being concerned with a users ability to recognise and decode the affective state of an agent based upon utterances designed to convey certain emotions, inference studies are more concerned with uncovering the underlying mechanisms for each acoustic feature of an utterance that influence affective expression, communication and interpretation (Scherer, 2003). This somewhat reflects the strategies through which authors of the previous work have affectively charged their utterances. In the majority of cases, studies have drawn upon insights from previous work across various fields to inform the design of stimuli. While this approach is solid in that it draws from previous results, it also makes the assumption that there is enough overlap between parameter modelling and configuration of previous work and the current work. This is more likely the case with gibberish speech than this assumption stands than with NLUs, given the heavy reliance of TTS technology and the use of human voice recordings. Not all studies have adopted this approach however. Affective charging has also been achieved via simulating affective states by recording actor portrayals and using these as templates. However there are questions over how genuine actor portrayals of emotions actually are (Scherer, 1986; Banse and Scherer, 1996; Scherer, 2003).

2.4.3.2 Subjects and Evaluation Settings

Another notable observation from the previous work is that the general age range of subjects has also been somewhat constant across all the studies that have been referred to (see table 2.4). All evaluations have been with adult subjects, and been conducted within a lab setting, with the exception of the work by Yilmazyildiz et al. (2013) who performed their evaluation with both adults and young teenagers.

Table 2.3: Acoustic parameters used to generate utterances in previous work, the emotions that were portrayed and the means of affective measurement.

Study	NLU/Gibberish	Acoustic Parameters	Emotions Portrayed	Measurement
Breazeal (2002)	Gibberish	Accent Shape Average Pitch Contour Slope Final Lowering Pitch Range Pitch Base Speech Rate Stress Frequency Breathiness Brilliance Laryngealization Loudness Pause Discontinuity Pitch Discontinuity Precision of Articulation	Categories: <i>Anger</i> <i>Fear</i> <i>Disgust</i> <i>Happiness</i> <i>Surprise</i> <i>Sorrow</i> <i>Neutral</i>	Categories: <i>Anger</i> <i>Fear</i> <i>Disgust</i> <i>Happiness</i> <i>Surprise</i> <i>Sorrow</i> <i>Neutral</i>
Oudeyer (2003)	Gibberish	F0 mean F0 variance F0 max F0 contour Last word contour Last word accent Accent probability Mean duration Duration variance Volume	Categories: <i>Happiness</i> <i>Sadness</i> <i>Anger</i> <i>Comfort</i> <i>Calm</i>	Categories: <i>Happiness</i> <i>Sadness</i> <i>Anger</i> <i>Comfort</i> <i>Calm</i>
Jee et al. (2007)	NLU	Tempo Key Pitch Melody Harmony Rhythm Volume	Categories: <i>Happy</i> <i>Sad</i> <i>Fear</i> <i>Dislike</i>	Categories: <i>Happy</i> <i>Sad</i> <i>Fear</i> <i>Dislike</i>
Jee et al. (2009)	NLU	Tempo Key Pitch Melody Harmony Rhythm Volume	Categories: <i>Joy</i> <i>Distress</i> <i>Shyness</i> <i>Irritation</i> <i>Pride</i> <i>Dislike</i> <i>Expectation</i> <i>Anger</i>	Categories: <i>Joy</i> <i>Distress</i> <i>Shyness</i> <i>Irritation</i> <i>Pride</i> <i>Dislike</i> <i>Expectation</i> <i>Anger</i>
Jee et al. (2010)	NLU	Intonation Pitch Range Timbre	Categories: <i>Happiness</i> <i>Sadness</i> <i>Affirmation</i> <i>Denial</i> <i>Encouragement</i> <i>Introduction</i> <i>Question</i>	Categories: <i>Happiness</i> <i>Sadness</i> <i>Affirmation</i> <i>Denial</i> <i>Encouragement</i> <i>Introduction</i> <i>Question</i>
Komatsu (2005)	NLU	Pitch Frequency Frequency Envelope/Pitch Slope Duration	Categories: <i>Disagreement</i> <i>Hesitation</i> <i>Agreement</i>	Categories: <i>Disagreement</i> <i>Hesitation</i> <i>Agreement</i>
Komatsu and Yamada (2007, 2008); Komatsu et al. (2011)	NLU	Pitch Frequency Frequency Envelope/Pitch Slope Duration	Categories: <i>Positive</i> <i>Negative</i> <i>Undistinguishable</i>	Categories: <i>Positive</i> <i>Negative</i> <i>Undistinguishable</i>
Komatsu et al. (2010); Komatsu and Kobayashi (2012)	NLU	Pitch Frequency Frequency Envelope/Pitch Slope Duration	Robot's Confidence: <i>Confident</i> <i>Not Confident</i>	Perception of Confidence: <i>Confident</i> <i>Not Confident</i>
Tuuri et al. (2011)	NLU	Pitch Contour Utterance Duration Voice Intensity	Categories: <i>Slow Down</i> <i>Urge</i> <i>Ok</i> <i>Reward</i>	Categories: <i>Slow Down</i> <i>Urge</i> <i>Ok</i> <i>Reward</i>
Yilmazyildiz et al. (2006)	Gibberish	Pitch Timing	Categories: <i>Anger, Joy</i> <i>Sadness, Fear</i>	<i>No evaluation took place.</i>
Yilmazyildiz et al. (2010)	Gibberish	Mary TTS parameters	Categories: <i>Happy, Sad</i>	Categories: <i>Happy, Sad</i>
Yilmazyildiz et al. (2011)	Gibberish	Recorded actor samples were used	Categories: <i>Neutral</i> <i>Anger</i> <i>Disgust</i> <i>Fear</i> <i>Happiness</i> <i>Sadness</i> <i>Surprise</i>	Categories: <i>Neutral</i> <i>Anger</i> <i>Disgust</i> <i>Fear</i> <i>Happiness</i> <i>Sadness</i> <i>Surprise</i>
Yilmazyildiz et al. (2013)	Gibberish	Recorded actor samples were used	Categories: <i>Neutral</i> <i>Anger</i> <i>Disgust</i> <i>Fear</i> <i>Joy</i> <i>Sadness</i> <i>Surprise</i>	Categories: <i>Neutral</i> <i>Anger</i> <i>Disgust</i> <i>Fear</i> <i>Joy</i> <i>Sadness</i> <i>Surprise</i>

Also, the number of subjects has generally quite low. It appears that there has been no work probing the interpretation of NLUs by young children (ages less than 10). Another interesting observation is that there have been no studies that have addressed subjects with social disorders such as autism, while efforts have been undertaken into understanding how robots can be used to investigate such social disorders (e.g. Robins et al. (2004)).

Child-Robot Interaction (cHRI) is an area of HRI that has shown great promise in recent years and is currently gathering momentum as a subfield of HRI as evidence through research efforts such as the ALIZ-E project (Belpaeme et al., 2012). The primary reasons for this are the willingness that children shown to engage in HRI, and suspect their disbelief (Breazeal, 2003a; Ros Espinoza et al., 2011; Salter et al., 2008).

Given the increasing number of potential application areas of cHRI, NLUs may have a particular amount of promise for the the use in this area, however there have currently been no efforts to explore this as of yet. This is one particular facet that the work in this thesis seeks to address, if only initially. Furthermore, there has been no work addressing how adults and children differ in their perception of NLUs and whether they have the same affective inferences from utterances. This too is something that this research seeks to address where possible.

2.4.3.3 Going beyond Affective Charging

The vast majority of the previous work has focused upon how NLUs and gibberish speech can be generated, presenting a variety of different methods and techniques. A limited number of utterances are then created using these methods, designed to convey a given affective state. Unfortunately, these studies become evaluations of specific methods for creating utterances which means that the results have a limited capacity for generalisation as to the application of NLUs and gibberish speech more broadly. Furthermore, some of these evaluations were conducted without the use of a real robot (e.g. Oudeyer (2003), Jee et al. (2007), Yilmazyildiz et al. (2006), Yilmazyildiz et al. (2010) and Yilmazyildiz et al. (2011)). The

Table 2.4: Number of subjects and subject age ranges in reviewed previous work.

Study	# Subjects	Age Range
Breazeal (2002)	9	23-54
Jee et al. (2007)	20	“undergraduates”
Jee et al. (2009)	NA	NA
Jee et al. (2010)	20	20-25
Komatsu (2005)	23	20-42
Komatsu and Yamada (2007)	9	21-24
Komatsu and Yamada (2008)	20	19-24
Komatsu et al. (2010)	19	22-25
Komatsu et al. (2011)	20	19-24
Komatsu and Kobayashi (2012)	20	21-28
Oudeyer (2003)	8	“adults”
Tuuri et al. (2011)	12	23-29
Yilmazyildiz et al. (2006)	NA	NA
Yilmazyildiz et al. (2010)	10	24-37
Yilmazyildiz et al. (2011)	11	27 - 32
Yilmazyildiz et al. (2013)	35	10 - 14

problem is that such evaluations focused upon a specific technique for creating utterances has no real input regarding how NLUs/gibberish speech can be used during real HRI.

This is something that sets the works by Komatsu *et al.* aside. Their research has focused upon a very simple set of utterances, but explored how these utterances are perceived by people when presented through different robots, and more importantly, how these utterances influence a real interaction that is not context free. Such knowledge something that the overall field is lacking and is in desperate need of if both NLUs and gibberish speech are unlock the potential benefits that they have to offer during real world HRI. The results of obtained from their simple experiments with simple utterances have been fascinating and provided initial valuable insights showing that within a interaction, NLUs indeed can have an important influence over how people behave.

However, the drawback with their work is that they have employed very simplistic utterances that have been hand crafted and have no really addressed affective meaning beyond simple positive or negative valence. The world of vocal affective displays is far richer than this, and a clear area that requires further exploration is how more acoustically rich NLUs are capable of conveying more

complex affective states, how these can be used during HRI, and how these can influence people during an interaction as well as how the interaction can influence the utterances. Highlighting and addressing such potent questions is another the prime goal of the research presented in this Thesis.

2.4.4 Review Summary

The review of previous work presented above serves to highlight a few important points regarding the field of NLUs and gibberish. Firstly, and perhaps most importantly, is that the field is young and not very well established, with only a few research efforts having been active, and in some cases, spread out of over the last decade or so. The review has covered both NLUs as well as gibberish speech as it is considered that there two means of expressive vocalisations are very much related and share a common underlying goal: to allow robotic agents to make expressive displays through sound without the need to rely on natural language.

In the case of NLUs, methods for creating NLUs have mainly revolved around the use of music synthesisers, and either drawing upon insights from the field of musicology to provide an affective charging to utterance, or taken samples of human expressive vocalisations and extracting a basic prosodic template and mapping these to musical MIDI notes and then synthesising utterances, and creating simple sine wave signals which have either a rising, falling or flat pitch contour. With respect to gibberish speech, the methods for creating these have all revolved around borrowing the methods and technologies developed by the world of speech synthesis and using these to create utterances that resemble human speech, but do not contain any linguistic content. Where TTS technology has been used, the developments in synthesising emotional speech have been adopted and directly utilised, however in some cases authors have decided to create their own synthesisers by recording voice portrayals of expressive speech and reimplementing methods for using these to re-synthesise new utterances.

All of the previous work has used the *decoding* paradigm for performing evaluations, where utterances were created to portray basic emotions and people had

to identify which emotion they thought the utterances conveyed. As such, all the evaluations report confusion matrices which generally show good accuracy in decoding the affective meaning of utterances. Furthermore, the accuracy of the decoding is increased when NLUs are combined with facial gestures, highlighting the importance of synchronising and aligning affective displays across multiple modalities.

However, the main drawback with this is that there are few insights that inform how the specific parameters of the specific synthesis methods influence affective meaning, rather only holistic insights overall different synthesiser parameter configurations are gained. No evaluations have been performed using a *inference* paradigm where specific features of an utterance are manipulated in order to identify how these different features contribute to different affective interpretations in the eyes of subjects.

In light of the review of previous work, a number of important observations have been made with regarding to the types of utterances that have been studied, the evaluations methodologies have been performed, the types of subjects that have been used in these evaluations, and the overall exploration of how NLUs and gibberish can be applied to HRI in a broader sense, rather than just evaluation different methods for creating utterances. As such, many of these observations are taken on board in the work described in this thesis. Specifically, this thesis seeks to address the following gaps:

- NLUs consisting of simple, single sine waves has been covered well in the literature. However, concatenating such sounds together to form sentence like structures has not been addressed as of yet (while the majority of work in gibberish speech has). NLUs in this thesis are designed to consist of multiple, concatenated single sine waves as they are thought to provide a richer repertoire of utterances.
- Moving beyond creating utterances to convey a small number of basic emotion labels and asking subjects to decode these, this thesis moves toward an inference study paradigm where utterances are systematically manipulated

in order to gain deeper insights regarding how the different properties of an utterance impact the affective interpretation people have of the utterances.

- Furthermore, the work here moves away from the common use of affective labels and adopts representations of emotions that are continuous in nature rather than discrete. The method through which this is done is presented in more detail in the next chapter.
- There has been a lack of focus upon how NLUs and gibberish speech can be applied to Child-Robot Interaction, with no evaluations performed at all. The work in this thesis addresses this and where possible has performed evaluations with both adults and young children in order to gain comparative insights.
- Some research efforts have gone beyond simple evaluations of the inference of affective meaning from the utterances, but has sought to explore what aspects of HRI might impact this, such as the robot morphology and the contexts within which utterances are used. The work in this thesis seeks to further explore this also.

2.5 Are NLUs a language?

Finally, a word must be said about NLUs and their potential relation to *language*. This thesis has opted to view NLUs as not having two fundamental properties of a natural spoken language, namely the lack of an established set of social and cultural conventions regarding a vocabulary of discrete and arbitrary symbols, and a grammatical structure and rules outlining how the symbols should be used together to convey meaningful, semantic information. However, it is possible to argue that NLUs could be used in such a manner as to form an artificial language. Assuming that there were established conventions on a vocabulary and grammar, NLUs would have the following fundamental properties of a language, as proposed

by Hockett (1960)⁸:

- *Semanticity*, where the ties between the meaningful elements of a message and their meaning can be arbitrary or non arbitrary.
- *Displacement*, allowing agents to communicate about events and objects that are not present in the environment, across difference tenses (i.e. the past, present and future).
- *Arbitrariness* in the vocabulary containing the symbols (and sounds associated with these) used to denote objects in the world.
- *Productivity*, in that the potential combination of symbols in the vocabulary is unique.
- *Discreteness*, in that the acoustic sounds have their own discrete meaning.
- *Duality* in that the “words” in an utterance have a discrete meaning, and each word is made.
- The language may be acquired through *cultural transmission*, i.e. through other agents present in the environment.

The rationale for this thesis not considering NLUs as language is that given that language is a large, complex field of study, adopting the view that does consider NLUs as a language introduces a level of complexity that would most likely draw the direction and focus of attention of this body of work away from the primary goals, as potentially make the achievement of these goals cumbersome to a degree that is impractical. Furthermore, as stated above, as there is no cultural agreement regarding a vocabulary of grammar regarding the use of NLUs, not all the properties of a language currently present. This is of course not to say that NLUs do not have the potential to become a language. Rather, on the contrary,

⁸In the original article, Hockett (1960) proposes in total 13 properties that are universal to language, however the remaining properties (the vocal-auditory channel, broadcast transmission and directional reception, rapid fading, specialisation and total feedback were the listener can reproduce what they hear) relate specifically language through vocal expression, and in the light of *artificial* languages such as sign language, their value with respect to the broader concept of language is deemed as limited.

they indeed do hold this potential, however, given that not all the properties of a language are currently met, viewing NLU as an artificial language of their own is a premature notion at this time.

However, recent work by Mubin et al. (2009, 2010a,b, 2012) has begun to address this. Their work seeks to tackle the problems associated with ASR in a direct manner, suggesting that the use of artificial languages may provide a means to increase ASR performance. They have developed the Robot Interaction Language (ROILA), an artificial language (that greatly resembles gibberish speech) designed to strike a balance between minimal learning effort for a human, and maximum performance in speech recognition.

2.6 Summary

This chapter has served to provide a more concrete description of what NLUs are, and what they are not. It has also introduced a similar modality, *gibberish speech*, as describing both these modes of expression together help provide more accessible and tangible examples of the underlying motivations and potential utility of each is. The chapter began with a more detailed explanation of the term *Non-Linguistic Utterances*, followed by examples of their use in both real and fictional robots. The motivations for studying these kind of utterances, and the potential utility that they have in HRI were outlined and discussed.

This was followed by a review of affective displays through sounds, and particularly in the human voice, but also through music. The commonly overlooked observation that NLUs and gibberish speech, and more specifically the means through which they are created, arguably have their roots in the field of psychology where similar techniques have been used to create expressive vocalisations that have masked or distorted semantic/verbal context, was also highlighted. The acoustic correlates of emotional speech was also reviewed from a high level, given the large body of research that exists, and the links with emotional expression through music highlighted as these all have relevance to the study of NLUs and gibberish speech.

Having outlined the nature and motivations of NLUs and gibberish, and an overview of the acoustic correlates of emotional expression through the human voice provided, the (limited) existing body of literature regarding NLUs and gibberish speech has been covered in detail, charting the developments that have been made, and the general different approaches that have been taken when creating utterances. This review was then discussed and a number of gaps in the current research highlighted. Namely, that there have been no studies aimed at young children, studies have focused on creating utterances to convey simple basic emotions and performing decoding studies in context-free settings, and that there have been very few efforts in studying how aspects of HRI impact how people perceive NLUs and how their potential utility in HRI can indeed be harnessed and applied. These gaps are then taken on board and have helped determine the direction that this research has taken.

Finally, the view that NLUs do not currently constitute a language is justified. NLUs currently lack the fundamental, universal properties that a language has, and as such this is deemed to be a premature notion, however the potential for NLUs to indeed become an artificial language of their own in the future is by no means dismissed.

Chapter 3

Methods

Summary of the key points:

- A custom method for creating and characterising NLUs is detailed. This method takes inspiration from related work on NLUs, gibberish speech and literature on the acoustic correlates of emotional expression in the human voice and music.
- The Nao humanoid robot is described. This is the robot that has been solely used as the platform in which NLUs have been embodied and studied in this thesis.
- Categorical and dimensional representations of affect are described and discussed, as is their relation to their use in synthetic systems with an affective component, and tools designed for measuring affect from humans.
- The affective measuring tool of choice, the AffectButton is described in detail, as is the application to this thesis.

This chapter serves to outline the tools that have been used in this body of research with respect to how NLUs have been created, the robotic platform in which they have been embodied, and the means through which peoples' affective interpretations of these have been captured. Specifically, the custom method that has been developed for characterising and synthesising NLUs is presented first. Then the Nao humanoid robotic platform is described. Finally, the tool used to capture affective ratings is described, as well as a brief overview and discussion of the issues surrounding its use in this thesis.

3.1 Creating NLUs

The previous chapter highlighted that currently no method exists for creating, specifying and systematically manipulating NLUs, while this is the case for gibberish speech. In order to investigate the use of NLUs beyond single tones, and into the realm of multiple, concatenated tones, a method was developed to generate and synthesise utterances in order to produce the acoustic stimuli used in experiments. This section serves to provide a detailed description and overview of the design and implementation of this method, covering the conceptual design and parameterisation of NLUs. A pseudo-code algorithm is outlined in Appendix A to allow others to reimplement this into a programming language of their choice.

As we shall see, this method allows for a rich variety of NLUs to be created and manipulated, however, in this research, this full potential is not exploited and only a limited number of parameters are systemically manipulated. This is done in order to maintain a manageable handle upon the exploration of the NLUs.

3.1.1 NLU Anatomy and Parameterisation

In this work, the notion of using single tones of sine waves with modulated frequency as used by Komatsu and Yamada (2011) and Komatsu et al. (2011) is used as an initial foundation for conceptual design of utterances, and is extended to include and cater for multiple, concatenated sine waves. Doing this has required a formalisation of parameters that are used to characterise utterances such that

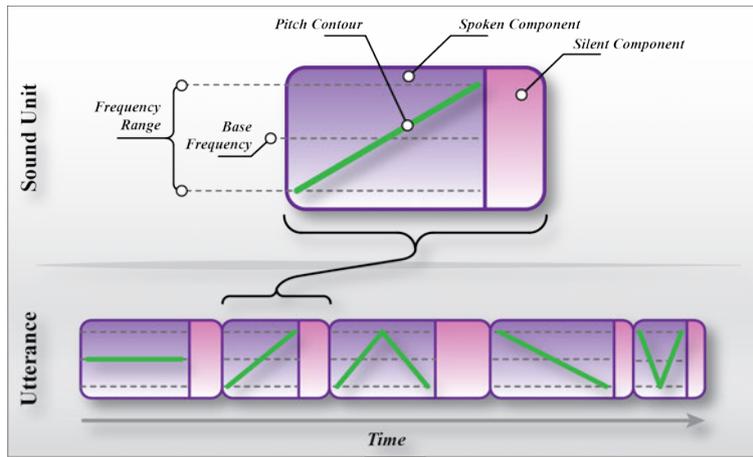


Figure 3.1: Utterance with 5 concatenated sound units, each with a different pitch contour and pause ratio, and a rhythm rhythm value that is less than 1.

acoustic features may be modulated and specified in a controlled and systematic manner.

Fundamental to the utterance anatomy is the pairing of a carrier signal (or acoustic *tone*) with a duration of silence, which are referred to as *Sound Units*, and are concatenated to form sentence-like *Utterances* (see figure 3.1). Sound units, have two basic components: a *spoken* component and a *silent* component, each with a specified temporal duration. The spoken component is essentially a container for a carrier signal, that is modulated using a frequency and amplitude envelope. Each envelope consists of a temporally ordered array of n *Nodes* sequentially connected by $n - 1$ *Edges*. These nodes and edges lay in a 2 dimensional (bi-) normalised space, where the dimensions represent time (covering the range $[0\ 1]$), and frequency (covering the range $[-1\ 1]$) or amplitude (covering the range $[0\ 1]$) for the respectively envelopes. In the case of the frequency envelope, it is the shape of this array of nodes with respect to time that is referred to as the *pitch contour*.

While this representation of an envelope caters for a wide variety of custom specifications, this can be considered a hinderance as it can produce an overwhelming number of variables which may be difficult to account for during experiments. As such, in this work, both the frequency and amplitude envelope shapes have been limited in order to allow for a systematic characterisation and manipulation. The shape of the frequency envelopes have been limited to 5 characterisations:

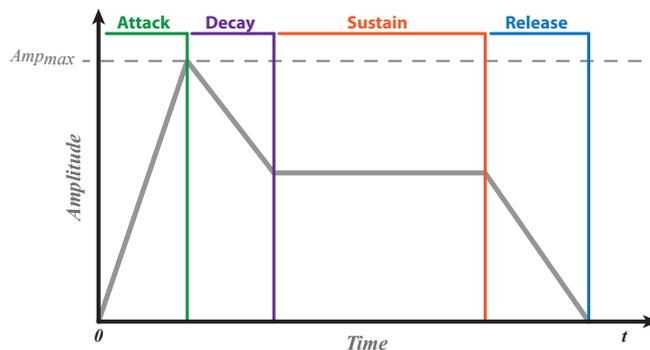


Figure 3.2: Schematic of a common Attack Decay Sustain Release (ADSR) amplitude envelope with linear transitions.

Table 3.1: Values of the five nodes for the amplitude envelope located in the bi-normalised space for each carrier signal in each sound unit (see figure 3.2). These values have been held constant throughout the work described in this thesis.

Node	Time value	Amplitude value
n_1	0	0.05
n_2	0.1	1
n_3	0.2	0.7
n_4	0.9	0.7
n_5	1	0.05

flat, rising, falling, rising-falling and *falling-rising* (as shown in figure 3.1). These characterisations however do not limit the size of the node array used to create the pitch contour. With respect to the amplitude envelope, this has been kept constant, following the classic linear Attack Decay Sustain Release (ADSR) format (figure 3.2) used in many commercial music synthesisers. As a result, the envelope consists of five nodes and four edges, where the nodes represent the transitions between the components of the envelope, and the edges represent the components themselves. These values are shown in table 3.1, with the values being relative to the normalised space of each envelope in a sound unit.

This representation has the primary benefit in that it is easy to scale the values of the nodes from their (bi-)normalised values to real world values of frequency and (mili-)seconds, without the need to alter the specification of the envelope themselves.

The transform of this normalised representation of an utterance to real world units is facilitated via a variety of parameters, some of which may be applied

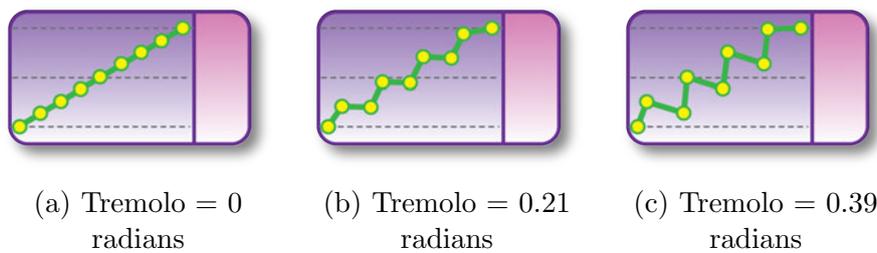


Figure 3.3: Example of how the Tremolo parameter applies to the frequency envelope of a carrier signal in a sound unit.

locally to individual sound units, and others that are applied *globally* to all sound units within an utterance. These are summarised in table 3.2. An example of an exclusive global parameter is the *sound unit count* which specifies the number of sound units that are to be concatenated to form an utterance while the *wave count* parameter controls how many different carrier signals¹ are to be contained within each sound unit and is an example of a parameter that may be specified uniformly for all sound units, or have a unique value for each sound unit individually.

The frequency envelope has 14 parameters associated with it. The *wave type* is the type of carrier signal (sine, saw, square, etc) that is to be modulated by the envelope, allowing for a variety of *timbres* to be used for each sound unit. The *node count* is the number of nodes that are to be used to create each pitch contour, with the minimum number of nodes required to create all of the five specified pitch contours being 3. As well as being able to alter the type of carrier signal, a tremolo effect may also be applied. This is done by modulating the frequency value of a node by the tangent of the *tremolo* angle (see figure 3.3).

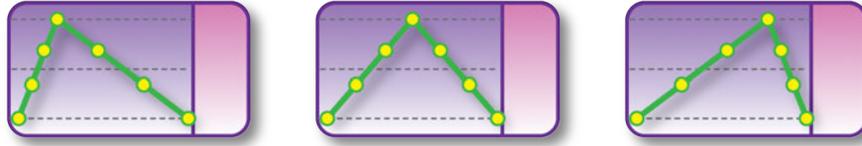
Given the five classifications of pitch contour, there are an additional two parameters that key be defined, and which only apply to the rising-falling and falling-contours. These are the *Node Ratio* and the *Skew Ratio*. The Node Ratio is used to specify how many nodes (of the number specified by the *Node Count*) are to be placed *before* the maximum/minimum value of the contour, and how many *after* this point (see figure 3.4). The *Skew Ratio* is used to control the location of this maximum/minimum point in the sound unit along the time dimension (see figure 3.5).

¹Each sound unit can have a number of different carrier signals that can all be unique. However, for this the work in this thesis, each sound unit only has a single carrier wave.



(a) Node Ratio = 0.9 (b) Node Ratio = 0.5 (c) Node Ratio = 0.1

Figure 3.4: Example of how the Node Ratio parameter applies to the frequency envelope of a carrier signal in a sound unit.



(a) Ratio = 0.25 (b) Ratio = 0.5 (c) Ratio = 0.75

Figure 3.5: Example of how the Skew Ratio parameter applies to the frequency envelope of a carrier signal in a sound unit.

The *Base Frequency* and *Frequency Range* parameters are used to specify the frequency and range through which the nodes in the pitch contour are to be scaled to transform them to real world values (in *Hz*). In a similar fashion, the *Volume Intensity* parameter is uniformly applied as a scalar to all nodes within an amplitude envelope to either increase or decrease the final acoustic volume of the sound unit. All three parameters may be applied both at the *global*, utterance level, or may be applied to *locally* each sound unit individually.

With respect to the temporal dimension of an utterance, there are three parameters; *rhythm*, *pause ratio* and *speech rate*. The duration of the spoken component is pseudo randomly determined using the rhythm parameter which controls the range of values from a set value of 1 ($rhythm \leq 1$) in which a duration value is randomly selected (equation 3.1). Once the duration of the spoken component has been determined, this value is multiplied by the pause ratio to calculate the duration of the silent component (equation 3.2). At this stage, there are no real world units (*ms*) associated with these values, rather these values serve to outline the *proportional* durations of sound units with respect to each other. The speech rate is used to specify how quickly all the sound units in an utterance should be articulated, and thus provides a means through which the abstract duration values may be scaled to real world values. Practically, the speech rate is used in

conjunction with the sound unit count to calculate a scalar constant, k (equation 3.3), that is applied to duration values of both the spoken and silent components of each of the sound units (equations 3.4 and 3.5).

$$dur_{spoken} = rhythm + [(1 - rhythm) \cdot rand()] \quad (3.1)$$

$$dur_{silent} = dur_{spoken} \cdot pauseRatio \quad (3.2)$$

$$k = \lceil \frac{soundUnitCount}{speechRate} \rceil \cdot \frac{1}{dur_{total}} \quad (3.3)$$

$$dur_{spoken} = dur_{spoken} \cdot k \quad (3.4)$$

$$dur_{silent} = dur_{silent} \cdot k \quad (3.5)$$

Here d_{spoken} is the duration of the spoken component, d_{silent} is the silent duration of the sound unit, and $rand()$ is a pseudo random number generator where $0 \leq rand() \leq 1$. With this arrangement, when $rhythm = 1$, all sound units within an utterance will have a spoken duration with the value 1, while if the $rhythm = 0.5$, the spoken duration is randomly assigned from the range where $0.5 \leq dur_{spoken} \leq 1$. An important point to highlight here is that the $rhythm$ parameter is the only parameter that when not equal to 1 introduces a random element to an utterance - two utterances with identical parameter values and where $rhythm < 1$ are not guaranteed to be identical. When this has been done, the *Base Frequency* and *Frequency Range* are then used to scale the nodes in the frequency envelope to their real world values.

Table 3.2: Summary of the parameters for characterising NLUs in this work.

Parameter Name	Level		Values (x)	Description
	Utterance	Sound Unit		
Wave Type	✓	✓	[sine, saw]	The type of carrier signal wave form.
Pitch Contour	✓	✓	[flat, rising, falling, falling-rising, rising-falling]	The shape of the frequency envelope of a given sound unit.
Base Frequency (Hz)	✓	✓	$x \in \mathfrak{R} 500 \leq x \leq 1500$	Frequency around which the Pitch Contour of a sound unit is shaped.
Frequency Range (Hz)	✓	✓	$x \in \mathfrak{R} 500 \leq x \leq 1500$	Frequency range either side of the Base Frequency that determines the minimum and maximum frequency values of the Pitch Contour.
Speech Rate	✓	-	$x \in \mathfrak{R} 1 \leq x \leq 6$	The number of sound units synthesized per second.
Sound Unit Count	✓	-	$x \in \mathbb{Z} 1 \leq x \leq 5$	The number of sound units within an utterance.
Pause Ratio	✓	-	$x \in \mathfrak{R} 0.05 \leq x \leq 1.5$	The ratio of the duration of the spoken component to the silent component in a sound unit.
Rhythm	✓	-	$x \in \mathbb{R} 0 \leq x \leq 1$	The amount of random variance allowed in the duration of a sound unit.
Volume Intensity	✓	-	$x \in \mathbb{R} 0.5 \leq x \leq 5$	Scalar value applied to all sound units within an utterance to increase or decrease the overall acoustic volume.
Node Count	✓	✓	$x \in \mathbb{Z} 3 \leq x \leq 15$	The number of nodes that are used to construct the frequency envelope of a sound unit.
Tremolo (rad)	✓	✓	$x \in \mathfrak{R} 2\pi/16 \leq x \leq 2\pi/16$	The angle at which the nodes are offset from a linear line intersection the first and last nodes, this is perceived as a tremolo on the utterance.
Skew Ratio	✓	✓	$x \in \mathfrak{R} 0 \leq x \leq 1$	Proportional location of a Pitch Contour maxima/minima in a given sound unit.
Node Ratio	✓	✓	$x \in \mathfrak{R} 0 \leq x \leq 1$	Proportion of Nodes either side of the a Pitch Contour maxima/minima in a given sound unit.
Envelope Count	✓	✓	$x \in \mathbb{Z} 1 \leq x \leq 5$	The number of frequency/amplitude envelope pairs within a given sound unit.
Utterance Duration	-	-	-	Total temporal duration of an utterance.

3.1.2 Utterance Generation

In this work, NLU are synthesised using *SuperCollider* (McCartney, 2002), a software tool/synthesiser initially developed for computer music audience. The SuperCollider system allows users to dynamically code sounds and synthesise them in realtime, across multiple computers on a network, which has an appealing trait in that it does not need to run locally on a robot, but rather the synthesiser can stream audio to the robot via a network connection.

While SuperCollider has been used in this work, it is only a technical solution, and as such a protocol for converting the “blueprints” (i.e. the characterisation of an NLU as specified by the method detailed above) of NLUs that has been described above into a format that SuperCollider can use to synthesise utterances is a specific solution. As a result, this particular protocol is not presented. The reason for this is that there are many other musical synthesisers that are commercial available and this work does not wish to limit the synthesisers that this particular method for creating and describing utterances in an abstract manner can be applied to.

3.1.3 Remarks on the design of NLUs

From a broad, global perspective, utterances take inspiration from natural language in that they consist of a concatenation of multiple acoustic sounds in fashion that is loosely analogous to how phonemes may be concatenated to form words, and how words may be concatenated together in order to form sentences. One noteworthy comment here is that unlike natural spoken language, NLUs in this work are not subject to any constraining rules such as the phonological or grammatical rules that are found in spoken language. This is of course something that can be introduced, but given the obvious lack of culturally established rules for this, doing this at this stage serves no productive purpose for this research.

The NLU design detailed above provides a versatile solution for utterance specification that is capable of catering for a very wide variety of utterances. While this is an attractive trait of the design with respect to variety and versatility, from

a scientific perspective, it would be unwise to immediately utilise the full potential of this design as the number of variables to control and account for would be very large and cumbersome to manage. As such, all parameters, aside from the pitch contour, have been applied at the global level, uniformly to all sound units within an utterance.

As should be evident, this method for describing and characterising NLUs serves to have a number of analogies with the manner in which the human voice is characterised in both speech synthesis and in psychological literature. For example, the Base Frequency and Frequency range parameters are analogous to the F0 Mean and F0 variability/range values that are reported in the psychological literature, as are the Speech Rate and Pause Ratio parameters, which outline how speech of utterance articulation, and the proportion of silence to sound in an utterance (see Scherer (2003)). The tremolo value is intended to be analogous to the “jitter” or tremolo effect that has been referred to in both human speech and music (see Juslin and Laukka (2003)).

There are also a number of parameters that are specific to this method of describing NLUs, and their purpose is to provide an extensive range of parameters that can be explored to create rich and vibrant sounding NLUs, and it is hoped that this particular method will be adopted by others in the field in order to further study robotic sounding NLUs in a similar, systematic and parameterised approach.

3.2 The Nao robot

The Nao robot (figure 3.6), produced by the French company Aldebaran Robotics, is a versatile 60cm tall humanoid robot designed primarily for, and currently marketed to, researchers investigating social HRI, as well as areas of science that are impacted by this (such as Cognitive/Developmental robotics). This has been the sole platform used throughout this body of research, as not only is it a cost effective platform to use, but as chapter 4 shows, it is also has been deemed by subjects’ to be an appropriate platform through which to embody NLUs. There



Figure 3.6: The Aldbaran Nao robotic platform used throughout this body of research.

is also an added benefit in that much of the research presented in this thesis has been utilised as part of the EU FP7 funded ALIZ-E² project (see Belpaeme et al. (2012)), which also uses the Nao as the sole research platform, and which this research contributes to, and is a part of.

The robot has 25 degrees of freedom, and boasts a wide variety of sensors and actuators that allow it to both sense its surrounding environment as well as manipulate this environment, and are all housed within a plastic external shell. The sensory arsenal includes two sonar sensors, four microphones, two high definition cameras and touch sensors located on top of the head. The actuators are magnetic absolute encoders that actuate the joints and limbs of the robot. The robot also has two built in audio speakers located in the head and an array of RGB Light Emitting Diodes (LEDs) that serve to represent and animate two eyes. There is an onboard embedded computer that runs a custom Linux distribution which plays host to *NaoQi*, a pseudo operating system used to provide both a high and low level interface to the onboard resources. Also included in *NaoQi* is a built in Speech Recognition engine, Text-To-Speech engine and Computer Vision libraries (which provide onboard face detection and recognition, and object recognition).

²www.aliz-e.org

3.2.1 Programming Nao

Aldebaran provide a number of computational solutions for the Nao platform. Firstly, as the onboard computer runs a Linux based operating system, it may be programmed using both the popular Python and C++ languages, for which an Software Development Kit (SDK) is provided. Secondly, a graphical Integrated Development Environment (IDE) called *Choreograph* is also provided allowing less experienced programmers to harness the onboard resources and create behaviours for the robot with relative ease.

Aldebaran also support and develop (in house, having acquired the french robotics company, *Gostai*, in 2012) a custom programming solution in the form of the Universal Real-Time Behaviour (Urbi) middleware that is specialised for real-time behaviour orchestration and remote computation, with gearing toward *Cloud Computing* solutions for robotics This comes in the form of the cross-platform Urbi Software Platform³. The package is comprised of two components: Urbi middle-ware, which resides predominantly on the computer on-board the Nao (interfacing with NaoQi), but also on remote computational resources (i.e. a laptop computer). The second component is the *UrbiScript*⁴ programming/scripting language. In essence this provides the interface between the Urbi Middle-ware and the programmer. UrbiScript also introduces remote computational processes that are known as *UObjects*. Written in C++, these shared objects can be run either on the on-board processor, or be run remotely via a Wi-Fi network, allowing processed that are computationally demanding to be perfumed remotely on a remote computer with more computational resource, thus not allowing this process to become burden on the robots onboard computation resources. Finally, UrbiScript provides a new paradigm for programming in which parallel and serial processes can be run both in a synchronous, or an asynchronous manner, at the choice of the programmer, and is a feature that makes Urbi particularly promising for programming multi-modal interactive robots where behaviours need to be executed in a asynchronous manner.

³Urbi Software Platform: <http://www.urbiforge.org>

⁴UrbiScript is best described as a variation on *Python*, with the syntax of *C++*.

3.2.2 How Nao has been used

In this body of work informing this thesis, the Nao robot has been used as the social agent through which NLUs have been embodied.

With respect to the onboard resources, Nao was programmed using the Urbi solution which provides an interface to the NaoQi SDK. All the experiments presented (with the exception of the experiment in chapter 4) the Nao robot was either physically present in the room with subjects when playing NLUs (chapters 5, 6 and 7), or was video recorded with these videos being presented via internet based experiments (chapters 8 and 9).

In all of these experiments, the robot was programmed to behave in a manner where it exhibited natural-like behaviours. For example, the LED eyes blinked, the robot's weight was shifted from foot to foot, the robot gazed around the room and looked in the direction of the subjects when it was touched on the head (to play an NLU). The reason for doing this was to minimise the chance that the robot was perceived as being a static entity. Rather, it was desired that subjects perceived the robot as being socially competent by exhibiting lifelike, natural and basic social behaviour through its movement when idle and reacting both visually and audibly to physical interactions and external events in the environment (see chapters 8 and 9). Furthermore, all NLUs were generated and pre-recorded using the method detailed above, and played back through the onboard speakers, again reinforcing the notion that the NLUs were indeed embodied in, and made by the robot as *it* wanted to express something.

3.3 Measuring Affect

Much of this body of work is concerned with identifying how NLUs are able to convey affect to people, and particularly with respect to the NLU generation algorithm described in section 3.1, how the different parameters of NLUs generated using this impact the affective interpretation that is elicited by the utterances. As such, it is important to outline the approach that has been adopted here to

facilitate the capture subjects' affective interpretations, as there are many ways through which this can be done. This section does just that. It begins with a brief overview of the two schools of thought that surround how emotions and affect may be represented, as there are many different tools that have been developed, with their design being influenced by these schools of thought. This is followed by an overview of a collection of measuring tools that were considered for use in this research, with a discussion regarding their pros and cons. Finally, the measuring tool that has been adopted - the AffectButton - is described in relative detail, as the underlying design of the tool has impacts upon how the results of experiments conducted in the subsequent chapters have been performed and presented.

3.3.1 Representations of Affect

When it comes to representing emotions or affective states in synthetic systems/agents, there are generally two schools of thought that have been informed by the various theories on emotions: discrete categorical labels, and continuous dimensional affect spaces. For the interested audience, there are a number of detailed and rich reviews of the issues surrounding the world of emotional representation and measurements (e.g. Plutchik (1994) and Cowie and Cornelius (2003)).

3.3.1.1 Categorical Labels

Categorical labels (e.g. "happy", "sad", "angry", "scared", etc.) are the most familiar way in which people are able to relate and refer to different affective states due to their common everyday use in natural language. Given this, these labels are self-evidently, assumed to have a coherent understanding between people and are thus the easiest ways in which to describe different emotions and states (Cowie and Cornelius, 2003), and reflects the natural tendency for people to discretise their sensory input from the surrounding world into manageable chunks as outlined by James (1890). In the majority, focus of affective labels has been around what has been termed the "basic six" emotions (Schröder, 2001; Plutchik, 1994; Scherer,

1986; Banse and Scherer, 1996; Cowie and Cornelius, 2003), which has primarily been due to the prominent theories surrounding the notion of *basic emotions*: *happiness, sadness, surprise, fear, anger* and *disgust* (e.g. Ekman (1992) and Izard (2007)).

With respect to measuring emotion from humans and representing emotions in affective systems, there are a number of drawbacks. Firstly, given the links with natural language, the use of linguistic labels of measurement requires caution as, specifically in cases dealing with emotional human speech, these can serve as a considerable bias as the stimulus can carry semantic linguistic information regarding the emotional labels (Plutchik, 1994).

There is also an issue of resolution: emotional labels inherently do not provide a granular measure or indication how intense an emotion is. They are not able to capture subtle, but important differences between affective states. In natural language, a listener is able to utilise a variety of different cues regarding this through the nature of multi-modal interaction.

Finally, there is the issue of the number of labels that is to be used during measurement. In the case where there are only a few labels, which has been a common practice in a number of fields, the rating of stimuli is more akin to a discrimination task rather than an identification task (i.e. subjects are more likely to provide ratings based upon what the stimulus is *not*, rather than focusing upon what it *is*), as Banse and Scherer (1996) and Scherer (1986) have highlighted. This can be overcome by introducing many more affective labels (Schröder, 2001), however this can make the experimental process notably longer, but has had the benefit of allowing assessment of the how many different affective labels can be broken down into more fundamental underlying components such as affective dimensions, as demonstrated by Russell (1980) with the Circumplex model of emotions.

With respect to their representation in synthetic systems, emotional labels have the benefit in that each affective state that is modelled can have an *activation* level, which allows multiple affective categories to be active at the same

time, something that has been shown to be useful in the design of systems that recognises and represent multiple complex mental states from the human face for example (Kaliouby and Robinson, 2004). However, the inherent lack of granularity is also a problem in that it means that in the eyes of recognition systems, people can “jump” from state to state, which is not representative of the how the behaviour or mental state of a person changes as an interaction unfolds. Furthermore, in systems that are designed to express affect or just act upon it, categorical representations of affect with respect to the modelling of an input, or the internal state of the system itself can lead to large changes in behaviour due to the tendency to also jump between affective states, which is also generally undesired (Schröder, 2003b).

3.3.1.2 Dimensional Affect Spaces

Dimensional representations seek to identify ways in which emotional/affective states may be represented in continuous manner in spaces that have a small number of dimensions. There are multiple facets that make this approach appealing not only to the field of psychology, but also to fields concerned with creating synthetic systems that deal with affect (for example, the field of Affective Computing (Picard, 1997), and HRI (Breazeal, 2002)). For example, one of the main attractions is that dimensions provide a way in which affective states can be described in a more tractable manner, but can also be translated into and out of common verbal descriptions commonly used by people (Fontaine et al., 2007). This translation is possible as emotion related words can be mapped to different affective dimensions (e.g. Russell (1980)), and thus referred to specific locations within these dimensions (Cowie and Cornelius, 2003). Thus, dimensions are able to not only capture subtle differences in affect to a high resolution, but it is also possible to interpret the dimensions into more coarse regions which can form the basis of a categorical representation also (Schröder, 2004), making them useful when investigating what effects subtle changes to a stimulus (e.g. an emotional face, or a vocal utterance) has upon how people affectively interpret these (Cowie and

Cornelius, 2003). Furthermore, given that dimensions provide a numeric representation, they lend themselves to world of machine learning, which exploits a variety of mathematical tools used to manipulate numeric data.

This approach however is not without problems and shortcomings. Firstly, and perhaps more importantly, is that as with the basic emotion theories, there are disagreements with respect to both the number of dimensions an affect space should consist of, but also *what* the different dimensions represent. This is a practical problem in that in situations where there are only two dimensions, certain states such as *Fear* and *Anger*, and *Excitement* and *Surprise* are difficult to differentiate (Fontaine et al., 2007; Zeng et al., 2009). As such, this has resulted in a large number of different affect spaces, with ongoing debate as to which spaces are most optimal. An issue that still remains very much open (Cowie and Cornelius, 2003).

A further drawback is that in dimensional spaces, only a single affective state can be modelled at a given moment in time, whereas with categories, the number of states is determined by the number of categories, each of which can have a self contained level of *activation*. This means that if any co-occurring affective states arise simultaneously, only one of these may be represented in the affect space.

3.3.2 Capturing Affect from People

Capturing affect from human subjects comes in two flavours: *implicit* and *explicit* methods (Broekens et al., 2010; Isomursu et al., 2007). Implicit methods measure behavioural characteristics of a person (heart rate, respiration rate, skin resistance, etc, see Picard (1997) and Zeng et al. (2009) for overviews), while explicit methods require that subjects self report and input data directly (suggesting or choosing emotional labels or adjectives, selecting an emotional face, etc.). The body of work presented in this thesis has solely employed the latter, as not only would the use of implicit measures introduce a cumbersome aspect to the experimental process due to the requirement for placing many sensors in various locations on the body⁵,

⁵This has some very sensitive ethical issues associated with it when experimentation with children is conducted.

particularly with young children, but also it limits the amount of comparison that may be made the the related literature on both NLUs and gibberish speech in social agents (none of the previous work has used physiological measures of affect), but also work in affective expression via the human voice and music.

The previous section outlined the two main approaches that have been established with respect to how emotions can be represented in synthetic systems that have an affective component: categorical labels and affective dimensions, and discussed their respective benefits and drawbacks. These representations affect have also fed into the design of methods for how affect can be measured and captured from people.

A final noteworthy point to highlight is that while affective dimensions have a number of properties that make them appealing, particularly with respect to the goal of addressing the research questions outlined in chapters 1 and 2, measuring tools based around emotional categories are easy to explain to naive subjects both young and old, and tools based around affective dimensions are not. This is a weightily point to consider as some of the experiments undertaken in this thesis have been performed with children. As a result, time and care has been taken to familiarise the subjects with the measuring tools used, and to confirm that they indeed were able to the measuring tools in an appropriate manner.

There are a number of different tools that have been developed for affective measurement based around affect dimensions, namely the Self Assessment Manikin (SAM), (Lang and Bradley, 1994), FEELTRACE (Cowie et al., 2000), EMuJoy (Nagel et al., 2007) and the AffectButton (Broekens and Brinkman, 2009; Broekens et al., 2010; Broekens and Brinkman, 2013).

The SAM is a tool is a picture orientated tool that is designed to assess the Pleasure, Arousal and Dominance dimensions independently. Graphical images are shown to depict major points along each dimension. For the pleasure dimension, the images shown an agent with differing facial gestures ranging from a large happy smile to an unhappy frown. Arousal is depicted with a figure with a wide-eyed excited face to a sleepily and relaxed face. Dominance is shown with the

figure with varying physical size, which relate to the amount of control that the figure has with respect to the surrounding environment (the surrounding box in this case): a large figure translates to high control and thus dominance, while a small figure translates to the figure having little control.

FEELTRACE is a tool that was specifically designed for capturing peoples' affective ratings of emotional speech. It shows a two dimensional Activation-Valence space, with a number of different affective labels located in the space, to help guide the users understanding of the dimensions and how they relate to the affective labels that they are familiar with through the use of their natural language vocabulary. A coloured circular icon is controlled by a computer mouse and dynamically changes colour as it is targeted around the space following a specification of colour/affect mappings as proposed by Plutchik (1994). This tool also has a functionality that allows previous inputs to be stored and shown graphically so as to capture a *history* of affective measurements for a subject. This is partially useful when presenting subjects with stimuli that span across different time scales and that require more than just a *snapshot* in time measure.

EMuJoy is a tool that has been developed for capping peoples' affective ratings of musical pieces rather than speech. Though compared to FEELTRACE, that was designed for the measurement of affective speech, these two tools actually have a very similar end solution. With EMuJoy, again two dimensions are shown on screen, Arousal and Valence, with a cursor that shows the current position the two-dimensional affect space. The cursor takes the form of a small expressive face that dynamically changes as the cursor is moved around the input space, to represent the general affect of the current location. This tool also facilitates a history of affective measurements in the form of a *worm tail* which shows the previous inputs by the user in their chronological order.

The AffectButton is a tool that shows only an expressive face that changes dynamically as the mouse cursor is moved around the input space. Each face is also encoded into a three-dimensional coordinate where the dimensions correspond to Pleasure, Arousal and Dominance. What is unique about this tool compared to

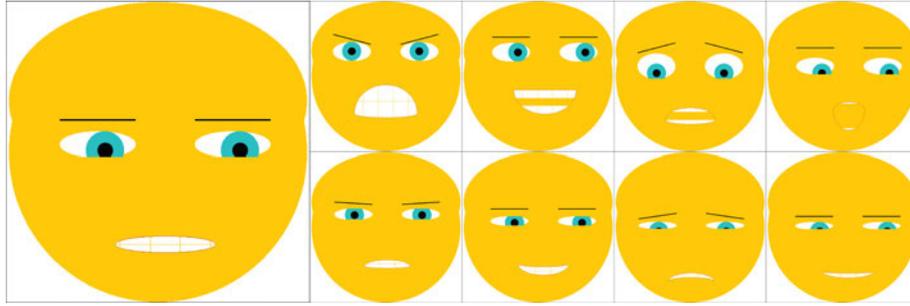


Figure 3.7: The AffectButton prototype facial expressions with PAD values. From left, clockwise: Neutral (0,-1,0), Angry (-1,1,1), Excited (1,1,1), Scared (-1,1,-1), Surprised (1,1,-1), Annoyed (-0.5,-1,0.5), Happy (0.5,-1,0.5), Sad (-0.5,-1,-0.5), Content (0.5,-1,-0.5). Adapted from Broekens et al. (2010).

the others outlined above is that the underlying affective dimensions are completely *hidden* from the subject, and thus there is no need to even mention the notion of affective dimensions to users. This is a key benefit (as discussed in the next section), and is why this tool was selected as the affective measuring tool of choice during the experiments presented in this thesis.

3.3.3 The AffectButton

The *AffectButton* (Broekens and Brinkman, 2009; Broekens et al., 2010; Broekens and Brinkman, 2013), figure 3.7, is an open-source facial gesture tool designed to facilitate the capture of affective interpretations from people, using explicit methods (i.e. people are directly asked to provide feedback).

To provide a high level description, the AffectButton is a tool that displays a simplistic cartoon like face on a laptop screen within a box with a mouse cursor. As the location of the mouse cursor changes, so too does the facial expression of the cartoon like face. Furthermore, co-ordinates of the mouse are also mapped to a single point co-ordinate within a three-dimensional *affect space* (PAD value), where the dimensions represent Pleasure, Arousal and Dominance. Pleasure relates to the positiveness verses negativeness of an affect, Arousal to the level of activation and Dominance to the degree that the environment is imposing influence. When a subject has selected their desired facial gesture, they can click the mouse and the PAD value is captured and stored.

The resulting numbers obtained from this mapping fall into the range $[-1\ 1]$

for each dimension respectively. These affect space triplets are represented by the dynamically changing expression on the face, allowing the user to select from a wide variety of different facial expressions and affect spaces values by moving the mouse within the box. The facial gestures are rendered in real time and therefore the user does not need to interpret the underlying affective dimensions, but rather provides an affective rating by selecting a facial expression that they feel matching their interpretation of a given stimulus.

There are 9 prototype facial gestures located within this affect space (figure 3.7), each corresponding to an affective label: happy, excited, annoyed, angry, sad, scared, content, surprised and relaxed (these labels are exemplary), as the affect triplet changes, the facial expression displayed interpolates linearly between these nine prototype expressions. This is done via a mechanism that is comparable to that used in the robot Kismet (Breazeal, 2002; Broekens and Brinkman, 2013).

3.3.3.1 Mapping 2D Input to 3D Output

As outlined above, the AffectButton essentially provides a mapping between a two dimensional input space (the laptop screen) and a three-dimensional affect space, which in turn is used to determine the facial expression that is displayed on the screen. The purpose of this section is to detail this internal mechanisms and rules that determine this 2D to 3D mapping. The reason for this is that this has an impact upon how the results in the aforementioned chapters have been analysed (i.e. the statistical tests that have been employed) and presented graphically.

The button consists of two parts: an outer border and an inner border, both of which are square in shape. The outer border defines the working range of the mouse cursor. The inner border spans from -0.55 to 0.55 along both the horizontal and vertical components of the input space. The horizontal (x -axis) and vertical (y -axis) components of the cursor location within the outer border are directly mapped to the Pleasure and Dominance dimensions of the affect space respectively, and thus are controlled independently of each other. These values are then scaled such that they fall within the range $[-1\ 1]$.

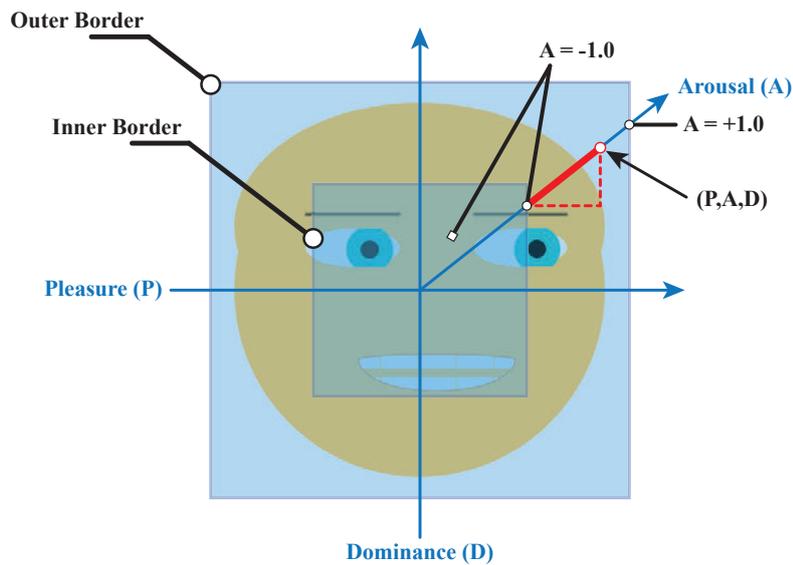
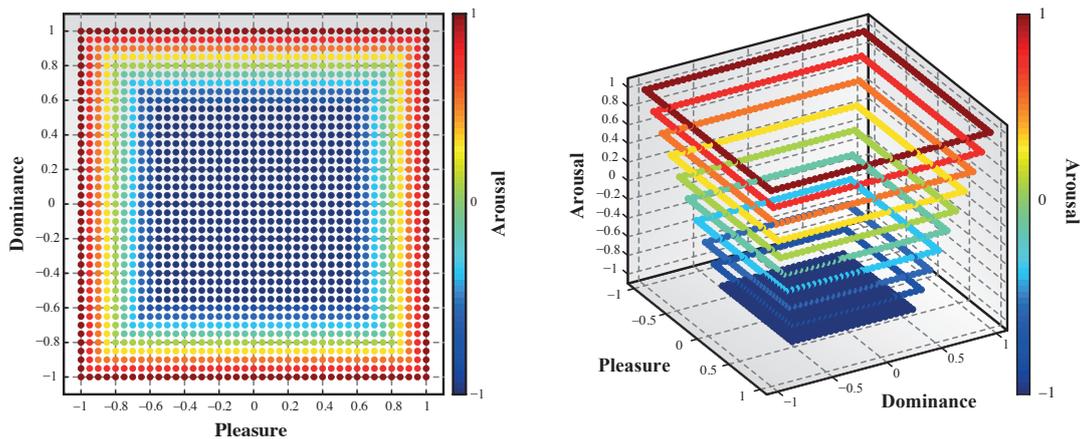


Figure 3.8: Example of how the Arousal value is calculated from the Pleasure and Dominance values in the AffectButton. Image adapted from Broekens and Brinkman (2013).

The Arousal value is a derived form the Pleasure and Dominance values, and is in essence the *radial* distance from the centre of the input space to the location of the mouse cursor. While the cursor location lies *within* the inner border, the Arousal value is held constant at -1 . When the cursor is located *outside* the inner border, the Arousal is lineally interpolated between -1 (the start of the inner boarder) and 1 (the start of the outer border), based upon the distance to the outer border (Broekens and Brinkman, 2013). More formally, it is characterised as the *hypotenuse* of the triangle that is formed from the edge of the inner border, the x co-ordinate of the cursor, and the y co-ordinate of the cursor. This is shown in figure 3.8, and the algorithm used to calculate the PAD values from the mouse coordinates and check the PAD values are outlined in algorithms 8 and 9 in Appendix A.

Given this description, there needs to be a clarification regarding the exact nature of the affect space, and the different PAD values that can be located in this space. While the Pleasure and Dominance dimensions are independent of each other, the Arousal is not an independent dimension - it is derived from the other two dimensions. As a result, the PAD values cannot be located *anywhere* within the cubic space defined by the working range of the three dimensions. Rather,



(a) 2D Plot of the possible PAD values along the Pleasure (x) and Dominance (y) dimensions.

(b) 3D plot of the possible PAD values. The mapping of these values to a Trapezoid surface can be seen.

Figure 3.9: Front and Side plots of the possible PAD values in the AffectButton PAD space.

PAD values that can be obtained via this mapping fall onto a trapezoid surface within this space, as shown in figure 3.9.

3.3.3.2 How the AffectButton has been used

Within the context of this work, the AffectButton has been used with both child and adult subjects to capture affective interpretations of NLUs, both in formal lab settings, and in more “wild” settings such as primary school classrooms. Specifically, the experiments presented in chapters 5, 6 and 7. It has also been reimplemented from the source Python code into C++ such that it could be integrated into the various pieces of software that were developed to conduct the various experiments in the aforementioned chapters.

The rationale for using this tool over the others is three-fold. Firstly, the tool provides an intuitive manner through which subjects are able to express their affective interpretations: via facial expressions. Secondly, as the Nao robot does not afford an expressive face and only acoustic stimuli have been used in this research, the use of facial gestures does not present a conflict between the modality through which the robot is expressing itself, and the modality through which the ratings of this expression are captured. Finally, from a practical perspective, given that the affective dimensions underlying the button are *hidden* from the subject,

one does not need to be concerned with the cumbersome task of describing the nature of these dimensions, and thus the experimental process is made easier. This latter point is particularly appealing as describing how affective dimensions relate to the subjects' affective interpretation of a stimulus is a tricky business and can be the source of considerable confusion, even for adults, let alone children.

However, while the AffectButton has been validated in a number of different ways, as described by Broekens and Brinkman (2013), this validation does not include young children. Thus, it cannot be assumed that they understand how to use the tool in the same way that teenagers and adults do. To address this, during experiments in which the AffectButton has been used in this thesis, care has been taken to allow subjects (both young and old) to become familiarised with how the tool works (i.e. how to move the button to produce different facial expressions) and the variety of different facial gestures that can be produced. This has been done via two methods (after subjects have been given time to explore the button). Either by asking subjects to assign facial gestures for different affective labels (chapters 6 and 7), or by presenting subjects with each of the prototype facial expressions, and asking subjects to move the mouse cursor the location in the button such that the AffectButton matched the prototype that has been presented (chapter 7).

Finally, a note on how the results obtained from experiments using the AffectButton are presented. Given that the Pleasure and Dominance dimension are the only two independent dimensions, with the Arousal values being derived from these, graphical plots of the PAD values are shown as two dimensional plots with Pleasure being represented as the horizontal dimension, and Dominance as the vertical dimension (see figure 3.9a). This is the same representation as is used in the button, and thus makes it easy to identify exactly where on the screen a subject clicked to capture an affective rating, and has the aim of making the graphs more intuitive to understand in this respect.

3.4 Summary

This chapter has served to provide details regarding methodological tools that have been employed in during the work informing this thesis. Firstly, a custom method for describing and characterising NLUs was presented. This makes it possible to create NLUs that have a sentence-like structure and is characterised using parameters that are analogous to those that are used to describe the acoustic correlates of affective expression in both the human voice and in music. Next, the Nao humanoid robotic platform was described as was its use in this work as the sole platform through which NLUs were embodied and studied. Finally, following a brief description and discussion regarding the two main schools of thought regarding representations of affect categories and affective dimensions, issues surrounding the measurement of affect and tools developed to do this were presented. This chapter ended with a detailed description of the affective measuring tool of choice, the AffectButton, and how it has been used in this work.

Chapter 4

Alignment of NLUs with Agent

Morphology

Summary of the key points:

- An online experiment is conducted to examine whether the morphology of a robot biases how people interpret the affective meaning, intentional meaning and appropriateness of utterances that the robot makes.
- People were shown an image of a Nao humanoid robot and an Aibo dog robot, and heard either a human-like utterance, animal-like sound or an NLU and were asked to rate these with respect to affect, intention and appropriateness.
- People are not coherent in the interpretations the affective or intentional meaning of utterances, and the morphology does not matter in this regard.
- There does need to be an alignment between the type of vocalisation a robot makes and the physical morphology of the embodiment in order for people to deem the combination as *appropriate*.
- People deem it acceptable for the Nao robot to make NLUs. This serves as a justification for the use of the Nao as the robot in which NLUs are embodied and studied in this thesis.

As this work has proposed the use of a Nao humanoid robot as the platform in which NLUs are to be embodied, it is important to confirm that users deem this combination of the embodiment and NLUs as *appropriate*. This issue holds weight as research has revealed that a miss alignment between a robot’s morphology and its behaviour can lead to adverse reactions to the robot (MacDorman and Ishiguro, 2006). This alignment is the focus of the *Uncanny Valley* hypothesis (figure 4.1) proposed by Mori (1970) which states that as the anthropomorphism of an agent converges to that of a human being, the reaction of users will tend toward an *affinity* with the agent (Moore, 2012), until a point where the physical resemblance of the agent is such that it begins to evoke an adverse response, due to aspects of the appearance and behaviour differing from the human norm provoking a sensation of strangeness (MacDorman and Ishiguro, 2006). As such, Mori proposed that robot designers use the hypothesis as a guideline for when designing robots, encouraging the design of robots to reside on the left side of the valley, rather than striving to create robots with a high degree of human resemblance.

While this hypothesis, in its more famous manifestation, applies primarily to physical morphology and movement of an agent, research from the field of computer animation has shown that there are also cross-modal effects in which vocalisations also have influence (Tinwell et al., 2011). Similarly Mitchell et al. (2011) created videos with combinations of visual/audio pairs with robot and human faces, and human and synthesised speech, finding that cross-modality mismatches resulted in a greater sense of *eeriness*, concluding that the physical and acoustic aspects of the robot should “match”. While the Uncanny Valley is not universally accepted in the current form (Tinwell et al., 2011; Bartneck et al., 2009), the basic premise that there is a relation between the degree of anthropomorphism, behaviour and user perception remains a relevant issue in HRI. Komatsu and Yamada (2011) provide a tangible example of this work, having found that subjects have significant differences in their interpretations of their *artificial sounds* depending on the physical appearance of an agent. The findings supporting the

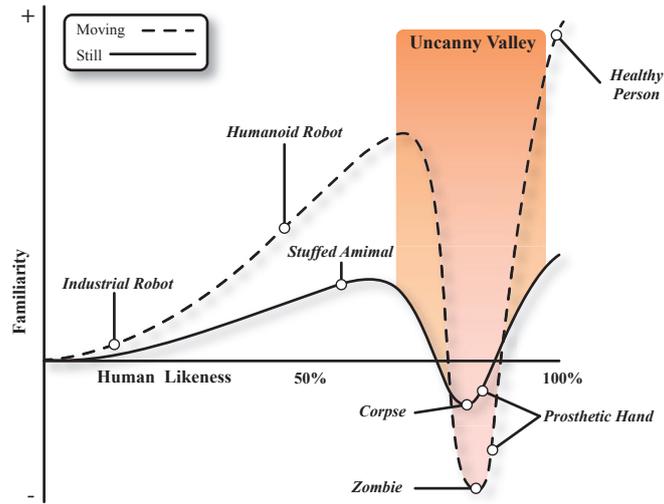


Figure 4.1: Graph of relationship between human likeness and perceived familiarity, as proposed by Mori (1970): familiarity increases with human likeness until a point is reached where subtle deviations from the human appearance evoke an adverse response. This is known as the *Uncanny Valley*. Figure adapted from MacDorman and Ishiguro (2006).

Uncanny Valley hypothesis strengthen the need to test the perception of the morphology/NLU alignment as applied to the Nao platform in order validate the use of the platform. Furthermore, the findings that a different embodiment may evoke a different interpretation of the same utterance (Komatsu and Yamada, 2011), and that the morphology of the embodiment alone can evoke substantially different reactions from subjects (Hwang et al., 2013) highlights to need to retain the same robot throughout the body of this research.

To this end, this chapter presents an experiment aimed at probing how the morphology of a robot influences the perception of appropriateness and the affective interpretation of NLUs, in comparison to more characteristic forms of utterance that may be associated with the particular morphology.

4.1 Experiment Setup

This experiment set out to test the following hypotheses:

H_1 : Users are coherent in their affective interpretations of utterances made by a robot.

H_2 : The physical appearance of a robot has an influence upon the interpretations.

H_3 : The physical appearance of a robot has an influence upon the perception of appropriateness of utterances made by the robot.

These hypotheses were tested through an online experiment, where subjects were asked to rate image and utterance stimulus pairs in terms of their emotional and intentional interpretations of utterances as well as provide a rating of the appropriateness of each stimulus pairing. In total, 20 utterances were collected together - 5 utterances recorded from a human source, 6 utterances from an animal source, and 9 sounds recorded from technology sources (e.g. analogue computers, mobile phones, etc.). Each acoustic stimulus was presented to subjects twice, once paired with an image of a Nao robot and once with an image of a Sony Aibo robot (figure 4.2), thus in total, 40 stimulus pairs were presented to subjects. For each utterance pair, subjects were asked to select from a list of which emotion they felt that the robot had conveyed through the utterance. They were also asked to guess the communicative intent, and to judge the appropriateness of the utterance and robot image pairing. All responses were forced choice. To conclude the experiment, subjects were asked whether they were pet-owners and if they came into contact with robots on a regular basis.

Two versions of the experiment were created with counterbalanced robot-utterance stimulus presentation, and subjects were directed to a URL that forwarded them to one of the two experiments.

4.1.1 Utterance Stimuli

The 20 utterances were grouped into three broad categories: human (6), animal (5) and technological (9). Audio samples were collected from a variety of sources including the FreeSound¹ online data base, and self recordings. No pre-testing of the sounds in order to ascertain rough affective interpretations was performed.

Human utterances consisted of recordings of utterances such as “hmm”, “ahhh”, and other recordings of sounds that can be produced using the human vocal tract. Animal utterances consisted of sounds such as a cat’s purr, or a small dog growl-

¹www.freesound.org

ing, and were selected to cover a small range of sounds that one might expect to hear from such animals. Technological utterances (which are essentially NLU) came from a broad range of sources, such as mobile phones and unusual daily sounds such as windows being wiped clean.

The motivation behind this selection of utterances was to capture a broad variety of stimuli with respect to sound source and acoustic parameters (intonation, pitch, speed, ect.), rather than providing systematic and controlled differences in acoustic profile. Understanding correlates in acoustic features in utterances and their interpretations was not the focus of this experiment. Rather, the focus was to query the impact of varying agent morphologies and validate that the Nao is indeed an appropriate platform on which to conduct NLU research. As such the selected utterances were not intended to portray any particular affective states or have any particular meaning.

The 40 stimuli were presented in a pseudo-random order, with the constraint that the repetition of each utterance was to be separated by at least 14 others, all of which were to be different. Doing this avoided the sequential repetition of an utterance, with the aim to minimize the chance that subjects would not only recognize the acoustic stimulus, but also recall the response that they provided with the first presentation.

4.1.2 Visual Stimuli

The two robots to be presented were carefully considered due to the wide variety of robots currently available, both in the commercial and research domains. The primary concern was to make a comparison between two robots that have the same design theme, and avoid introducing unnecessary noise to the results by comparing two robots with differing underlying aesthetic design themes. Robots such as the Paro (Wada and Shibata, 2006) or MIT's Leonardo (Coradeschi et al., 2006) are designed to resemble living creatures (evident through their *soft, furry* exteriors), while robots such as the Nao or Aibo have a more prominent industrial design theme and resemble technological artefacts (evident through their rigid,



(a) Sony's Aibo dog robot.



(b) Aldebaran's Nao humanoid robot.

Figure 4.2: Images of the two robots used in the experiment.

plastic exteriors). If comparison were made between the Nao and Paro platforms for example, this would not only be testing morphology, but life-like aesthetic also, and thus would be a less fair comparison and may have confounded results.

Making a comparison between the Nao and the Aibo was deemed suitable on three rationale. Firstly, both the robots have similar aesthetic characteristics providing a fairer comparison. They are both very clearly robots, evident through the plastic exterior². Furthermore, they both facilitate capabilities for displaying affect through an array of LED's that represent eyes, reinforcing the technological theme of the design. Secondly, both platforms are or have been commercially available (though the Aibo has been discontinued for some time, they can be purchased from various second hand sources), thus are likely to be known to a wider audience, however, this was not assumed. Finally, since the two robots represent what may be considered technological artefacts, it was deemed likely that users would be open to the interpretation of a broad range of utterances exhibited due to milder preconceived expectations³ (Komatsu and Yamada, 2007).

²Robots with a soft fur exterior may cause confusion with cuddly toys in static images.

³Though there may be an influence due to popular Culture and media.

4.1.3 Emotional States

Subjects were presented with a forced choice of 9 affective labels: *happiness*, *sadness*, *relaxation*, *anger*, *affection*, *fear*, *interest*, *boredom* and *disgust*. This selection differs from the affective labels that are commonly used in the literature for studies addressing the recognition of emotional content in acoustic cues and speech, from both robots and humans. Most studies employ the “basic emotions” (Plutchik, 1994): *joy*, *sadness*, *fear*, *anger*, *disgust* and *calmness* (*c.f.* Breazeal (2002); Bryant and Barrett (2008); Oudeyer (2003)). While these basic labels provide subjects with a short and intuitive list to choose from, there are drawbacks. Scherer (1986) highlights that using a small number of labels studies *discrimination* rather than *recognition* between the options. There is also the concern of familiarity and alignment of the labels understanding between the subjects - are all the subjects familiar with the various labels presented, and do the labels have the same meaning to each of the subjects?

Each of the states may also be loosely paired with another so that one falls into a “positive” classification and the other into a “negative” classification (table 4.1). This has been done in an attempt to maintain an even balance between “positive” and “negative” options. In the case of disgust, at the time of the study design, no basic affective state could be readily identified that would produce a suitable counterpart to make a pairing.

Table 4.1: Pairing of Emotion/State Classifications.

Pair	Positive	Negative
1	Happiness	Sadness
2	Relaxation	Anger
3	Affection	Fear
4	Interest	Boredom
5	-	Disgust

4.1.4 Communicative Intents

Aside from querying affective interpretation, subjects were also asked to provide an intentional interpretation. Five intentional categories were selected for this and

presented in a forced choice paradigm: approval, attention, prohibition, comfort and neutral. These are the basic communicative intents that have been employed in adult and infant-directed speech in both HRI studies (Breazeal, 2002) and psychological studies (Bryant and Barrett, 2007; Fernald, 1989).

4.2 Results

In total, 61 participants responded to the study, 27 males (mean age = 26.66, std = 8.3) and 34 females (mean age = 35.65, std = 12.03). There were 32 pet owners (13 male, 19 female). 12 participants (11 male, 1 female) reported that they used robots on a regular basis. 55 participants reported that they lived in the United Kingdom, 46 of whom were British natives.

Due to the use of forced choice nature of this experiment, there are limits as to the statistical methods that may be employed. Krippendorff's α (Hayes and Krippendorff, 2007) was used as a measure for Inter-Rater Reliability (IRR) as this metric caters for nominal data. The α holds a value between -1 (for complete *disagreement* between subjects) and $+1$ (for complete *agreement* between subjects). One-way χ^2 tests were performed to determine whether individual ratings (e.g. the appropriateness ratings for the Human Utterance/Nao robot pairs) were statistically above chance levels. Two-way χ^2 tests were used to check for differences between subject sub-groups (gender, and pet ownership) while Stewart-Maxwell tests were employed to test for differences between the different robot morphologies (as all subjects provided ratings for both conditions).

This section will present an overview of the results focusing on the affective, intentional and appropriateness ratings individually, providing an overview of the overall ratings, comparisons between subject groups and the robot images, and the subject agreement based on Krippendorff's α values.

4.2.1 Affective Ratings

Figures 4.3a, 4.3b and 4.3c show the overall percentage of affective ratings for each of the different affective categories, for each of the utterances classes respec-

Table 4.2: Krippendorff’s α values showing the agreement between subjects in their affective (Happiness, Sadness, Relaxation, Anger, Affection, Fear, Interest, Boredom and Disgust) ratings of the different classes of utterances.

Robot	Subjects	Utterance Class		
		Human	Animal	Tech
Nao	Overall	0.2922	0.0948	0.1271
	Females	0.3250	0.1101	0.1369
	Males	0.2529	0.0964	0.1169
	Non-Pet Owners	0.2878	0.1204	0.1033
	Pet Owners	0.2591	0.0910	0.1419
Aibo	Overall	0.3196	0.0955	0.1243
	Females	0.3370	0.0882	0.1341
	Males	0.3035	0.1218	0.1192
	Non-Pet Owners	0.2911	0.0914	0.1034
	Pet Owners	0.3458	0.0945	0.1381

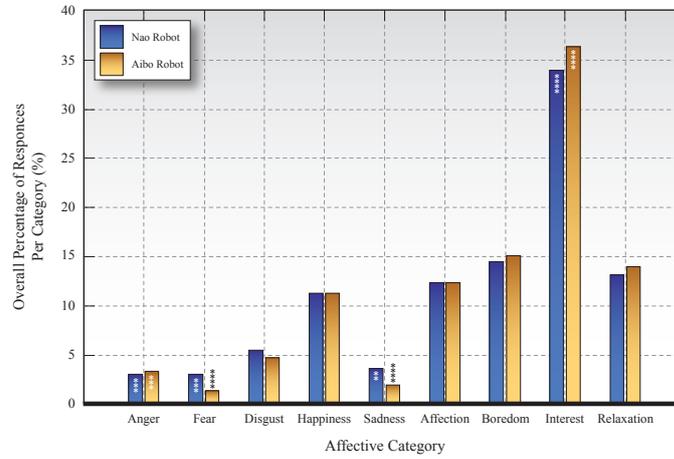
tively (these results are also summarised in table B.1), indicating the frequency of different rating categories for class of utterance. The graphs also indicate whether each of the ratings were statistically above chance levels (50%), and if so, to what degree of statistical significance.

It can be seen that in the majority of cases, the distribution of ratings are not statistically above chance level (see table B.1). The graphs also reveal that the distributions of ratings between each of the utterance classes are different, however this is to be expected as the stimuli were not intended to portray any particular affect.

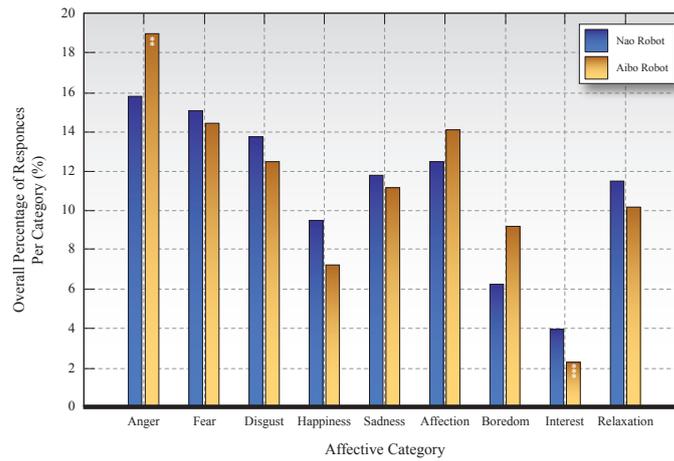
Krippendorff’s α values were calculated for the affective ratings, for all the subjects as a whole, and for the subgroups (males, females, pet owners and non-pet owners), for each robot individually. The calculated values (table 4.2) reveal that overall, there is little agreement between subjects in their affective interpretation of the utterances. This corresponds with the general trend for ratings to be at chance level also. It is also interesting to see the general trend that the Human class of utterances has the highest α values, and that the Animal class of utterances has the lowest. However, all values are indeed low.

Figure 4.3: Overall Percentage of Affective Ratings across both Robots and Utterance Categories (within bar markings indicate statistical significance that is above chance).

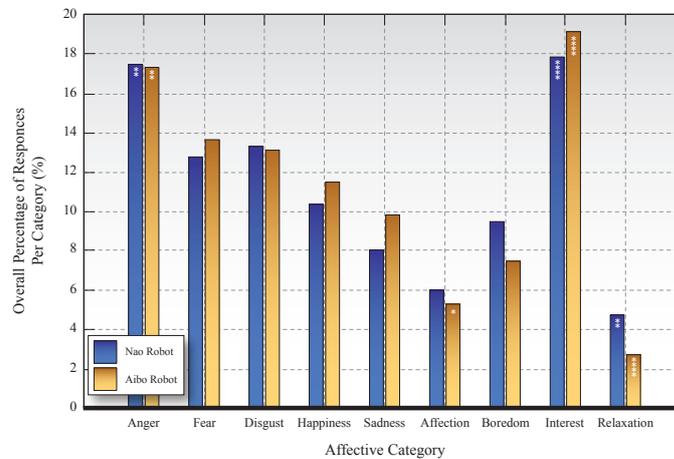
* : $p < 0.05$, ** : $p < 0.025$, *** : $p < 0.01$, **** : $p < 0.005$. See table B.1.



(a) Human Utterances



(b) Animal Utterances



(c) Technological Utterances

4.2.2 Intentional Ratings

Similar to the affective ratings, the intentional ratings were also generally at chance level. Figures 4.4a, 4.4b and 4.4c show the distributions of ratings spread across the 5 intentional labels from all the subjects for both the Nao and Aibo robots, for each of the utterance classes respectively (these plots are summarised in table B.4).

It can be clearly seen from the graph that overall the utterances were predominately rated as having an “Attention” intention at levels that are above chance for all the utterance classes, while there are also labels that received low percentages at levels that are also above chance indicating that subjects did not feel that the stimuli conveyed these intentions, namely the “Prohibition”, “Comfort” and “Approval” labels (see table B.4 for χ^2 and p values).

Stewart-Maxwell tests were performed to check for differences in the overall distributions across the labels due to the robot image that was presented with the utterances, revealing that there were no significant differences due to the robot image for any of the utterance classes. These tests revealed no differences between the ratings due to the different images that were presented with the utterances, with this being true for all the utterance classes (see table B.4).

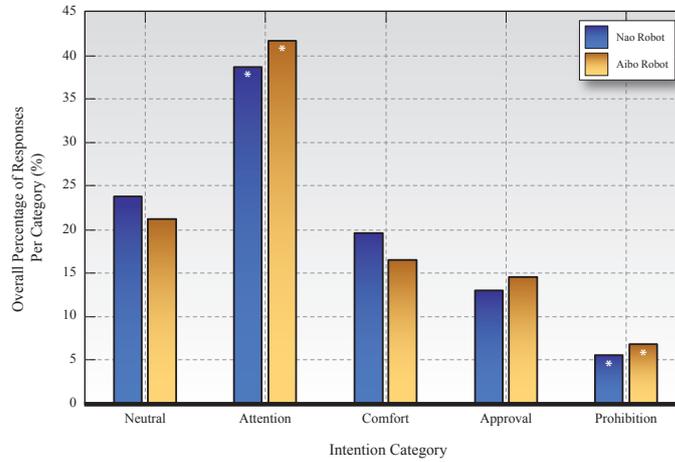
The Krippendorff’s α values (table 4.3) are again low, indicating that there was little agreement between subjects in their ratings. Similarly to the Affective ratings, this is reflected in the general chance level ratings that have been found.

4.2.3 Appropriateness Ratings

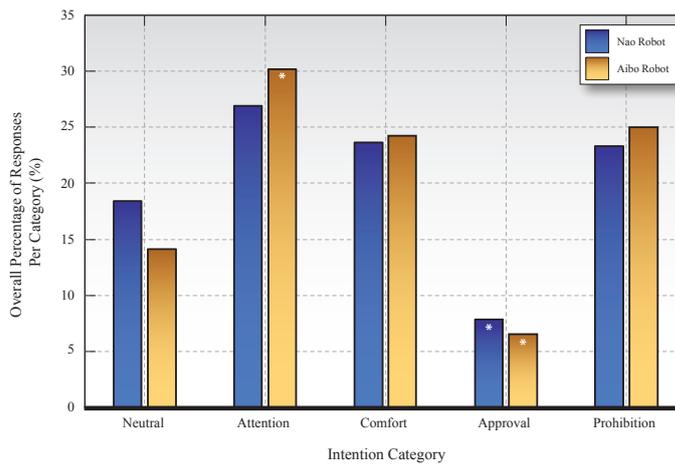
Figure 4.5 plots the percentage of “Yes” responses to the appropriateness of each stimulus pair. An initial visual inspection shows that in the case for the Human and Animal utterance classes, subjects showed higher approval ratings for the utterance class that matched the morphology of the robot - human utterances were deemed as more appropriate when presented with the Nao (74.5%) robot than the Aibo (57.9%), and animal utterances were deemed more appropriate when presented with the Aibo (68.2%) robot than the Nao (46.2%). In the case

Figure 4.4: Overall Percentage of Intention Ratings across both Robots and Utterance Categories (within bar markings indicate statistical significance that is above chance).

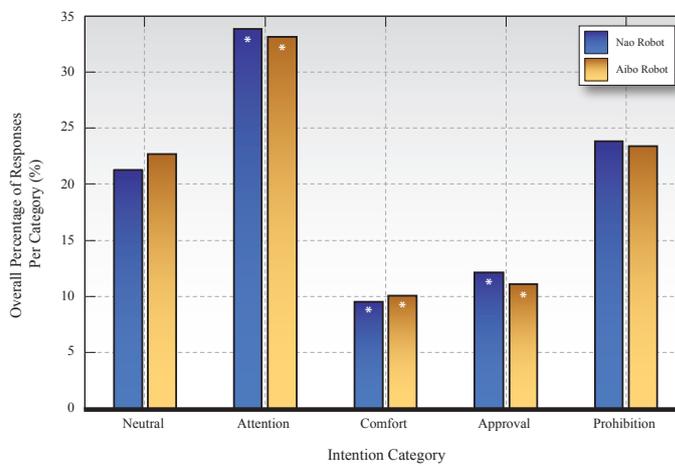
* : $p < 0.005$. See table B.4.



(a) Human Utterances



(b) Animal Utterances



(c) Technological Utterances

Table 4.3: Krippendorff’s α values showing the agreement between subjects in their interpretation ratings (Approval, Attention, Prohibition, Comfort and Neutral) of the different classes of utterances.

Robot	Subjects	Utterance Class		
		Human	Animal	Tech
Nao	Overall	0.0875	0.0612	0.0834
	Females	0.0675	0.0590	0.0833
	Males	0.1229	0.0585	0.0702
	Non-Pet Owners	0.0977	0.0507	0.0818
	Pet Owners	0.0776	0.0671	0.0796
Aibo	Overall	0.0735	0.0567	0.0622
	Females	0.0801	0.0397	0.0545
	Males	0.0722	0.0968	0.0787
	Non-Pet Owners	0.0440	0.0649	0.0392
	Pet Owners	0.0964	0.0683	0.0756

of technological utterances, as in the subjects rated these are more appropriate for the Nao (59.7%) than the Aibo (54.8%) also, but with a far smaller margin of difference. As some of the percentages lay near 50%, one-way χ^2 squared tests were performed to test the likelihood that any of the results were due to chance. The ratings for the Animal utterances presented with the Nao robot and the Technological utterances presented with the Aibo robot were both found to be due to chance, while all other ratings were found to be above chance (see table B.7, Chi Squares Test column).

McNemar⁴ tests were used to test for statistical significance in the differences between the appropriateness ratings between the robot type. It was found that there were significant differences in the ratings due to the robot morphology of the Human utterances ($\chi^2(1, N = 61) = 27.480, p \leq 0.005$), Animal utterances ($\chi^2(1, N = 61) = 34.299, p \leq 0.005$) and Technological utterances ($\chi^2(1, N = 61) = 4.198, p \leq 0.05$). These are summarised in table B.7.

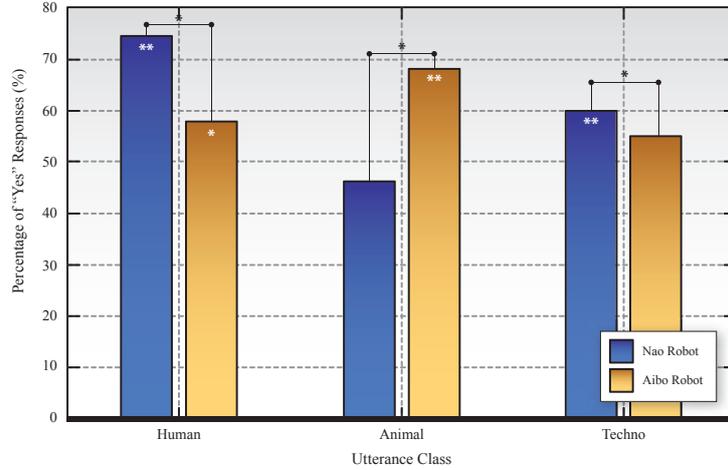
4.2.3.1 Gender Differences

From figure 4.6 and table B.8 it can be seen that both genders rated Human-Nao pairs as appropriate (males = 73%, females = 75%) to a degree that was above chance, and that there was no significant difference between the genders

⁴The McNemar test is similar to a Stewart-Maxwell test, but only caters for two factors with two conditions, while Stewart-Maxwell tests cater for two factors with multiple conditions.

Figure 4.5: Percentage of Appropriate Responses for the Robot Types Across the Utterance Classes.

* : $p < 0.05$, ** : $p < 0.001$



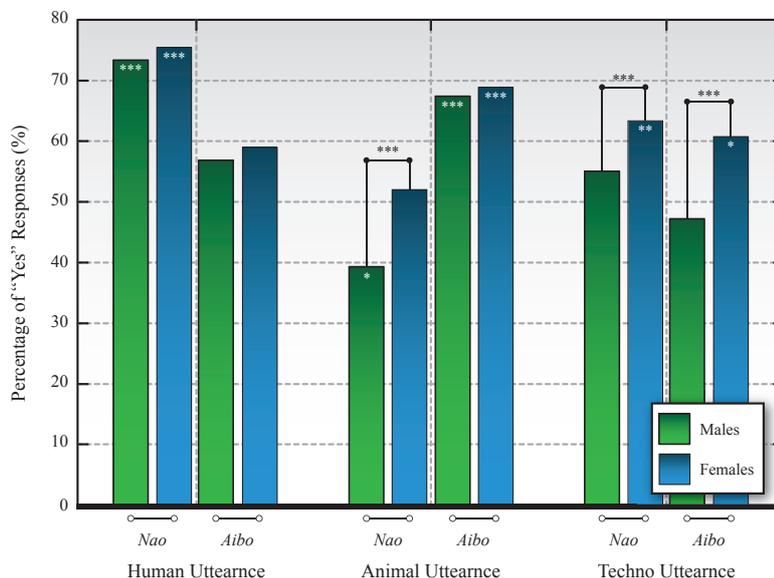
($\chi^2(1, N = 61) = 0.197, p > 0.05$). Human-Aibo pair ratings (males = 56%, females = 58%) were found to be at chance level for males ($\chi^2(1, N = 27) = 1.494, p > 0.05$) and females ($\chi^2(1, N = 34) = 2.526, p > 0.05$) with no significant differences ($\chi^2(1, N = 61) = 0.153, p > 0.05$).

Animal-Nao pairs were rated as less appropriate by males than females (males = 39%, females = 53%) and this difference was found to be significant ($\chi^2(1, N = 61) = 6.361, p < 0.05$). However, the male ratings were found to be above chance ($\chi^2(1, N = 27) = 3.316, p < 0.05$) while the female ratings were not ($\chi^2(1, N = 34) = 2.172, p > 0.05$). Similarly to the Human-Nao pairs, Animal-Aibo pairs were rated as appropriate (males = 67%, females 68%) with no significant difference between the ratings ($\chi^2(1, N = 61) = 0.070, p > 0.05$), with the ratings being above chance for both the males ($\chi^2(1, N = 27) = 8.182, p < 0.05$) and females ($\chi^2(1, N = 34) = 10.552, p < 0.05$).

In the case of Techno-Nao pairs, it was found that the females rated these pairs as more appropriate than males (males = 55%, females = 63%) and that this difference was statistically significant ($\chi^2(1, N = 61) = 3.847$), however, the male ratings were at chance level ($\chi^2(1, N = 27) = 1.286, p > 0.05$) while the female ratings were above chance ($\chi^2(1, N = 34) = 7.235, p < 0.05$). A similar trend was also found for the Techno-Aibo pairs where a significant difference ($\chi^2(1, N = 61) = 9.907, p < 0.05$) between the genders was observed (males

Figure 4.6: Percentage of “Yes” Responses for the male and female subjects across the robot types and utterance classes. Plot also shows whether results are due to chance (marked within each bar), and whether there are significant differences between the genders (marked between two bars).

* : $p < 0.05$, ** : $p < 0.01$, *** : $p < 0.005$



= 47%, females = 60%). However the male ratings were again found to be at chance level ($\chi^2(1, N = 27) = 0.348, p > 0.05$) while the female ratings were not ($\chi^2(1, N = 34) = 4.449, p < 0.05$).

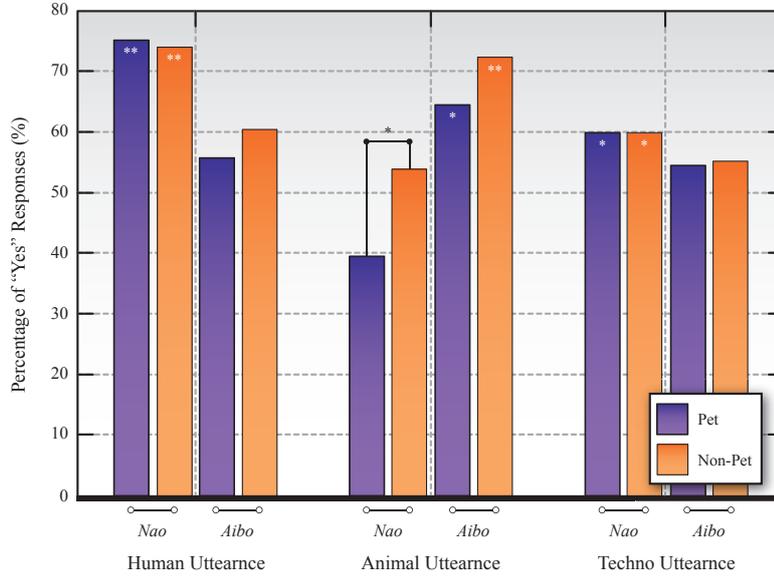
4.2.3.2 Pet Ownership Differences

Form figure 4.7 and table B.9 it can be see that overall, there was little difference in how the pet and non-pet owners rated the appropriateness of the utterance classes across the different robot morphologies. In the case of the Human-Nao pairs, both pet (rating = 55%, $\chi^2(1, N = 32) = 24.00, p < 0.025$) and non-pet owners (rating = 60%, $\chi^2(1, N = 29) = 22.321, p < 0.025$) gave similar ratings that were both found to be above chance and were not significantly different. The ratings for the Human-Aibo pairs were notably lower also with no significant differences, but were not above chance level.

In the case of the Animal-Nao pairs, pet owners rating these are less appropriate than non-pet owners, to a statistically significant degree, however, both the ratings were not found to be above chance, casting doubt upon the significance of

Figure 4.7: Percentage of “Yes” Responses for the pet owners and non-pet owners subjects across the robot types and utterance classes. Plot also shows whether results are due to chance (marked within each bar), and whether there are significant differences between the subjects groups (marked between two bars).

* : $p < 0.025$, ** : $p < 0.005$



the difference found ($\chi^2(1, N = 61) = 6.361, p < 0.025$). Animal-Aibo pair ratings followed the same trend as the gender difference, both groups ratings were above chance and were not significantly different ($\chi^2(1, N = 61) = 2.266, p > 0.05$).

Techno-Nao pairs were rated as appropriate at above chance levels by both pet (rating = 59%, $\chi^2(1, N = 32) = 5.444, p < 0.025$) and non-pet owners (rating = 59%, $\chi^2(1, N = 29) = 5.236, p < 0.025$), with no significant differences between the groups ($\chi^2(1, N = 61) = 0.000, p > 0.05$). Techno-Aibo pairs were found to be at chance level for both pet (rating = 54%, $\chi^2(1, N = 32) = 1.174, p > 0.05$) and non-pet owners (rating = 55%, $\chi^2(1, N = 29) = 1.274, p > 0.05$), with no significant differences ($\chi^2(1, N = 61) = 0.024, p > 0.05$).

4.2.3.3 Inter-Rater Agreement

Table 4.4 shows the Krippendorff’s α values for the appropriateness ratings for the different utterances classes, robot images and subject groups. All the values are low and near a 0 value, and thus indicate that there was a low overall agreement between subjects in which individual utterances they found to be appropriate for

Table 4.4: Krippendorff’s α values showing the agreement between subjects in their judgement of the appropriateness of different classes of utterances with the robot image.

Robot	Subjects	Utterance Class		
		Human	Animal	Tech
Nao	Overall	0.0469	-0.0066	0.0825
	Females	0.0413	-0.0120	0.1220
	Males	0.0343	0.0181	0.0250
	Non-Pet Owners	0.0187	-0.0175	0.0692
	Pet Owners	0.0533	-0.0214	0.0748
Aibo	Overall	0.0065	0.0673	0.0880
	Females	0.0053	0.0610	0.0658
	Males	0.0096	0.0938	0.1072
	Non-Pet Owners	-0.0046	0.0865	0.0829
	Pet Owners	-0.0046	0.0346	0.0804

a given robot image, regardless of the subject group.

4.2.4 Summary of Results

The results have shown that overall, subjects were not coherent in their ratings regarding the affective meaning of utterances, or the intentional meaning of utterances, and while the distributions of the ratings were different across the three utterance classes, the majority of ratings were not above chance levels (thus they can be said to be random). Furthermore, there were no significant differences found in ratings due to the different robot morphologies, nor were there significant differences found between the two genders in their ratings, or between subjects who did or did not have a pet.

With respect to the appropriateness ratings, it was found that when the utterance class was aligned with the morphology of the robot (Human utterances presented with the image of the Nao robot, and Animal utterances presented with the image of the Aibo), the ratings were significantly higher than when the morphology and utterance class were misaligned. In the case of the Technological utterances, subjects rated the combination with the Nao robot as more appropriate than the combination with the Aibo robot. The results also show that females tended to have a higher appropriateness rating than male subjects.

4.3 Discussion

It was unexpected to find little coherence between subjects in their responses. These results falsify the hypothesis (H_1) that users are coherent in their interpretations of utterances. Explanations for this are that the lack of context throughout the experiment made it difficult for subjects to assign affective states to what may be considered rather abstract sounds. If this is the case, it highlights the important role that context may play in helping narrow the scope of interpreting social cues emitted by robots. An alternate explanation may be that the stimuli themselves did not convey the affective states that were in the forced choice list, or any affective states at all.

The second hypothesis, that the morphology of the robot impacts the users' interpretation, was also found to be dismissible as no significant differences were found between the ratings for utterances when presented with the two different robot images. This is likely due to the lack of coherence between subjects and majority of chance level results also. In this light, H_2 is dependant upon H_1 - if subjects are not coherent in general, then it is unlikely to find significant differences between specific sub sets of data.

Addressing the significant results that were obtained, arguably the most important result from this experiment is the finding that subjects do have a sensitivity and preferences for the alignment between the morphology of a robot (which supports H_3 - that the morphology impacts the perception of the appropriate of an utterance), and the *type* of utterance that is made, a notion that falls somewhat in line with the basic premise of the Uncanny Valley whereby a mismatch between two or more facets of an agent results in an adverse response from people. While it is rather intuitive that the subjects showed preference for human type utterances made by the Nao humanoid and animal like utterance made by the Aibo dog robot, it is interesting to see that there is a weak, but statistically significant preference for the technological class of utterances made by the Nao robot. While it is difficult to isolate the exact cause of this result, this result does build confidence in the notion of using the Nao platform as the means through which

utterances in subsequent experiments will be embodied.

More broadly, these results also suggest that there are limits as to how NLUs may be used in robotic agents, and that some of these limits, at least, are manifested through the physical morphology of the embodiment. Thus, this highlights the need for designers of (social) robots to be aware of the impacts that their physical designs have upon how other aspects of the system, such as the use of NLUs, and visa-versa. Furthermore, it is clear that the designers should also be aware of the potential limitations the physical design of a robot has upon the other parts of the system with respect to the types of behaviours that are likely expected and deemed as acceptable from an end users perspective.

A prominent gender difference was also found for the appropriateness ratings for both the Animal and Technological utterance classes, and not the Human class. This may be in part due to the imbalance of male and female subjects, but also that 12 of the male subjects self reported to have frequent contact with real robots. This may be the source of significant bias as a result. In the case of the Technological utterances, the results show that the female subjects found the use of NLUs more acceptable than males, while the results from the male subjects were not above chance. An implication of this for the use of NLUs in robots is that, if the robot is both aware of the gender of a user (say through the use of a *User Model*) and able to select from a variety of means of acoustic expression, it may be beneficial to know that the use of NLUs may be more appropriate with female users than male users. However, due to the limitations of this study with respect to ecological validity, the weight behind such implications should be kept in consideration.

4.3.1 Methodological Remarks

There are some shortcomings with regard to the methodology within this experiment. Firstly there was a small number of total utterance stimuli used, with these being unevenly spread across the three classes of utterance. The lack of specification and variation in acoustic properties does not allow for insights to be gained

with respect to how certain aspects of an utterances impacted the interpretation. However, this was not the main focus of the study.

Due to the online nature of the experiment, there is the issue of ecological validity in that not only were images of robots used, there was also no situational context presented. It may be that the inclusion of situational context can help narrow down the potential scope of interpretations of a given utterance (this is addressed in chapter 8). The lack of situational context however was a deliberate choice. The rationale for this was to remove all possible sources of biasing that may impact the subject responses, as it is possible that factors such as situational context may colour, or possibly override completely, any information or content carried by the utterances alone. With respect to the use of images over videos or real robots, images also provide a neutral context due to their static nature. This too was done in an effort to avoid confounding any raw effects of the utterances. Though, this approach comes with the risk of reducing ecological validity. While some authors argue that online studies that employ videos do retain ecological validity (e.g. Lohse et al. (2008a); Woods et al. (2006a,b); Walters et al. (2011)), it is unlikely that this same argument extends to the use of static images.

4.4 Summary

This chapter has presented an experiment aimed at confirming that the Nao robotic platform is indeed an appropriate platform through which the research into Non-Linguistic Utterances may be embodied into a real world social agent. This issue stems from the Uncanny Valley hypothesis which has the basic premise that the physical morphology and behaviour of an agent has an impact upon the overall empathic response of a person to that agent. As such, this chapter has sought to confirm that people do not have an adverse response to the Nao robotic platform when it employs the NLU modality, in comparison to a similar type of robot, the Sony Aibo. This experiment also allowed testing of three hypotheses: that subjects are coherent in their interpretations of utterances, that the robots physical morphology has an impact upon this interpretation, and that the

morphology also impacts whether subjects deem the given morphology/utterance combination as appropriate.

The first two hypothesis were not supported as little coherence was found between subjects in their ratings, as well as no significant differences in the affective and intentional interpretations of utterances. The results of this experiment are somewhat in agreement with those of Komatsu and Yamada (2011) in that the subjects had a different appropriateness rating of an utterance when it was presented with a different robot, thus the morphology of an utterance indeed had an affect, which supports H_3 .

While the experiment in this chapter employed acoustic samples obtained from an internet based source, in the majority of cases, this is not an approach that is suitable or sustainable for the rest of this thesis. As such, the rest of this thesis uses the utterance generation algorithm detailed in chapter 3 to produce the NLU stimuli used in the experiments.

Finally, with respect to the main aim, the experiment has shown that though NLUs were not deemed as appropriate to the same level as the Human utterances, they were deemed as appropriate to a degree that is above chance. This finding provides the initial confidence in the notion that subjects in subsequent experiments will be accepting of the Nao robot making NLUs, and as such, the Nao has been used as the sole platform used for the social embodiment in all experiments presented in this thesis.

Chapter 5

Collecting Training Data for Machine Learning

Summary of the key points:

- An experiment designed to systematically explore the different parameters of the NLU generation algorithm and their impact upon affective interpretation is conducted with young children.
- Data collected is intended to be used as *training data* for Machine Learning later in this thesis to learn an affective mapping between the different NLU parameters and affective interpretation.
- Results show that different parameter values do indeed evoke different affective interpretations of NLUs, but that the relationships between parameters and affective meaning is complex, subtle and noisy.
- While overall coherence between subjects' in their affective interpretations is low, children do readily rate NLUs as having distinct affective meanings. Furthermore, there are indications that subjects' interpretations are subject to a perceptual magnet effect/categorical perception.

In chapter 3, a custom algorithm for generating NLUs was described, taking inspiration from research into the human voice and gibberish speech, with respect to both the structure of utterances and the parameters that are used to specify/charge an utterance. While there is an established literature surrounding the acoustic correlates of conveying affect through the human vocal system as well as music, it is not wise to assume that the findings and insights from this body of literature are directly applicable to the generation algorithm in order to affectively *charge* or *colour* utterances in the same manner (i.e. the same basic, acoustic features of an NLU may not evoke the same affective response as a human utterance). The rationale for this is two-fold. Firstly, with respect to the human voice, the utterances generated via the algorithm are not only very simplistic in comparison (i.e. utterances consist of simple single carrier wave modulations with clear constraints, while the human voice is composed of complex wave signals), the manner in which the generation parameters interact are only analogous to those of the human voice and are likely to not have enough overlap to facilitate a direct mapping. Secondly, with respect to the world of affective expression via music, in comparison, both NLUs and the human voice operate in a notably shorter time frame, and while there may be similarities in the underlying characteristics describing the acoustic signals (e.g. frequency ranges, the rhythm, the melody/pitch contours, etc.) that are used to encode an affective meaning (Weninger et al., 2013), how these characteristics are used functionally is different (Scherer, 1995), though high level comparisons may be made (see Juslin and Laukka (2003) for an extensive review of this).

As such, in order to be able to use the NLUs from the algorithm to convey the robot as having a specific and desired affective state, it is necessary to ascertain how the parameters of the generation algorithm translate to, and impact, different affective interpretations of the utterances that can be generated by the algorithm. In essence, a mapping between generation parameters and affective interpretation is required, making it possible to estimate the parameter configuration of an utterance in order to evoke a desired interpretation.

This chapter presents an experiment in which local school children ($n = 42$) were asked to affectively rate, using the AffectButton measuring tool, utterances generated by the NLU generation algorithm, with the aim of collecting data that could help reveal the relationships between the parameters of the generation algorithm and coordinates within an Pleasure, Arousal, Dominance *affect space*.

It is important to stress from the offset that the work presented in this chapter had one primary goal: to collect a broad variety of *training data* that was used in chapter 7 to train Artificial Neural Networks in order to learn the aforementioned mapping. Given that a large amount of data is collected, the opportunity to perform an analysis checking for any initial insights was also taken, however as explained below, the experimental arrangement designed to meeting the primary goal did hamper the raw analysis of the data collected.

5.1 Identifying an Affective Mapping

Searching for such a mapping requires that as many possible combinations of the different parameters be presented to, and rated by subjects. This is challenging task for two linked reasons. Firstly, there are a large number of parameters controlling the NLU generation algorithm, and secondly, the majority of parameters are continuous and thus discrete samples must be taken for each utterance, leading to an issue of *sampling resolution*. This latter issue holds relevance as it means that a practical trade off must be made between the number of utterances required to uniformly sample the high dimensional parameter space, the number of subjects required to rate these utterances, and the number of utterances each subject is able to rate as the subjects in this experiment were young school children with a limited attention span.

In total, the algorithm has 14 controllable parameters, which presents a considerably large, high dimensional space¹ in which to identify potential relationships between parameters and their affective interpretations. To reduce the size of the

¹This high dimensional space of the algorithm parameters will be referred to as the *parameter space*.

parameter space, it was decided that a number of parameters providing fine tuned control of utterances would be held constant. Namely, the Envelope Count, Volume Intensity, Node Count, Skew Ratio and Node Ratio were kept at constant values² as they introduce a degree of complexity that is unmanageable given the number of subjects available. While this reduced parameter space of nine dimensions (Wave Type, Base Frequency, Frequency Range, Pitch Contour, Sound Unit Count, Pause Ratio, Speech Rate, Rhythm and Tremolo) is still large, it does encompass the primary parameters that specify an utterance and are parameters that are found in the related literature also (allowing room for potential insights with respect to the relevance of the findings from that body of literature to NLUs). As such, it was deemed unwise to reduce the space any further as this would neglect the basic features of an NLU. It was decided that sparsely sampling the parameter space would be an acceptable trade off between sampling resolution and the number of utterances, provided that the sampling uniformly cover the full working range of each of nine the continuous parameters.

As it was not assumed that insights from the body of literature on affect in music and the human voice may be applied to NLUs, utterances that were generated for this study were not assumed to have any particular affective interpretation. Rather, as outlined above, the utterances were generated to provide a board, uniform sampling of the parameter space. The rationale for this is three-fold: firstly, no assumption is made regarding prior understanding of how generation parameters are related and interact with each other as well as counter part representations of affective ratings (i.e. the AffectButton affect space).

Secondly, the debate of which affective states should be represented by the stimulus is effectively side-stepped. Recent developments in emotion research have proposed a variety of theories of emotion (see chapter 3), generally with limited underlying agreement between theories. For example, a recent theory proposes families of emotions (Ekman, 1992) in which it is acknowledged that similar classes of emotion (e.g. “hot” and cold” anger) have notably different acoustic correlates

²These parameters have been kept constant for all other studies presented later in this body of work also.

in the human voice (Banse and Scherer, 1996; Juslin and Laukka, 2003; Scherer, 2003). As a result, it becomes important to ensure that the desired emotional state is indeed represented both in the encoded stimulus as well as in the measuring tool when subjects are decoding the stimulus. Failure to do so may lead to results that disagree with others within the literature (Scherer (1986); Banse and Scherer (1996); Scherer (2003)), making it difficult to compare results. Side stepping this issue with respect to the affective encoding of an NLU, and using an affective measuring tool that does not explicitly rely on distinctions such as “hot” and “cold” anger is able to avoid such issues.

Finally, no assumptions are made about the coverage of the affect space given the generated stimuli, allowing clear assessment of whether the stimuli provide full coverage of the affect space, and whether a uniform sampling of the parameter space equates to a uniform distribution of data points in the affect space.

Given the exploratory nature of this chapter, no hypotheses are presented or tested.

5.2 Experimental Setup

The study was organised in collaboration with a local primary school, where subjects were recruited through two school classes - year 2 (6-7 years old) and year 3 (7-8 years old). As well as recruiting subjects through the school, a school class room also hosted the study setting, to help promote the notion of a real world evaluation environment - an environment that was familiar to the subjects. The study was conducted with pairs/trios of children, with each group being pulled out of class time and each child taking turns to listen to and rate the stimuli. The experimental duration for each group was between 20 and 30 minutes, with each child taking an average of 10 minutes to complete the ratings. The overall duration of the experiment spanned over two school days, with one class being completed each day.

Subjects were seated in front of a laptop running the AffectButton measuring tool, with the Nao stood behind it facing them (Figure 5.1). Utterances were



Figure 5.1: Image of the experimental setup in the classroom with two children and the experimenter.

played (by tapping the Nao on the head) and the subjects asked to guess how the robot was feeling, inputting their response via the AffectButton. Subjects were presented with 5 practice utterances followed by 30 experimental utterances, with their affective interpretation being captured after each utterance by selecting a facial expression that matched their interpretation using the AffectButton. Repetition of the current utterance was permitted, while repetition of previous utterances was not.

The Nao was programmed to operate in an autonomous fashion such that the control of the experiment was achieved through physical interaction with the robot itself via the three touch sensors on the head. This presented an opportunity to invite the other children present to instruct the robot to play the next utterance, thus integrating them into the interaction scenario and minimising the amount of influence that they may have over the children rating the utterances. Also, in an effort to present the robot as an “alive”, reactive agent, rather than a static object, a neutral behaviour was implemented, where the Nao gazed around the room randomly, blinking by turning the eye LEDs off and on, shifting weight from leg to leg and moving its arms, wrists and fingers in controlled and subtle, yet random manner.

5.2.1 Utterance Specification

In total, 84 utterances were generated using the NLU algorithm, specifically addressing the variance of each of the nine generation parameters outlined. These 84

utterances, and the manipulation of the nine different parameters were organised into five *mini* experiments, each addressing the systematic manipulation of utterance parameters with intuitive relationships. For example, the speech rate and pause ratio both relate to temporal aspects of an utterances and thus were studied in one mini experiment, while the base frequency and frequency were studied together in another. A further 60 utterances were generated without the systematic variations to parameters (these were not subject to analysis in this chapter). The purpose of these 60 utterances was to increase the variety of different parameter configurations that were rated, helping provide a data set of affective ratings for utterances that sampled as much of the parameter space as possible.

The 5 mini experiments are described below:

Experiment #1: This experiment was designed to probe the role of the *Pitch Contour* and the *Tremolo* parameters with the aim of gaining insight as to what effects the first and last sound unit contour shape may have. Utterances consisted of 3 Sound Units only, with the first and last sound unit each having either a flat, rising or falling contour shape, while the middle sound unit had a fixed flat contour. All possible combinations of the first and last sound unit contours were generated (see table 5.1). These contour profiles were synthesised for utterances with Tremolo values of 0 *rad* and 0.34 *rad*, thus resulting in 18 utterances in total.

Experiment #2: Here the role of the *Rhythm* parameter was investigated. The Rhythm parameter could have the values of either 0, 0.5 or 1, where a value of 1 generates sound units that all have exactly the same duration, while a value of 0 generates sound units that have large differences in duration. The Sound Unit Count was again set to 3 and 5, and for each value five utterances were synthesised (resulting in 10 utterances in total). Rhythm values were set to either 0 (2 utterances), 0.5 (2 utterances) or 1 (1 utterance). The reason for having two utterances for both the 0 and 0.5 values is that sound unit duration is set pseudo-randomly and utterances with the same rhythm value (not set to 0) are likely to be different from each other.

Experiment #3: Here both the Base Frequency and Frequency Range were varied, each parameter having a value of 500 Hz, 1000 Hz, and 1500 Hz. All combinations of these parameters were used, and synthesised for both 3 and 5 word utterances, producing a total of 18 utterances.

Experiment #4: In a similar manner to Experiment #3, only the speech rate (2,4 or 6) and the pause ratio (0.25, 0.5 or 0.75) were varied, again with the Sound Unit Count being set to both 3 and 5. Thus, 18 utterances were produced.

Experiment #5: The final experiment probed the role of the contour profile and sought to clarify whether different parameter configurations, other than the *Pitch Contour* could be utilised to sway subject interpretation. This question stems from an ongoing debate regarding differing opinions over role of intonation and the contour profile: some authors argue that contour profile plays a vital role (Fernald, 1989), while others argue that the contour is likely to be influenced by language (Banziger and Scherer, 2005; Grandjean et al., 2006) placing less importance upon this feature. To address this, two different Parameter Configurations (PC1 and PC2) were specified (Table 5.4) and were used to generate utterances with 3 and 5 Sound Units. These Parameter Configurations had different Base Frequency, Frequency Range and Speech Rate values. Five different Pitch Contour specifications for both utterance lengths were also specified. This arrangement addresses two things: allowing comparison between the different *Pitch Contours* for a given *Sound Unit Count* with the same parameter configuration, and comparison be made for the same *Pitch Contours* across different parameter configurations.

Overall, all the utterance parameter configurations were selected in such a way as to maximise the sampling of the entire parameter space. Figure 5.2 shows a parallel plot of all the different combinations of parameters, with the exception of the Pitch Contour, to illustrate the sparse, but even sampling of the parameter space. The Pitch Contour is a more difficult parameter to manage as it not only

Table 5.1: Overview of the nine different Pitch Contour profiles for the utterances in experiment #1. Sound unit pitch contours are encoded as follows:
 F = flat contour, U = rising contour and D = falling contour.

Combination No	Contour Profile
1	F-F-F
2	F-F-U
3	F-F-D
4	U-F-F
5	D-F-F
6	U-F-U
7	U-F-D
8	D-F-U
9	D-F-D

Table 5.2: Specifications of the Utterance Parameter Configurations (PC) for the five mini experiments.

Parameter	Experiment #				
	1	2	3	4	5
Sound Unit Count	3	3/5*	3/5	3/5*	3/5
F0 Base (Hz)	1250	750/ 1000*	500/1000/1500	750/ 1000*	500/1500
F0 Range (Hz)	1250	1000/ 750*	500/1000/1500	1000/ 750*	500/1500
Tremolo (<i>rad</i>)	0/0.34	0	0.17	0.34	0
Speech Rate	4	3	5	2/4/6	2/6
Pause Ratio	0.25	0.75	0.5	0.25/0.5/0.75	0.75
Rhythm	1	0/0.5/1	1	1	1
Wave Type	Sine	Sine	Saw	Saw	Sine
#Utterances	18	10	18	18	20

* Values in bold text were used for 5 word utterances, while values in normal text were used for 3 word utterances.

has a large number of different combinations, but this number of combinations is dependant upon the Sound Unit Count. As such, the Sound Unit Count was limited to three of five sound units, depending on the mini experiment.

The utterances were randomly divided into five different utterance lists, with each list containing approximately 30 utterances. Subjects in a given pair/trio were each presented with a different utterance list, avoiding issues of subjects in the same group hearing the same stimuli. The presentation order of each utterance list was randomised to minimise ordering effects.

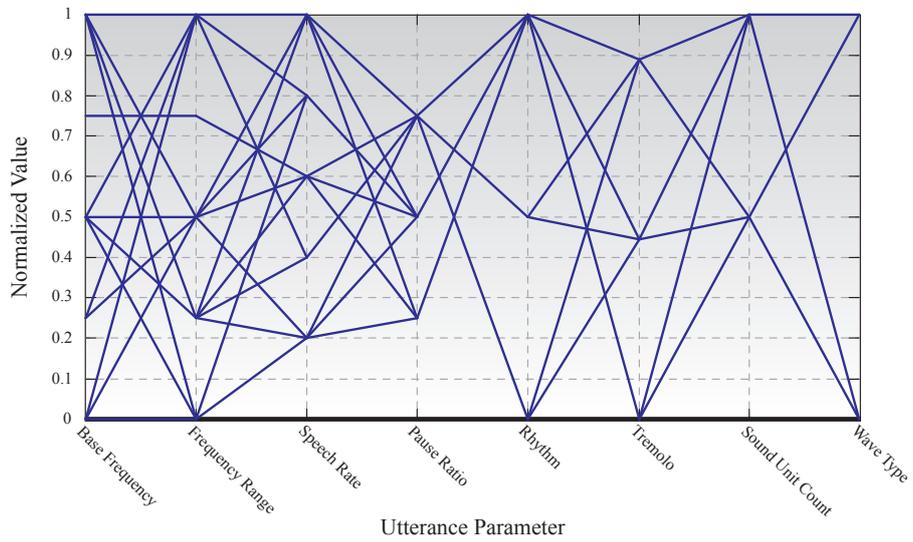
Table 5.2 provides an overview of the *parameter configurations* used in each of the different experimental arrangements while Table 5.3 outlines the Pitch Contour specifications, illustrating the range of different parameter configurations and pitch contours that were employed.

Table 5.3: Pitch Contour specifications for Experiment #2 through to Experiment #5, across the different sound unit counts. Contours are encoded as follows: F = Flat, U = Rising, D = Falling, Ud = Rising-Falling, Du = Falling-Rising.

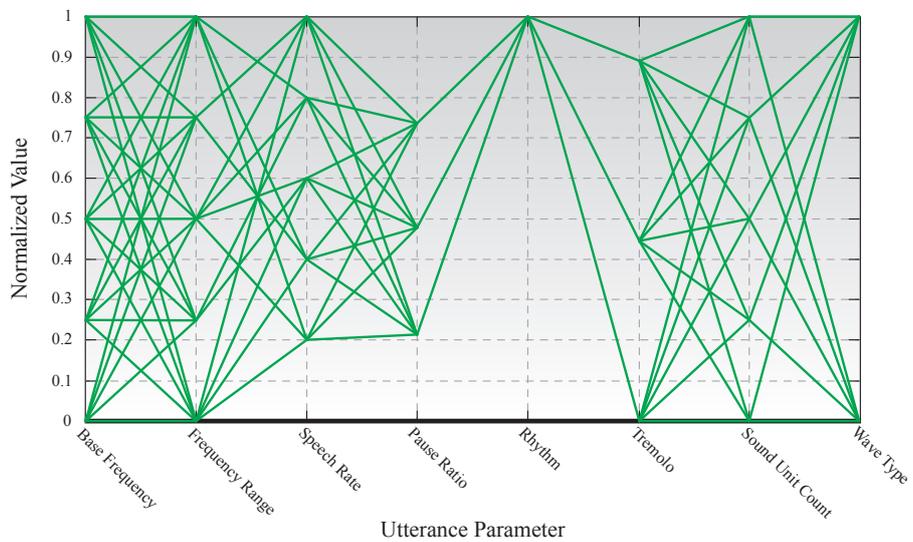
Experiment	Sound Unit Count	
	3	5
Exp 2	Ud-U-Du	U-F-Du-D-Ud
Exp 3	D-Ud-U	D-Du-U-F-Ud
Exp 4	Du-D-U	Ud-F-U-D-Du
Exp 5	F-Ud-U	F-D-U-Du-Ud
	U-F-Ud	U-Ud-D-F-Du
	D-U-F	D-Du-Ud-U-F
	Ud-Du-Du	Ud-F-Du-U-U
	Du-D-D	Du-U-F-D-D

Table 5.4: The two Utterance Parameter Configurations for mini experiment #5

Parameter	Parameter Config	
	PC1	PC2
Sound Unit Count	3/5	3/5
F0 Base	500	1500
F0 Range	500	1500
Tremolo	0	0
Speech Rate	2	6
Pause Ratio	0.75	0.75
Rhythm	1	1
Wave Type	Sine	Sine



(a) Parallel Plot of the 84 Experiment NLUs in the Parameter Space.



(b) Parallel Plot of the 60 remaining NLUs in the Parameter Space

Figure 5.2: Parallel Plots of the the NLU specifications in the Utterance Parameter space.

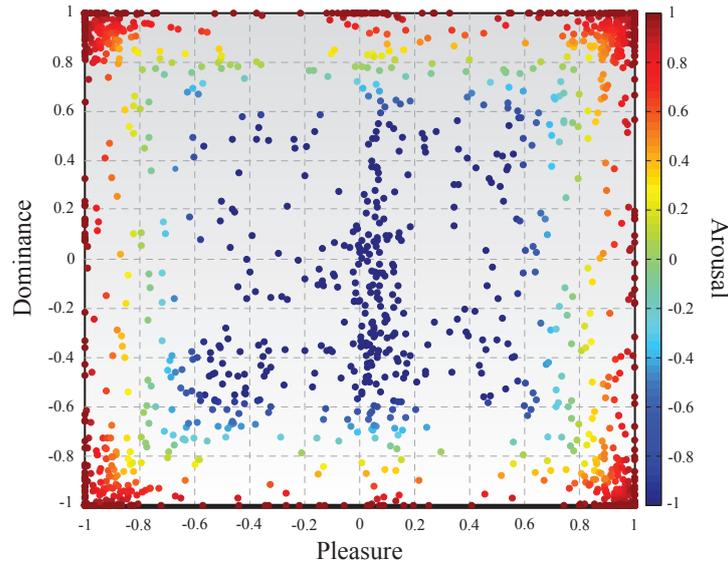


Figure 5.3: Plot of all the Affective Ratings for all the NLUs.

5.3 Results

In total, 54 subjects partook in the experiment, however not all of the data collected was used in the analysis. The data of six participants who were suspected of not engaging with the study and participant diagnosed with autistic spectrum disorder were omitted from the results analysis reported in this section. As such, the data collected from 48 children (26 girls and 22 boys) from the two classes (Y2 = 25 and Y3 = 23) was included in the results analysis.

Kruskal-Wallis (KW) tests were employed to perform one-way non-parametric ANOVAs, as the data did not fulfil the assumption of being normally distributed. These tests did not check for differences between the genders as the number of subjects in each group was not deemed to have a high enough frequency to hold any statistical validity and power.

The remainder of this section provides a brief statistical analysis of the five mini experiments.

5.3.1 Experiment #1

KW tests were performed to investigate the differences between the different Pitch Contours, along each affective dimension individually. This was done indepen-

dently for the utterances with the two different Tremolo values, as well as with all the utterances combined (regardless of the Tremolo values). Subject ratings were grouped in three different manners: by the unique Pitch Contour (9 groups), by the contour shape of only the first sound unit (3 groups), and by the contour shape of only the last sound unit (3 groups).

The tests found that there were no significant differences in the affective ratings due to any of the pitch contour groupings along the Pleasure, Arousal and Dominance dimensions, with only one exception. It was found that there was a significant difference in the Dominance ratings in the utterances with a tremolo value of 0, and grouped by the first contour shape ($\chi^2(2) = 7.57, p = 0.023$). Mann-Whitney's U post-hoc tests revealed that utterances beginning with a flat contour were found to be significantly lower ($U = 706.5, z = 2.637, p = 0.004$) than the utterances beginning with a rising contour. The same was also true for utterances beginning with a flat and falling contours ($U = 460, z = 1.780, p = 0.037$), while no significant difference ($U = 627, z = 1.114, p = 0.133$) was found between utterances beginning with raising and falling contours. These results are shown graphically in figure 5.4. The results of all these KW tests are shown in table 5.5.

To test the effect of the Tremolo values, KW tests were again performed, with the ratings being grouped into the tremolo values (0 *rad* or 0.34 *rad*), with all the pitch contours collapsed together. No significant effects were found due to the different tremolo values along the Pleasure ($\chi^2(1) = 0.7, p = 0.405$), Arousal ($\chi^2(1) = 0.67, p = 0.412$) or Dominance ($\chi^2(1) = 0.5, p = 0.477$) dimensions. These results are summarised in table C.1.

5.3.2 Experiment #2

With regard to the Rhythm parameter, the KW tests found that there were no significant main effects due to the rhythm parameter value, across any of the affective dimensions, for either the utterances consisting either of 3 or 5 sound units. However, when all the utterances (regardless of the sound unit count)

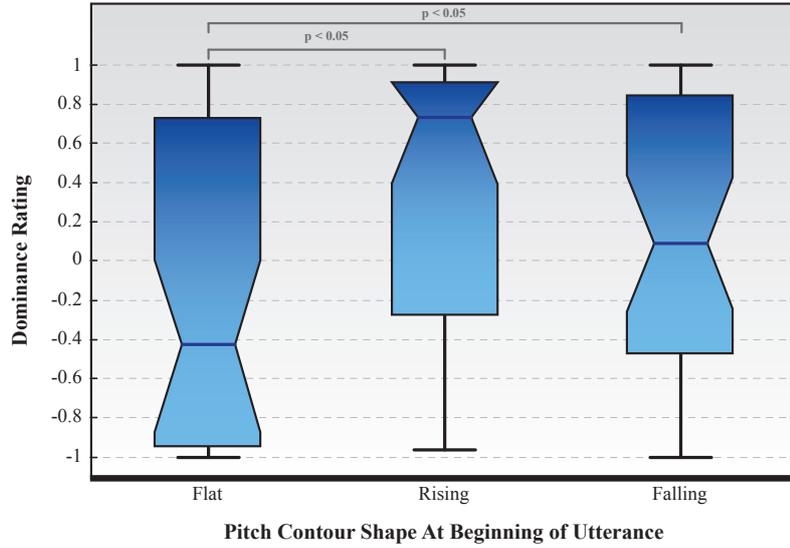


Figure 5.4: Box plot of the Dominance ratings for the utterances in Experiment #1 with a tremolo value of 0, grouped by the the first contour shape in the utterance (see table C.2 for a summary of the descriptive statistics).

Table 5.5: Results of Kruskal-Wallis tests for Experiment #1, comparing the Pitch Contour and Tremolo parameter specifications against the affective ratings. Pitch Contour has three possible groupings: by the whole pitch contour combination, by the pitch contour of only the first sound unit, or by the contour of only the last sound unit.

Contour Grouping	Tremolo	<i>d.f.</i>	Affect Dimension					
			Pleasure		Arousal		Dominance	
			χ^2	<i>p</i> value	χ^2	<i>p</i> value	χ^2	<i>p</i> value
All	Any	8	6.12	0.634	5.86	0.663	12.5	0.130
First		2	0.17	0.918	4.36	0.113	5.28	0.071
Last		2	2.19	0.335	0.58	0.748	0.79	0.674
All	0	8	5.18	0.738	8.02	0.431	11.88	0.157
First		2	1.4	0.496	2.3	0.317	7.57	0.023
Last		2	1.14	0.564	1.22	0.543	0.25	0.879
All	18	8	5.87	0.661	3.36	0.910	10.8	0.213
First		2	2.92	0.233	2.19	0.335	3.4	0.183
Last		2	1.38	0.50	0.01	0.997	0.88	0.643

Table 5.6: Results of the Kruskal-Wallis tests in Experiment #2, testing the influence of the Rhythm parameter, with the data grouped by the sound unit count.

Affect Dimension	Unit Count	<i>d.f.</i>	χ^2	<i>p</i> value
Pleasure	3	2	0.66	0.718
Arousal		2	1.54	0.463
Dominance		2	4.93	0.085
Pleasure	5	2	0.86	0.651
Arousal		2	5.16	0.076
Dominance		2	0.15	0.927
Pleasure	Both	2	0.24	0.887
Arousal		2	5.93	0.052
Dominance		2	1.86	0.395

collapsed together, the KW gets did identify a significant effect along the Arousal dimension ($\chi^2(2) = 5.93, p = 0.052$). The test results are summarised in table 5.6. Mann-Whitney U post-hoc tests found the utterances with rhythm value of 0 and 1 received significantly different ratings ($U = 742, z = 2.345, p = 0.009$), with the rhythm value of 1 receiving a higher rating with a notably smaller range of ratings than the utterances with a Rhythm value of 0. No significant differences were found between the Rhythm values of 0 and 0.5 ($U = 1227, z = 1.52, p = 0.064$), and 0.5 and 1 ($U = 618.5, z = 0.983, p = 0.163$). These results are shown graphically in figure 5.5.

Tests were also performed to check for differences in the ratings due to the two different sound unit counts (3 and 5), and found that there were no significant differences, across any of the affective dimensions. The results of these tests are summarised in table C.3.

5.3.3 Experiment #3

When collapsing the Frequency Range values and comparing only the ratings grouped by the Base Frequency values, no significant differences were found between the three conditions along any of the affective dimensions, with this being true for the utterances with 3 sound units, 5 sound units, and when all the utterances collected together (see table C.5 for these results).

Collapsing all the Base Frequency values and comparing the ratings grouped

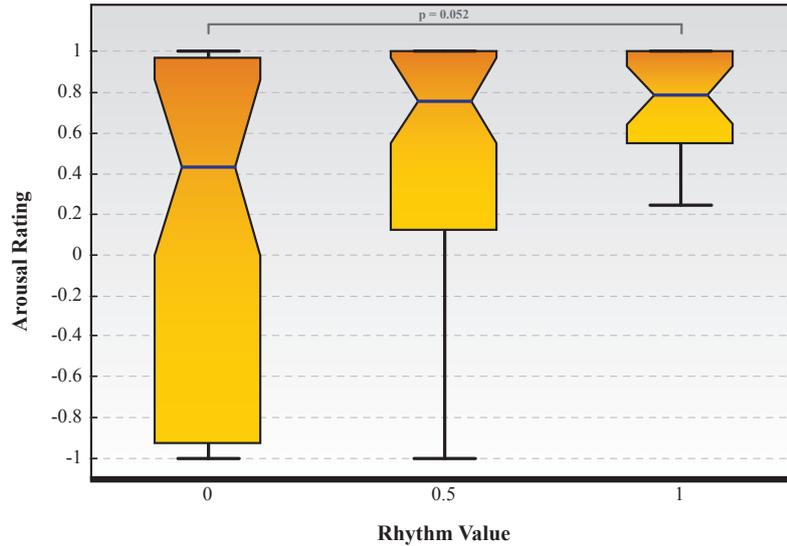


Figure 5.5: Box plot of the Arousal ratings for the all the utterances in Experiment #2 (regardless of the sound unit count), across the three different Rhythm values (see table C.4 for a summary of the descriptive statistics of the box plot).

by the three Frequency Range of the NLUs, KW tests found that for 3 sound units, there was a significant difference in the ratings along the Dominance dimension ($\chi^2(2) = 6.805$, $p < 0.05$), while for the 5 sound unit utterances there was a significant difference along the Pleasure dimension ($\chi^2(2) = 7.111$, $p < 0.05$). The results of the KW test for each dimension, across the two sound unit counts and these collapsed are summarised in table 5.7.

The Mann-Whitney U post-hoc tests found that for the NLUs of 3 sound hits in length, the utterances with a Frequency Range of 500 Hz received significantly lower Dominance ratings than both the utterances with a frequency range of 1000 Hz ($U = 421.5$, $z = 2.363$, $p = 0.009$), and 1500 Hz ($U = 607.5$, $z = 2.099$, $p = 0.018$), and that there was no difference in the ratings between the 1000 Hz and 1500 Hz conditions ($U = 522.5$, 0.117 , $p = 0.454$). This is shown in figure 5.6.

For the NLUs with 5 sound units, the post-hoc tests revealed that the utterances with a Frequency Range of 1500 Hz were rated significantly higher along the Pleasure dimension than both the 500 Hz ($U = 629.5$, $z = 2.123$, $p = 0.017$) and 1000 Hz conditions ($U = 788$, $z = 2.523$, $p = 0.006$), and that there was no difference between the 500 Hz and 1000 Hz conditions ($U = 661.5$, $z = 0.011$,

Table 5.7: Results of the Kruskal-Wallis tests in Experiment #3, collapsing the Base Frequency and testing only for the influence of the Frequency Range across the different sound unit counts and affective dimensions.

Affect Dimension	Unit Count	d.f.	χ^2	<i>p</i> value
Pleasure	3	2	0.975	0.614
Arousal		2	0.094	0.954
Dominance		2	6.805	0.033
Pleasure	5	2	7.111	0.029
Arousal		2	1.782	0.410
Dominance		2	1.479	0.477
Pleasure	Both	2	4.996	0.082
Arousal		2	0.768	0.681
Dominance		2	0.836	0.659

Table 5.8: Results of the Kruskal-Wallis tests in Experiment #3, comparing the difference in ratings across the two difference sound unit counts.

Affect Dimension	d.f.	χ^2	<i>p</i> value
Pleasure	1	0.34	0.558
Arousal	1	0.63	0.428
Dominance	1	3.61	0.057

$p = 0.496$). These results are shown in figure 5.7.

When both the Base Frequency and Frequency Range values were interleaved, providing 9 grouping variables for the ratings, no significant differences were found, along any of the affective dimensions, between the grouping variables, with this being true for the 3 Sound Unit NLUs, 5 Sound Unit NLUs, and when all NLUs were collapsed together. The results of these KW tests are shown in table C.6.

Comparing the ratings across the sound unit counts, the KW tests found that there was a *near* significant difference in the results due to the different sound unit count values in the NLUs, along the Dominance dimension ($\chi^2(1) = 3.61$, $p = 0.057$), while this was not the case along the Pleasure ($\chi^2(1) = 0.63$, $p = 0.428$) and Arousal ($\chi^2(1) = 0.3$, $p = 0.558$) dimensions (see table 5.8). These results are summarised in figure 5.8.

5.3.4 Experiment #4

When collapsing the Speech Rate parameter and testing only the three Pause Ratio values, the KW tests found that there were no significant differences in the

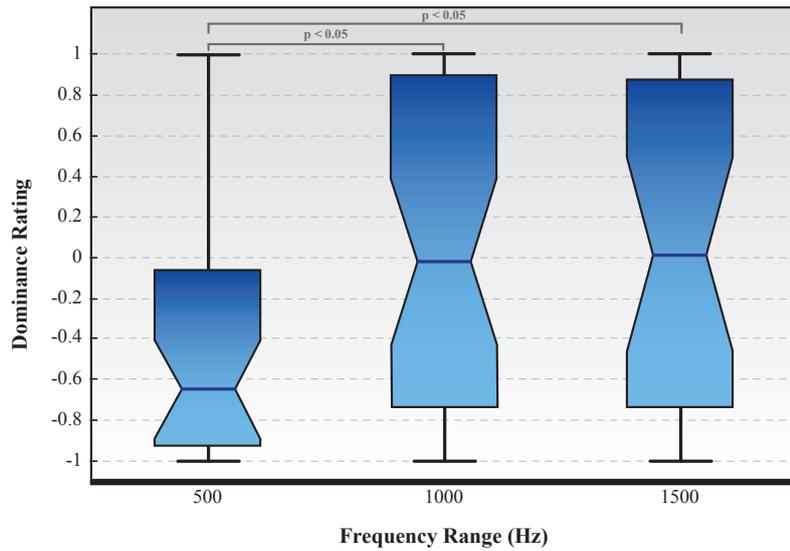


Figure 5.6: Box Plots showing the difference in ratings grouped by Frequency Range for NLUs with 3 Sound Units, along the Dominance dimension in Experiment #3 (see table C.7 for descriptive statistics for this figure).

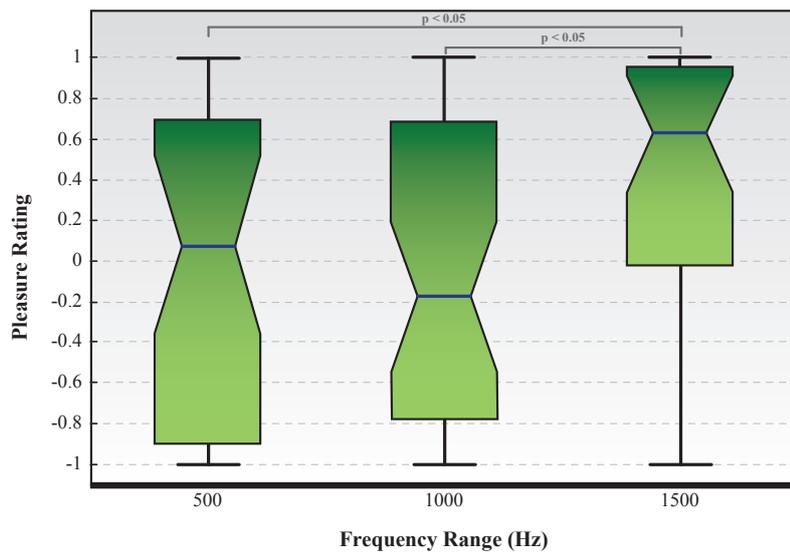


Figure 5.7: Box Plots showing the difference in ratings grouped by Frequency Range for NLUs with 5 Sound Units, along the Pleasure dimension in Experiment #3 (see table C.8 for descriptive statistics for this figure).

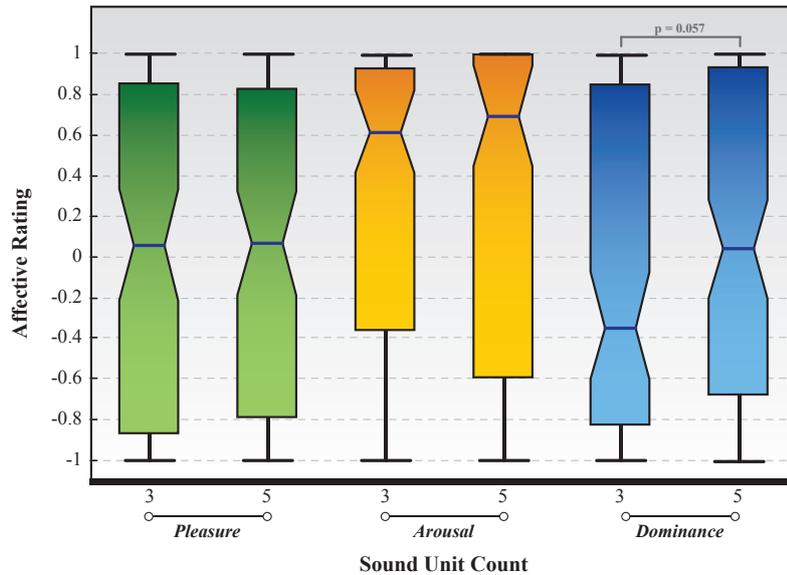


Figure 5.8: Box Plots showing the difference in ratings between the Sound Unit Count values, for each of the Affect Space dimensions in Experiment #3 (see table C.9 for descriptive statistics for this figure).

affective ratings, with this being true for the 3 sound unit NLUs, 5 sound unit NLUs, as well as when all the NLUs were tested together (see table C.10 for the KW test results).

Similarly, no significant effects were found along any of the affective dimensions when the Pause Ratio parameter was collapsed and the Speech Rate parameter tested, with this also being true for all the sound unit conditions (see table C.11).

When interleaving the Pause Ratio and Speech Rate parameters (resulting in 9 grouping variables), again no significant effects were found along any affective dimension. This was true for each of the sound unit conditions (3, 5 and both). These results are shown in table C.12.

A significant difference was found along the dominance dimension when checking for differences due solely to the two different sound unit count values ($\chi^2(1) = 8.13$, $p = 0.004$), see table 5.9. Figure 5.9 shows box plots of the ratings for the NLUs with 3 sound units and 5 sound units, for each of the affect dimensions. This shows that the utterances with 5 sounds units were rated as significantly higher than the utterances with 3 sound units along the dominance dimension.

Table 5.9: Results of the Kruskal-Wallis tests, checking for the differences in ratings due to the different Sound Unit Counts (3 and 5) of the utterances, along each affective dimension in Experiment #4.

Affect Dimension	<i>d.f.</i>	χ^2	<i>p</i> value
Pleasure	1	1.6	0.206
Arousal	1	0.01	0.904
Dominance	1	8.13	0.004

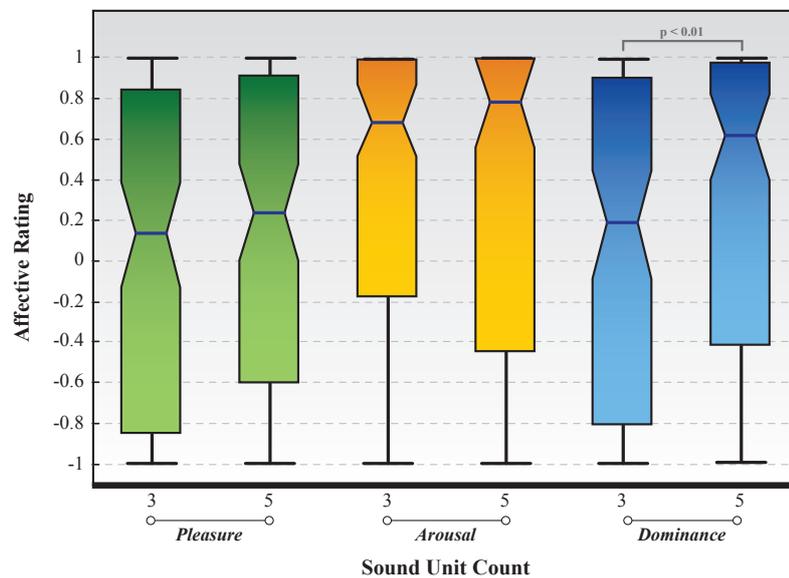


Figure 5.9: Box Plots showing the difference in ratings between the Sound Unit Count values, for each of the Affect Space dimensions in Experiment #4 (see table C.13 for descriptive statistics for this figure).

5.3.5 Experiment #5

Comparison between the ratings between utterances with PC1 and PC2 using the KW tests, for utterances with 3 sound units and 5 sound units independently, found that there were no significant differences in the ratings along any of the three affect dimensions. However, when all the sound unit counts were collapsed together, the KW tests found that there was a significant difference along the Dominance dimension ($\chi^2(1) = 4.52, p = 0.034$). The results of these tests are summarised in table 5.10, while figure 5.10 shows a box plot of the ratings for each Parameter Configuration and shows that the dominance ratings for utterances with PC1 are significantly lower than those for PC2.

As there were five utterances for each Parameter Configuration (PC1 and PC2), each with a different Pitch Contour specification, KW tests were performed to check for differences in the ratings (along each of the three affective dimensions) due to the five different Pitch Contour specifications. This was done for the five utterances with a Sound Unit Count of 3 and had been generated using PC1, and then also for the utterances which had been generated using PC2. The same was then done for the utterances with a Sound Unit Count of 5. Thus in total, 12 KW tests were performed. It was found through these tests that there were no significant differences in the ratings across the five different Pitch Contour specifications, across either Sound Unit Counts or Parameter Configurations. This shows that, contrary to the results of experiment #1, that the Pitch Contour specification did not have a significant effect upon how subjects rated the utterances. These results are summarised in table 5.11.

Finally, when grouping the data by sound unit count alone, the KW found that there were no significant differences along either the Pleasure ($\chi^2(1) = 1.82, p = 0.1778$), Arousal ($\chi^2(1) = 2.8, p = 0.094$) or Dominance ($\chi^2(1) = 0.7, p = 0.402$) dimensions, indicating that the different sound unit counts appeared to have no significant effect on the ratings that subjects provided.

Table 5.10: Results of the Kruskal-Wallis tests checking for differences in ratings due to the two different Utterance Parameter configurations (PC1 and PC2) in Experiment #5, with data grouped by the utterance Sound Unit Count (3, 5 and both). The table shows the degrees of freedom, χ^2 values and p values of the tests that were performed for each affective dimension individually.

Affect Dimension	Unit Count	<i>d.f.</i>	χ^2	<i>p</i> value
Pleasure	3	1	0.19	0.667
Arousal		1	3.31	0.069
Dominance		1	2.07	0.145
Pleasure	5	1	1.47	0.225
Arousal		1	0.13	0.724
Dominance		1	2.27	0.132
Pleasure	Both	1	0.2	0.652
Arousal		1	1.17	0.279
Dominance		1	4.52	0.034

Table 5.11: Results of the Kruskal-Wallis tests checking for significant differences between the five different pitch contour specifications in Experiment #5, with data grouped by Parameter Configuration (PC1 and PC2) and the Sound Unit Count (3 and 5) of an utterance. Tests show the degrees of freedom, χ^2 values and p values of the tests that were performed for each affective dimension individually.

Unit Count	Param Config	Affect Dim	<i>d.f.</i>	χ^2	<i>p</i> value
3	PC1	Pleasure	4	0.76	0.944
		Arousal	4	5.86	0.209
		Dominance	4	4.25	0.374
	PC2	Pleasure	4	1.47	0.833
		Arousal	4	3.01	0.557
		Dominance	4	3.67	0.452
5	PC1	Pleasure	4	4.21	0.378
		Arousal	4	0.39	0.984
		Dominance	4	6.88	0.142
	PC2	Pleasure	4	1.87	0.759
		Arousal	4	5.35	0.253
		Dominance	4	1.9	0.755

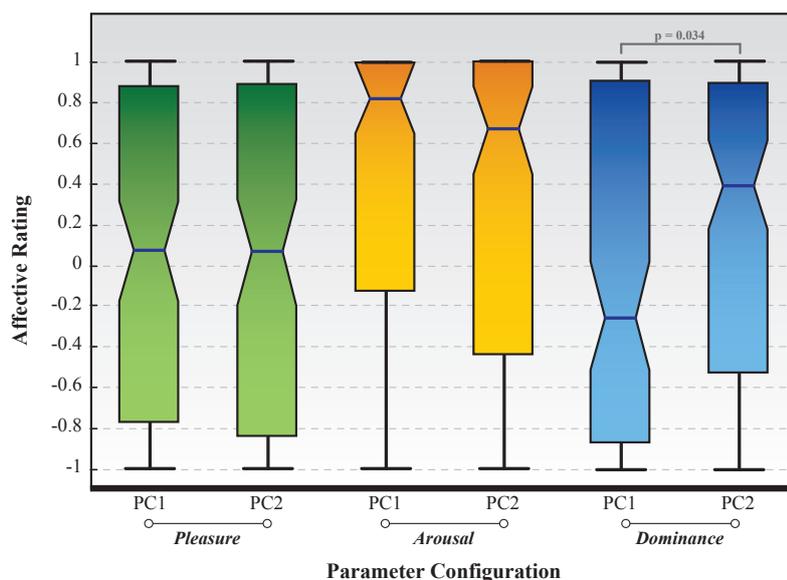


Figure 5.10: Box Plots showing the difference in ratings between Parameter Configurations 1 and 2 in Experiment #5, regardless of the Sound Unit Count, across each of the three affective dimensions (see table C.14 for the descriptive statistics for this figure).

5.3.6 Summary of Results.

Given the rather dense population of results that have been presented. This section provides an overview summary of the results that have been obtained via the analyses.

Experiment #1 probed the role of the Pitch Contour parameter across utterances consisting of three sound units. Nine different pitch contour combinations were tested, with utterances beginning with either a flat, rising or falling pitch contour, as well as ending with either a flat, rising or falling pitch contour. The middle sound unit had a fixed, flat, pitch contour shape. This was repeated for utterances with a tremolo value of 0 rad , and a value of 0.34 rad . It was found that utterances that began with a flat contour shape received a lower rating along the Dominance dimension than utterances that began with either a rising or falling contour shape, while no differences were found along the Pleasure or Arousal dimensions. No significant differences were found between the ratings due to the different tremolo values.

Experiment #2 tested the role of the Rhythm parameter, presenting utterances with either a rhythm value of 0, 0.5 or 1. This was done for utterances

consisting of both three and five sound units. While no significant differences were found due to the Rhythm values when the sound unit counts were isolated (3 or 5), a significant difference in the ratings along the Arousal dimension was identified when the sound unit count values were collapsed together, where utterances with a Rhythm value of 1 were found to have a significantly higher rating than the utterances with a Rhythm value of 0, and utterances with a Rhythm value of 0.5 were not found to have ratings that were significantly different from any of the other utterances.

Experiment #3 probed the influence and interaction between the Base Frequency and Frequency Range parameters, doing so for utterances with both 3 and 5 sound unit utterances. It was found that utterances consisting of 3 sound units, utterances with a Frequency Range of 500 Hz were rated significantly lower along the Dominance dimension than the other utterances. For the utterances consisting of 5 sound units it was found that utterances with a Frequency Range of 1500 Hz were rated significantly higher along the Pleasure dimension than the other utterances.

Experiment #4 investigated the relationship between the Pause Ratio and Speech Rate parameters, again for utterances consisting of 3 and 5 sound units. When considering only the Pause Ratio, no significant differences were found in the ratings along any of the affective dimension, regardless of the sound unit count. Similarly, no significant differences were found for the different Speech Rate parameters, again regardless of the sound unit count. However, a significant difference was found along the Dominance dimension when comparing the ratings across the two different sound unit counts (3 or 5). Specifically, it was found that utterances consisting of 5 sound units received a higher Dominance rating than utterances consisting of three sound units.

Experiment # 5 presented subjects with five utterances consisting of 3 and 5 sound units, repeated twice for two different groups of utterances with different utterance Parameter Configurations (PC1 and PC2). Also, utterances within each utterance parameter group had different Pitch Contour specifications. This

experiment sought to investigate whether two notably different utterance parameter configurations would evoke significantly different ratings from subjects, and whether utterances within the same group with respect to their parameter configurations but with different pitch contour specifications would have different ratings. The results show that the utterances with different Parameter Configurations did evoke ratings that were significantly different along the Dominance dimension, but not along the Pleasure or Arousal dimensions, regardless of the different Sound Unit Counts. With respect to the different Pitch Contours of the utterances, no significant differences were found in the ratings along any affective dimension. This result is contrary to that of experiment #1 which did find that there was a difference in how utterances were rated based upon the pitch contour of the first sound unit within an utterance.

5.4 Discussion

Overall, the results obtained in this chapter have only provided rather limited insights into how the nine different Utterance Parameters operate and interact, and how this may impact a child's affective interpretation of an NLU. Furthermore, some of these insights have been contradictory, particularly in the case of the Pitch Contour of an utterance. In experiment #1 it was found that utterances whose first sound unit had a *flat* pitch contour shape, received a lower affective rating along the Dominance dimension than when the first sound unit had either a *rising* or *falling* Pitch Contour shape. However, in experiment #5, it was found that there was no significant differences in the ratings for utterances that had different Pitch Contour shapes but the same Parameter Configuration.

These contradictory findings make it difficult to assess the general nature and influence that the Pitch Contour parameter has, and highlights the potential complexity that one can expect to face in attempting to investigate (through experimental design) and understand this feature of an utterance. This is further complicated given that as the number of sound units in an utterance increases, the complexity of search space grows exponentially (i.e. the more sound units in an

utterance, the more possible combinations of sound unit pitch contours). This is also further evidenced in the fact that across the five different mini experiments, experiments 3 and 4 found that there were (near) significant differences in the affective ratings due to the different sound unit count.

Setting the contradictory results aside, some findings have been informative such as those for the Frequency Range and Rhythm parameters. With respect to the Frequency Range it was found that utterances with a low value (500 Hz) were rated as significantly lower than utterances with a high Frequency Range (1500 Hz). This is an encouraging result as it coarsely follows the characteristics in the human voice that have been reported in the psychological literature: vocalisations with a high frequency range are commonly associated with an affective state that is high in Pleasure and Dominance (see Juslin and Laukka (2003); Scherer (2003)). However, no relationship was uncovered linking the frequency range to the base frequency (or fundamental frequency, in the human voice) in the results, while psychological literature do report this (Fernald, 1989; Banse and Scherer, 1996; Scherer, 1986, 2003). What is interesting about this result is that these differences occurred over different affective dimensions, depending upon the utterance sound unit count.

With respect to the Rhythm parameter, which controls the relative overall length of a given sound unit to all the other sound units in an utterance, the results have shown that utterances with a Rhythm value of 1 (i.e. all sound units in an utterance had the same duration) received a significantly higher rating along the Arousal dimension than utterances that had a Rhythm value of 0 (i.e. all the sound units in an utterance had a different duration). In essence this result shows that as the *beat* of an utterance increased, subjects' ratings were more toward the extremities of the AffectButton rather than the center. Overall, this is a result that is difficult to interpret and compare with the findings in both psychology and musicology as this particular parameter is not well defined in these two respective fields (Juslin and Laukka, 2003) and as such, the analogy within the utterance

generation algorithm is also limited³.

The results of experiment #5, aside from suggesting that the pitch contour has not influential role, have shown an important and fundamental characteristic, in that utterances with notably different parameter configurations do evoke significantly different affective interpretations, where slower, low pitch utterances were rated as having less Dominance than utterance that were fast and high pitched. This is a useful finding as it demonstrates that large differences in the utterance parameter confirmations do indeed evoke significantly different responses from subjects. Furthermore given that the two extreme parameter configurations lay at extreme ends of the parameter space, it shows that the working range of the parameters are indeed large enough to portray difference affective states, however this has only been found to occur over one affective dimension thus far.

While the findings of the results analyses have a limited overall scope with respect to the specific utterance parameters that been manipulated, an overview of all the data does show an interesting trend. From the plots in figures 5.2 and 5.3, it can be seen that a relatively even sampling of the parameter space has not resulted in an even distribution of affective ratings in the AffectButton affect space. Rather, it can be seen that data points appear to follow a non-normal distribution and are clustered in particular regions of the affect space. These regions also tend to coincide with the prototype regions of the AffectButton. It appears that the ratings are somewhat *binary* and have been pulled toward particular prototypical facial expressions. This shows that subjects seem to exhibit a form of categorical perception when interpreting the utterances, in that there is little subtlety. The robot was - in their eyes - either happy, sad, angry, scared, surprised or neutral, and seldomly interpreted the utterances in a more subtle manner. This is an interesting observation that is shown to have some far reaching implications with respect to the use of NLU's during social HRI (see chapters 6, 8 and 10).

³This lack of analogy with the fields of psychology and musicology is also why through the majority of this body of work the Rhythm parameter is held constant at a value of one in utterances generated using the Utterance Generation Algorithm.

5.4.1 Methodological Remarks

The general lack of clear and firm findings in the results of the analyses may well be attributed to a number of methodological shortcomings in the experimental arrangement. Firstly, it must be noted that the goal of attempting to provide as broad and even a sampling of the Utterance Parameter space in the interest of collecting training data has adversely affected the power of the statistical tests that have been performed. This is primarily due to the low number of subjects that rated each of the utterances in each of the mini experiments (ranging between 9 and 15 children for any given utterance).

Furthermore, given that the utterances for each experiment were randomly split into three groups of utterances, the experiential design became a mix between an independent samples and a repeated measures paradigm, as not all the utterances in each experiment were rated by all the same children. This has resulted in a child hearing only a small subset of the utterances in each of the experiments, but not hearing all the utterances related to each mini experiment. As such, this is a flaw in the experimental design that has made a statistical analysis difficult, as has the finding that the data did not follow a normal distribution (thus violating basic assumptions of an ANOVA based approach to the results analyses). This is why non-parametric tests have been employed. The Kruskal-Wallis tests have been employed (as opposed to the Friedman tests) as these perform analogously to an independent samples one-way ANOVA, and it was deemed that the experimental setup resembled more an independent samples arrangement than a repeated measures arrangement.

Finally, comment must be made regarding the manner in which the mini experiments went about testing the influence of the individual utterance parameters. On one hand it was a correct methodology to manipulate only the parameters that were being investigated in each mini experiment. However, it is the values of the other parameters that may be subject to criticism. The other values of the parameters were set with the overall goal of providing as broad a range of the overall parameter space as possible. Thus, each of the mini experiments was

performed in a different general *region* of the utterance parameter space. Herein may lie a problem in that the influence that the particular utterance parameter being examined (e.g. the speech rate) has upon the interpretation of the whole utterance may be dependant upon the general region of the parameter space within which it is being tested - i.e. if the value of the parameters that have been held constant were different, one might find that the parameter being examined may have a different influence. The underlying point to be made is that examination of a given parameter may have occurred in a region of the parameter space in which the parameter itself could have had little overall influence.

5.5 Summary

This chapter has described an experiment designed to provide initial insights regarding the how the Utterance Generation method described in chapter 3 operates, and how nine of the parameters (Base Frequency, Frequency Range, Speech Rate, Pause Ratio, Rhythm, Sound Unit Count, Wave Type, Tremolo and Pitch Contour) may interact and impact how child subjects affectively interpret utterances generated using this method. Subjects were young children from a local primary school and were asked to listen to the NLUs and use the AffectButton tool to assign a facial expression to the utterance, indicating how they thought that the robot felt when it made a given utterance.

It is stressed that the over arching goal of the experiments presented was to collect affective ratings of a large variety of utterances that, overall, provided a sparse, but uniform sampling of as many of the possible combinations of utterance parameters as possible. The purpose of this was that the data collected in this experiment has formed the large majority of the training data used to train Artificial Neural Networks to learn a mapping between the AffectButton affect space and the utterance parameter space, such that utterances may be coloured in such as manner as to evoke a particular desired affective interpretation.

While the main goal of this chapter has been the collection of training data, a brief analysis of the data is also presented, though methodological drawbacks are

acknowledged. The results show that two different utterance parameter configurations, at two opposite ends of the parameter space are indeed rated as having a significantly different affective interpretation. It was found that utterances that began with a flat Pitch Contour were rated as having less Dominance than utterances that began with either a rising or falling pitch contour. However, other results contradicted this, indicating that utterances with notable different pitch contours but the same parameter configurations did not receive significantly different affective ratings. It was also shown that as the Frequency Range of a Pitch Contour increases, this resulted in a significant increase in ratings along both the Pleasure and Dominance dimensions. Similarly, it was found that as the Rhythm parameter increased, subjects tended to provide more extreme ratings, by selecting facial expressions that were closer to the edge of the Affect-Button and thus having a higher overall Arousal rating.

Finally, it was observed that while the utterances provided a uniform sampling of the utterance parameter space, the children provided ratings that are suggestive of categorical perception in that their ratings were not subtle, but rather tended toward prototypical facial expressions of happy, sad, scared, angry, surprised and neutral. This suggests that while utterances may have subtle, uniform differences in their acoustic characteristics, these differences do not result in subtle differences in affective interpretations, but instead lead to significantly different affective interpretations that appear to follow the basic emotion categories. This particular insight has important implications regarding the use of NLUs during social HRI and thus is addressed in a more direct manner in the next chapter.

Chapter 6

Categorical Perception of NLUs

Summary of the key points:

- Adopting the experimental approach that has been matured in psychology, an experiment to assess whether peoples' affective interpretations of NLUs are subject to Categorical Perception is conducted with both adults and children.
- Using two different continuums of NLUs with equal, linear, physical differences between stimuli, subjects were asked to rate utterances as having either a *similar* or *different* affective meaning (discrimination task), and to assign a facial expression to each NLU using the AffectButton in order to identify the conveyed affective meaning (identification task).
- Results show that adults do indeed exhibit signs of Categorical Perception when affectively rating NLU, while children do not.
- The main implication for NLUs is that one does not need to invest a great deal of effort in creating NLUs that convey subtly different affective states as these efforts are likely lost due to subjects' perception of NLUs being drawn to well established affective prototypes.

The results from chapter 5 have suggested that NLUs were interpreted in a Categorical manner. To probe this further, in a more direct manner, this chapter presents the results of two experiments (one with adults, and the other with school children) aimed at uncovering whether the affective user interpretations of NLUs are categorical and to what degree this may be the case. This has important implications on how NLUs may be best used by a robotic agent, and also giving insights as to how humans are likely to perceive and in turn affectively interpret utterances. For example, if subjects do interpret utterances categorically, then it suggests that efforts directed at producing utterances to convey subtly different affective states may not equate to subjects interpreting utterances as having subtly different affective charges or meaning. Thus, efforts directed at evoking subtly different affective interpretations of NLUs may well hold little practical value. Furthermore, if utterances are found to be interpreted categorically, when presented in a context-free manner, it may well be the case that when they are used in a scenario that does have a clear contextual setting (for example, a chess game (Castellano et al., 2013; Leite et al., 2013b)) and within a multi-modal interaction, effects of categorical perception of the robot may be more vivid. Chapter 8 provides valuable insights with regard to this issue.

6.1 Categorical Perception

Categorical Perception (CP) (Liberman et al., 1957) is the phenomenon where sensory stimulation is sorted into discrete categories within the brain. In essence, where one may have a stimulus continuum with equal, linear physical differences (for example, the *hue* of a colour), people exhibit discretisation of the continuum where a stimulus in a given region of the continuum gains membership to a given discrete *category* (e.g “red”, “blue”, etc). The hallmark of CP is subjects showing greater sensitivity to a physical change that occurs over a perceptual *boundary* than when the same physical change occurs within a perceptual *region* (Harnad, 1987; Laukka, 2005). As such, stimuli that lay near such a boundary are commonly subject to a “magnet effect” (Kuhl, 1991), a mechanism whereby the

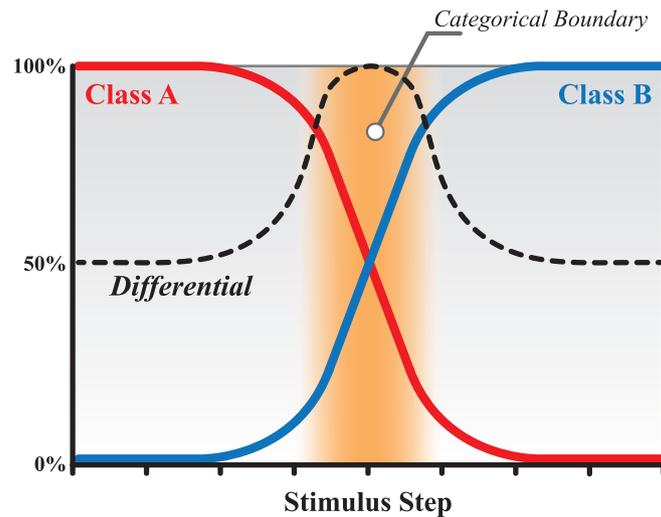


Figure 6.1: Example of the dynamics of class membership associated with Categorical Perception.

perception of the stimulus is *pulled* toward a particular well established category, thus resulting in a non-linear relationship between the stimulus continuum and the class membership of each stimulus within the continuum.

Figure 6.1 illustrates how two abstract *classes* are classically discretised, in a non-linear fashion, across a stimulus continuum consisting of linear physical differences between stimuli (see the red and blue lines). The figure also shows the *differentiation profile* of the two classes - as the stimulus steps approach the *categorical boundary*, people exhibit greater sensitivity to the differences between the stimuli.

The phenomenon of CP has been shown to take place during the processing of a broad variety of sensory information in both adults and children. Examples of this are during the processing of phonetic sounds (Liberman et al., 1957; Kuhl, 1991), colour (Bornstein et al., 1976; Franklin and Davies, 2004; Zhou et al., 2010), facial expressions (Bimler and Kirkland, 2001; Cheal and Rutherford, 2011) and affect in synthesised speech (Laukka, 2005). As such CP has been proposed as a fundamental foundation upon which human cognition is built (Harnad, 1987). For the interested audience, Repp (1984) provides a good overview of issues, experimental methods and findings surrounding the scientific study of CP.

The issue of CP also holds relevance for areas closer to HRI. Moore (2012)

has proposed that the Uncanny Valley effect (see chapter 4) may be a particular manifestation of CP, where, in the presence of multi-modal perceptual cues feeding into a category membership, conflicts in these cues could lead to the feelings of *discomfort* akin to those as described by Mori (1970). While this example specifically concerns the relation between the physical appearance of an agent and its physical behaviour, it may also be possible that similar effects occur between an agent’s physical appearance and the acoustic behaviour it exhibits, for instance, NLUs. This notion does however presuppose that NLUs are indeed subject to CP, the validity of which is investigated in this chapter, however not with respect to the relevance to the Uncanny Valley hypothesis, but rather to how subjects affectively interpret NLUs.

6.2 Experimental Setup

In psychological experiments, the typical methodology for testing CP involves producing a stimulus continuum in which there are at least two *prototype* perceptual categories represented, with all other members of the continuum providing equal, linear transitions between these prototypes. This presupposes that CP is occurring and that a readily recognisable (and established) categorical boundary exists at some point along the continuum. Subjects are then asked to complete two tasks: a *discrimination* task and an *identification* task. The purpose of the discrimination task is to determine whether subjects exhibit a perceptual *difference* between two stimuli: do the two (or more) stimuli fall into the same category or a different category, without explicitly declaring their class membership. This is done by presenting stimulus pairs (usually neighbouring stimuli) from the continuum and asking subjects whether they perceive them as *similar* or *different*. The identification task entails subjects assigning a category to each stimulus individually and explicitly declaring the class membership. This is done by presenting a single stimulus and asking subjects to rate it in some way, where the rating metric relates to the underlying representative categories (e.g. labels or sliders for dimensions, in the case of affect). The results for these two tasks together are

then used to assess whether CP is occurring.

From the discrimination task, indications of CP are that subjects rate neighbouring stimuli that cross a categorical boundary as different, while they rate neighbouring stimuli that sit within a categorical region as similar. Thus, one would expect to see the frequency of “different” ratings increase as the neighbouring stimulus pairs approach a categorical boundary, providing an inverted “V” differential profile (see the black differential profile line in figure 6.1).

In the identification task, indications of CP are when there are at least two clusters of neighbouring stimuli each situated near prototype stimuli that are closely rated, with clusters being separated by a sharp change (or step) in the average rating. This represents the crossing of a categorical boundary and is characterised by a clear step in the class membership profiles (see the red and blue class membership profiles in figure 6.1).

This experiment adopts the same basic methodology with some minor alterations to serve the focus on HRI, namely that utterances were embodied in the Nao robotic platform and a facial gesture tool was used for capturing affective ratings from subjects via the AffectButton tool (chapter 3). The remainder of this section details the experimental set-up, covering the stimuli, the three different tasks that were completed, and the overall experimental procedure.

Given these characteristics, it is possible to formulate the following conditions under which CP may be said to be occurring:

- C_1 : The two extreme/prototypes of the stimulus continuum receive class membership ratings during the Identification Task that are significantly different.
- C_2 : Subjects rate neighbouring stimuli in the continuum that are near a prototype stimulus as “different” to a degree that is not statistically above chance (this indicates the presence of a categorical region).
- C_3 : Subjects rate neighbouring stimuli that lay in the middle of the continuum as “different” to a degree that is above statistical chance, forming an

inverted “V” shape in the differential profile (this indicates the presence of a categorical boundary).

- C_4 : Stimuli that are near a particular prototype stimulus have the same class membership rating as the prototype and a significantly different rating to the *other* prototype stimulus in the continuum, forming a *step* function.
- C_5 : The peak in differential ratings during the Discrimination Task coincides with the step in the class membership rating in the Identification Task.

6.2.1 Utterance Stimuli

For this experiment, a stimuli set of a total of 12 utterances was produced, comprised of 2 continua (Set 1 and Set 2) each consisting of 6 utterances (Utter-0 to Utter-5), each with a different Utterance Parameter configuration (see chapter 3). Within each continuum there were two prototype utterances (Utter-0 and Utter-5) separated by 4 utterances with linear transitions in the four utterance parameters (table 6.1). Each utterance was comprised of five sound units, and across the two continua only the Pitch Contour specifications were different (see table 6.2), whilst within each continuum these combinations remained the same. The parameter specifications of the two *prototype* utterances came from the finding in chapter 5, where subjects were able to distinguish between the two different parameter configurations, which resulted in significantly different ratings along the Dominance dimension of the AffectButton (see section 5.3.5). As such, these two parameter configurations were used to represent the two extremes of each stimulus continuum. Spectrograms of Utter-0 to Utter-5 from Stimulus Set 1 and Set 2 are illustrated in figure 6.2. Note the differences in the Pitch Contour specifications as outlined in table 6.2.

To provide some rationale as to this Stimulus Set arrangement (having two sets rather than one), the human voice facilitates the efficient dual encoding of both semantic and affective information through the same acoustic channel (Picard, 1997; Banse and Scherer, 1996; Scherer, 2003), and has led to the question of what role does the pitch contour of natural language play in affective expression

Table 6.1: Utterance Parameter configurations for each utterance in both Stimulus Sets 1 and 2 (see figure 6.2).

Stimulus	Parameter Configuration			
	Base Freq	Freq Rang	S. Rate	P. Ratio
Utter-0	1500	1500	6	0.05
Utter-1	1333.33	1333.33	5.5	0.166
Utter-2	1166.67	1166.67	5	0.2833
Utter-3	1000	1000	4.5	0.4
Utter-4	833.33	833.33	4	0.5166
Utter-5	666.67	666.67	3.5	0.633

Table 6.2: Pitch Contour specifications for the utterances in Stimulus Set 1 and Set 2 (see figure 6.2).

Sound Unit	Stimulus Set	
	Set 1	Set 2
1	Flat	Rising-Falling
2	Falling	Flat
3	Rising	Falling-Rising
4	Falling-Rising	Rising
5	Rising-Falling	Falling

(Banziger and Scherer, 2005). Given the degree of acoustic simplification of NLUs in comparison to Natural Language, NLUs are generally considered to contain very little, if not any semantic content. This is an important distinguishing feature from Natural Language (chapter 2), and as a result NLUs do not afford this dual encoding of both semantics and affect. This raises an important issue: what is the role of the pitch contour in NLUs in portraying affect? This question has a practical motivation in that as this work has sights on the automated generation of NLUs (see chapter 7), there is a motivation to understand which properties of NLUs play a minor role in affective charging, and which do not. In essence, one is wanting to know which features of utterance, if any, play a minor role and can potentially be exploited through randomisation with the aim of generating *unique* and non-repetitive utterances that may carry similar affective charges. This is seen as a means of providing a wide variety of utterances that do not make the robot appear repetitive (and by extension, pre-programmed), as humans are quick to spot this, which can have negative impacts on HRI (Belpaeme et al., 2012). Given the large number of combinations of pitch contours for a given utterance with a given sound unit count, the pitch contour is an appealing feature with

which to start in this respect.

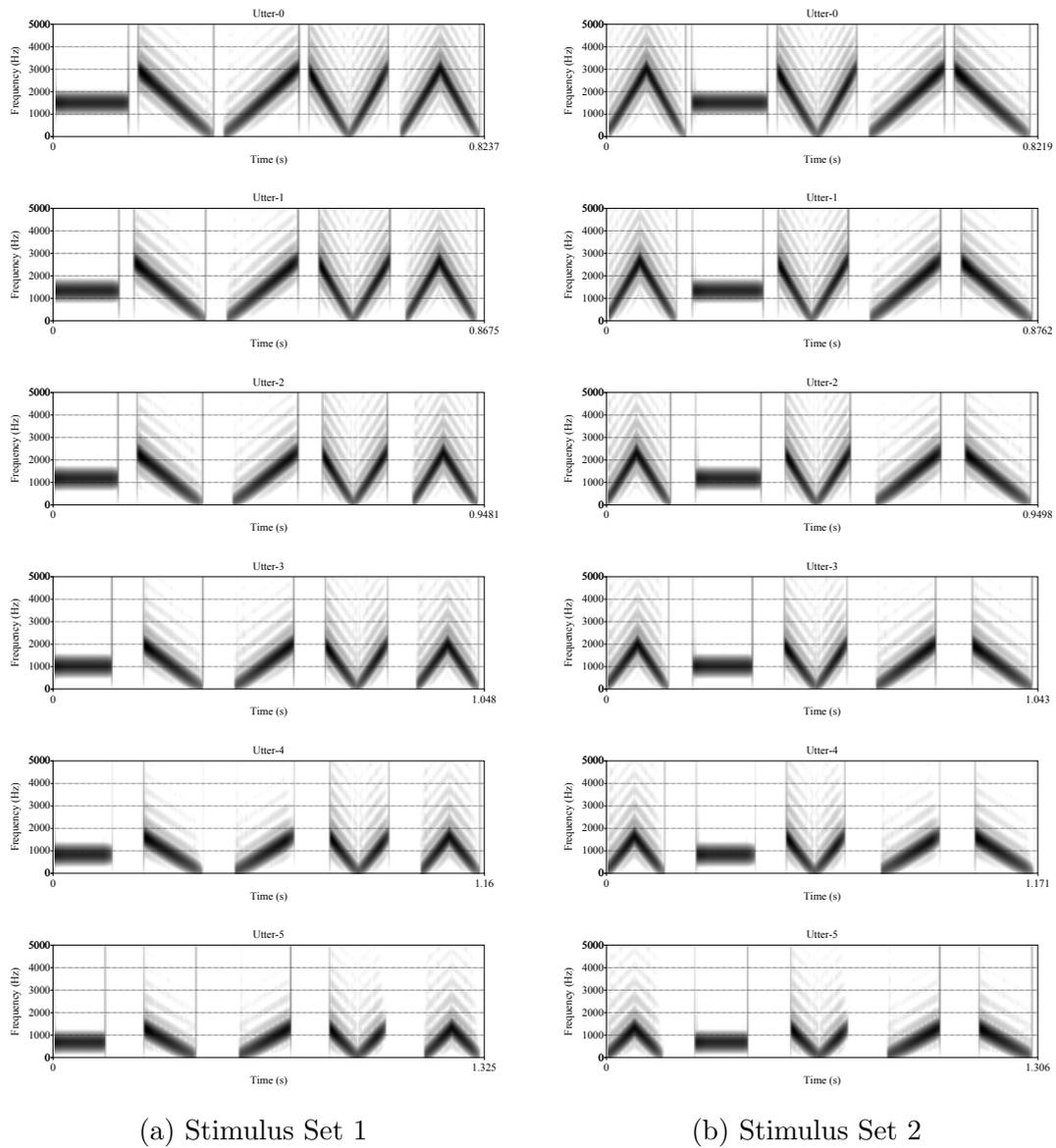


Figure 6.2: Spectrograms of the 12 NLUs used as the stimuli.

6.2.2 Labelling Task

In order to use the AffectButton as an effective tool for recording affective ratings, it is important to ascertain how coherent subjects are in their use of the tool. Thus, this task was aimed at forming an impression of the overall coherence between subjects in their use of the AffectButton. Subjects were given some time to familiarise themselves with the tool and explore the range of facial gestures that that can be displayed, and the associated mouse cursor locations onscreen.

Once familiarised, an affective label was then displayed on the laptop screen

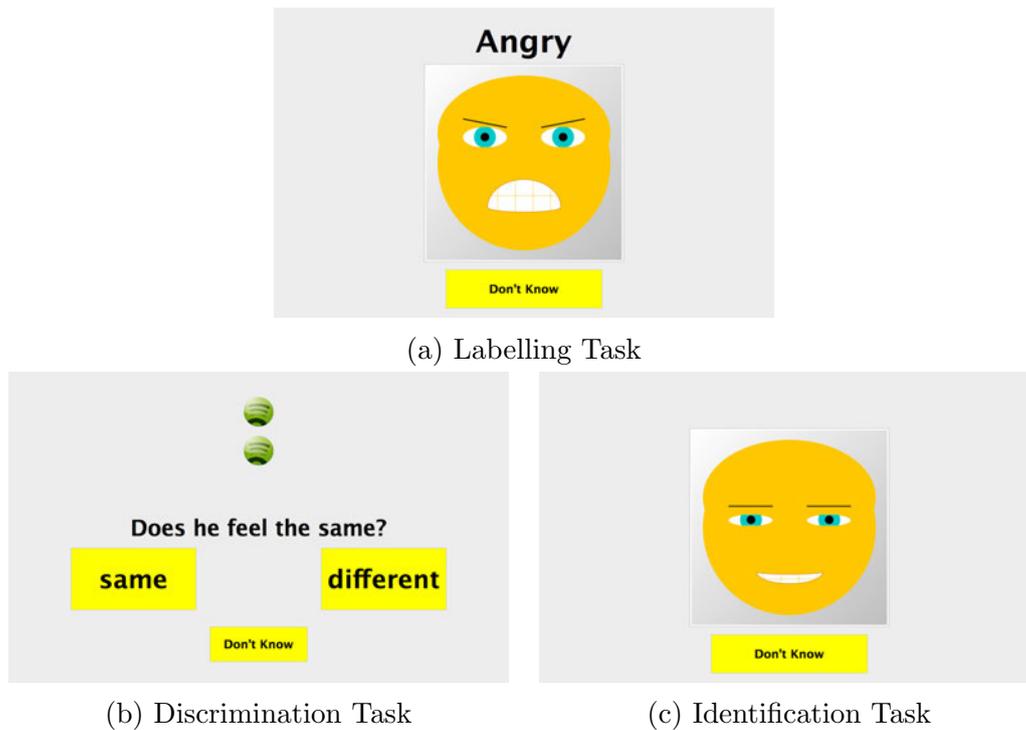


Figure 6.3: Images of the subjects' laptop screen during each of the three tasks in the experiment.

above the button, and subjects asked to match a face to the label (figure 6.3a). The affective labels that were used were: *Happy*, *Excited*, *Angry*, *Annoyed*, *Surprised*, *Scared*, *Sad*, *Calm* and *Relaxed* and were presented in a random order. This choice was motivated by the prototype facial expressions that are hard coded in the AffectButton (see chapter 3) and the overlap with the theory of basic emotions (Plutchik, 1994). As such, it was considered that a wide audience of subjects (e.g. children, adults, individuals from different cultural backgrounds and mother tongues, etc.) would also be familiar with these labels. The co-ordinates of the prototype facial expressions in the AffectButton PAD space and the associated affective labels are shown in table 6.3.

6.2.3 Discrimination Task

The discrimination task was performed using an AX discrimination paradigm (Cheal and Rutherford, 2011; Gerrits and Schouten, 2004), where two stimuli were presented in pairs, sequentially (but randomly ordered), and subjects were asked

Table 6.3: Affective co-ordinates in the AffectButton affect space of the labels (and associated prototypical facial gestures) used during the Labelling Task. Note that Calm and Relaxed are no prototypes used in the AffectButton.

Label	Affect Space Coordinate		
	Pleasure	Arousal	Dominance
Angry	-1	1	1
Annoyed	-0.5	-1	0.5
Happy	0.5	1	0.5
Excited	1	1	1
Sad	-0.5	-1	-0.5
Scared	-1	1	-1
Surprised	1	1	-1
Calm	-	-	-
Relaxed	-	-	-

to report whether they thought the robot *felt*¹ different or the similar between the stimuli. On their laptop screen, subjects could choose from either “same”, “different” or “don’t know” options. Stimulus pairs were presented by tapping the Nao on the head on the touch sensor. At each head tap, the subjects laptop screen went blank, and at the onset of each of the utterances a small visual symbol appeared on screen with the aim of aiding the subject to track which stimulus of the pair was being presented. Once both utterances had been played the options were then displayed on screen with the cursor reset to the centre of the screen. The latter was done in order to avoid subjects hovering the cursor over a particular response button (see figure 6.3b).

In total, each subject rated 13 pairs of utterances in this task. The first three pairs were test pairs whose order remained constant across all subjects (and were not used in the results analysis). These consisted of one stimulus pair with two extreme prototype utterances, one pair with identical utterances and one pair with neighbouring utterances. All of these stimuli were different from the actual experiment stimuli as they had a different *pitch contour*, and served the purpose to acquaint subjects with the format of the task.

The remaining 10 utterance pairs were neighbouring utterances (e.g. Utter-0 vs. Utter-1, Utter-1 vs. Utter-2, ect), which consisted of 5 pairs in each stim-

¹It was made explicit to the subjects that this was not a test of whether they could recognise whether stimuli were identical, but rather whether they felt that the sound portayed the same affective state.

Table 6.4: Overview of the neighbouring utterance pair comparisons in the Discrimination Task (note that A and X were randomly ordered).

AX Pair	Stimulus Set	Utterance Comparison	
		A	X
1	1	Utter-0	Utter-1
2		Utter-1	Utter-2
3		Utter-2	Utter-3
4		Utter-3	Utter-4
5		Utter-4	Utter-5
6	2	Utter-0	Utter-1
7		Utter-1	Utter-2
8		Utter-2	Utter-3
9		Utter-3	Utter-4
10		Utter-4	Utter-5

ulus set. This was done for both the utterances in Stimulus Sets 1 and 2, thus accounting for 10 utterance pairs in total. This is outlined in table 6.4.

6.2.4 Identification Task

This task involved presenting subjects with a single utterance stimulus and asking them to provide an affective interpretation by matching a facial expression on the AffectButton to their affective interpretation of the utterance (figure 6.3c). For this task, a simplified version of the AffectButton was used, where the facial gestures were limited to interpolate between only the *sad*, *neutral*, *happy* and *excited* prototypes as the mouse cursor was moved horizontally (vertical movement had no effect). In doing this, the Pleasure value was modulated via the horizontal cursor movement, and Dominance was then set equal to Pleasure (p) with both values falling in the range of $-0.5 \leq p \leq 1.0$. This Pleasure/Dominance mouse position mapping is shown in figure 6.4, with the corresponding prototypical facial expressions at these PAD coordinates in the AffectButton.

During typical psychological CP identification tasks it is common for subjects to select from a small set of category labels (e.g. happy, sad, angry), however doing this explicitly promotes the notion of splitting the stimulus continuum into two or more categories. The use of the AffectButton overcomes this by presenting subjects with a continuous scale of measurement. By using a continuum of

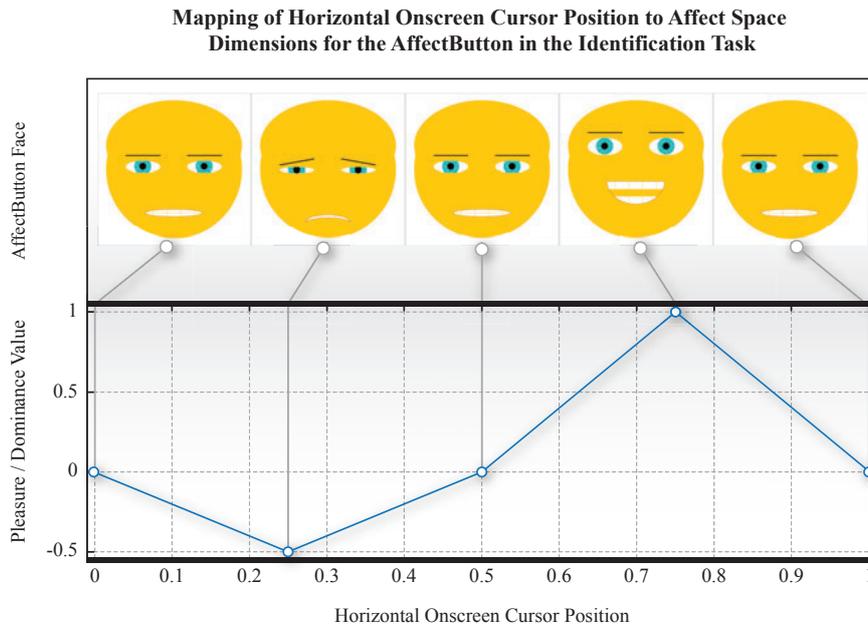


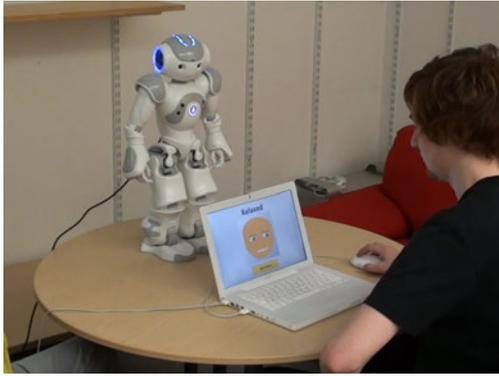
Figure 6.4: Plot of the Pleasure/Dominance values as a function of the horizontal onscreen cursor position and the resulting AffectButton prototype expressions associated with the PAD values (from left to right): *neutral* (0, -1, 0), *sad* (-0.5, -1, -0.5), *neutral* (0, -1, 0), *excited* (1, 1, 1), *neutral* (0, -1, 0).

possible facial expressions subjects are not forced to make an explicit categorical distinction, rather it leaves room for any CP to present itself in a more unrestricted manner.

Similarly to the discrimination task, the AffectButton face was hidden at the onset of an utterance and reappeared at the offset of an utterance in order to avoid any priming or habituation effects due to the presence of a face with an affective expression. Finally, by placing the extreme facial expressions at the upper and lower quartiles of the range of mouse movement, expression selection became a more cognitive task, avoiding subjects swinging to the extreme locations of the AffectButton (see figure 6.4).

6.2.5 Experimental Procedure

For the adult experiment, subjects were recruited (and performed the experiment individually) through advertisements located around the university campus, and the experiment took place within a lab setting and was conducted using two laptops and the Nao robot. The adult subjects were rewarded with £5 in cash at



(a) Adults



(b) Children

Figure 6.5: Image of the experimental setups for both the adult and children subjects.

the end of the experiment. The experiment with the school children was conducted at a local primary school, in a space classroom with children coming in pairs². These differences aside, the experiment was otherwise conducted in an identical manner between the adults and children.

All stimuli were played through the Nao’s built in speakers, embodying the utterances in a robotic agent. Furthering the notion of embodiment and agency, the Nao was also programmed to exhibit neutral behaviours (random gazing, shifting weight from foot to foot and subtly moving the arms and fingers) in order to provoke the “illusion of life” in subjects, and avoid it being perceived as a static object. Care was also taken to avoid subjects witnessing the robot starting up and loading the experiment software as this too could impact the perception of the robot.

A laptop was placed in front of the subject and was used to capture their responses for each task. Iteration through a given task was controlled via the Nao, where the touch sensors on the head were used to either play the next utterance(s) or repeat the current utterance(s). A second laptop was operated by the experimenter and was used to orchestrate and monitor the overall experiment from a global perspective, and also managing the information flow between the Nao and

²Children came in in pairs in order to set them at ease with being in a room with a stranger and a robot. While one child was performing the experiment, the other was invited to control the presentation of the utterances by tapping the robot on the head. Software was written such that a new utterance could only be presented when the child performing the experiment had provided a rating via their laptop. This avoided missing ratings for utterances. Thus, the child controlling the sounds had only limited control.

the subject's laptop and gathering all the data in one place. The experimental arrangements for both the children and adults is shown in figure 6.5.

Subjects were told that they needed to provide affective interpretations of the utterances made by the Nao, and that there were no correct or incorrect answers. They were also told that they should try and respond as quickly as possible and use their "gut feeling" so as to avoid over thinking the problem. The Labelling Task was completed first as this task was intended to enable the subjects to become familiarised with the AffectButton and place them in a frame of mind orientated around attributing affect to a robot. It is common practice in CP studies for the Discrimination Task to be completed before the Identification Task in order to avoid the process of assigning categories to stimuli biasing the process of discriminating between two stimulus pairs, a practice that was followed here. The total duration of the experiment was 20 minutes and once completed, subjects were free to ask any questions.

6.3 Results

In total, 27 school children (aged 7-8 years old) were recruited for the experiment: 15 girls, and 12 boys. 28 Adult subjects also partook: 17 females (mean age = 32.2, std = 10.6) and 11 males (mean age = 28.8, std= 6.8).

This section presents the results for each of the three tasks, beginning with the Labelling Task, then the Discrimination Task, and finally, the Identification Task.

6.3.1 Labelling Task Results

Figure 6.6 shows plots of the mean values and standard deviations for the ratings of each affective label, of both the adults and children. An initial visual inspection of the results revealed that both the adults and children provided a range of ratings that covered the majority of the AffectButton affect space. It is notable however that the adults provided more precise ratings in that they had smaller standard deviations, with the mean ratings that appeared to be closer to the coordinates in

the affect space where the prototype facial gestures associated with each label are located. The ratings of the children tended to have a larger standard deviation for each label, with the means tending more toward the centre of the affect space.

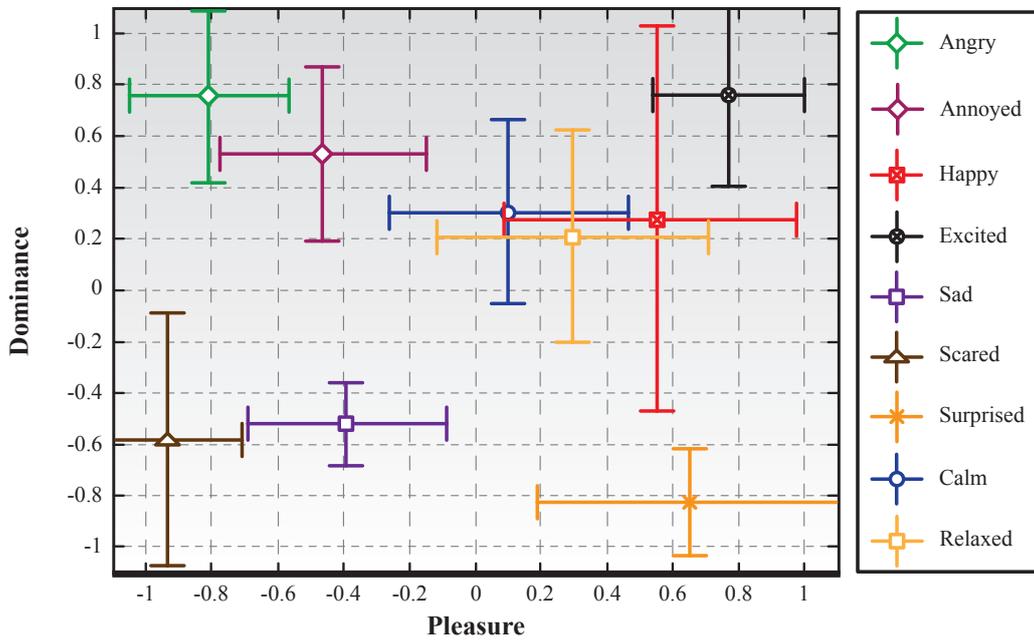
The ratings for some of the labels were found to *not* follow a normal distribution, and as such, non-parametric tests were employed to perform the analysis of the results in this task. Kruskal-Wallis tests were performed to check for overall differences between the genders in their ratings for each affective label, while Friedman tests were employed to identify whether there were significant differences in the ratings for the affective labels. Both the Kruskal-Wallis and main Friedman tests were followed up by post-hoc, pair-wise Friedman tests used to compare the ratings for two individual labels at a time. All of these tests were performed for each affective dimension individually.

6.3.1.1 Adult Subject Results

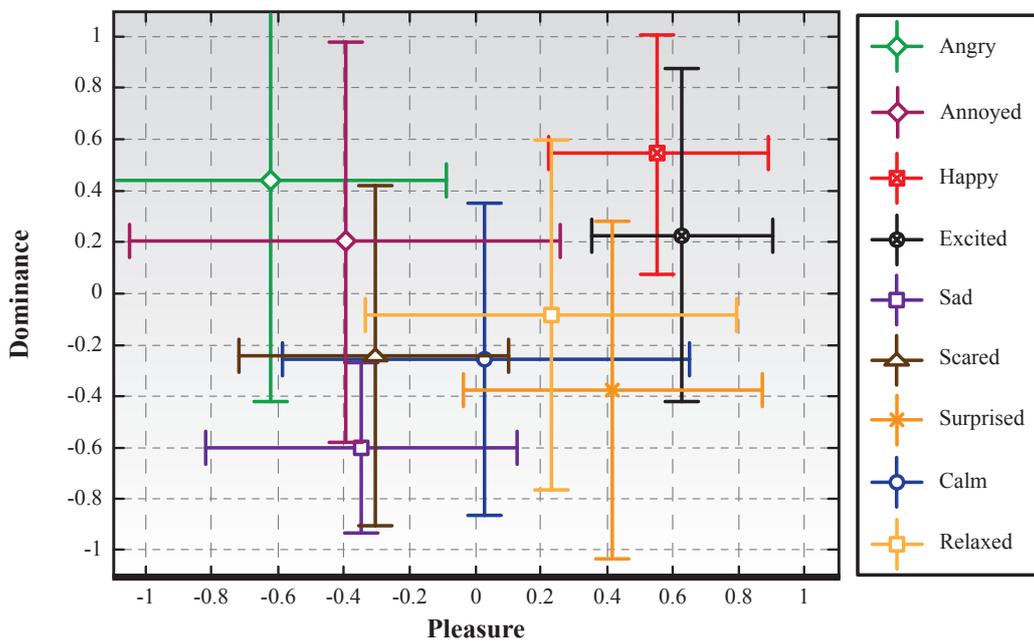
The Kruskal-Wallis tests found that there were no significant differences between the genders in their overall ratings along either the Pleasure ($\chi^2(1) = 1.29, p = 0.255$), Arousal ($\chi^2(1) = 0.09, p = 0.771$) and Dominance ($\chi^2(1) = 0.1, p = 0.751$) dimensions.

The Friedman tests, however, found that there were significant differences in how the affective labels were rated along the Pleasure ($\chi^2(8) = 184.35, p < 0.001$), Arousal ($\chi^2(8) = 143.55, p < 0.001$) and Dominance ($\chi^2(8) = 153.6, p < 0.001$) dimensions. Table 6.5 shows the χ^2 values calculated through the Friedman tests for pair-wise comparison. The values indicate that the adults provided significantly different ratings between the majority of the labels overall.

Perhaps what is more important to extract from these results, beyond the individual differences, is the demonstration that the subjects were able to clearly distinguish between the different affective labels to a degree of high statistical significance, along the dimensions that *differentiate* the (prototype) labels within the AffectButton affect space. For example, while the ratings for the *Happy* and *Angry* labels are not significantly different along the Arousal and Dominance



(a) Adult Subjects.



(b) Children.

Figure 6.6: Plots of the Mean values and Standard Deviations of the ratings for each affective label in the Labelling Task. These values are summarised in table D.1. Figures D.1 and D.2 show the AffectButton facial gestures for the mean values for the adults and children respectively.

dimensions (which is to be expected, see table 6.3), they were significantly different along the Pleasure dimension. Similarly, the *Surprised* and *Happy* labels were only significantly different along the Dominance dimension, again the key differentiating dimension in the AffectButton.

Such key, significant discriminations demonstrate the fact that subjects were indeed able to reliably use the AffectButton to assign significantly different facial gestures (and thus affective ratings) to the different affective labels, and that they were able to robustly and coherently rate the labels as different along the key affective dimension with which the labels are differentiated. For example, the *Happy*, *Excited* and *Surprised* labels are all associated with (and received) a high Pleasure rating, while the *Angry*, *Annoyed*, *Sad* and *Scared* labels are all associated (and received) low Pleasure ratings. Similarly, the *Happy*, *Excited*, *Angry* and *Annoyed* labels were all rated as having high Dominance and were all significantly different to the *Sad*, *Scared* and *Surprised* labels that all received low Dominance ratings. As such, subjects were clearly able to identify the differences between these affective labels, and robustly represent this via the AffectButton facial gestures. One result that does stand out however, is that there were no significant differences at all between the *Happy* and *Excited* labels, along any of the affective dimensions: subjects appear to have been unable to distinguish between these affective labels.

6.3.1.2 Child Subject Results

The Kruskal-Wallis tests comparing the two genders found that there was no significant difference between the genders in how they rated the different labels overall along the Pleasure ($\chi^2(1) = 2.77$, $p = 0.096$) and Dominance ($\chi^2(1) = 0$, $p = 1$) dimensions. However, there was a significant overall difference identified for the ratings along the Arousal ($\chi^2(1) = 14.36$, $p < 0.001$) dimension, with males (mean = 0.349, std = 0.654) providing a overall higher rating than the females (mean = -0.108, std = 0.933).

As with the adult subjects, the Friedman tests found that there were sig-

Table 6.5: Results of the post-hoc Friedman pairwise comparisons for the adults’s affective ratings for the affective labels in the Labelling Task. The table show the $\chi^2(1)$ results and indicate the associated p -value for each dimension of the AffectButton affect space independently.

Label		Ang	Ann	Hap	Exc	Sad	Scar	Sur	Calm	Rel
Angry	P	-								
	A									
	D									
Annoyed	P	14.29‡	-							
	A	17.64‡								
	D	12.45‡								
Happy	P	29‡	29‡	-						
	A	1.47	17.64‡							
	D	0.31	5.83*							
Excited	P	25.14‡	29‡	0.03	-					
	A	0.25	18.62‡	1.47						
	D	0.03	1.69	0.86						
Sad	P	12.45‡	0.31	29‡	25.14‡	-				
	A	22.15‡	2.13	27‡	24.14‡					
	D	25.14‡	29‡	25.14‡	15.21‡					
Scared	P	1.69	18.24‡	29‡	29‡	25.14‡	-			
	A	0.89	14.44‡	0.05	4.26*	20.57‡				
	D	25.14‡	25.14‡	25.14‡	15.21‡	2.79				
Surprised	P	25.14‡	25.14‡	0.03	1.69	25.14‡	29‡	-		
	A	0.6	18.62‡	2.88	0	24.14‡	2.25			
	D	29‡	29‡	29‡	21.55‡	15.21‡	7.76*			
Calm	P	25.14‡	18.24‡	25.14‡	21.55‡	15.21‡	25.14‡	15.21‡	-	
	A	24.14‡	6*	29‡	29‡	0	21.55‡	29‡		
	D	15.21‡	2.79	18.24‡	2.79	25.14‡	25.14‡	29‡		
Relaxed	P	25.14‡	21.55‡	11.57‡	9.97†	21.55‡	29‡	12.45‡	7*	-
	A	19.59‡	1.64	23.15‡	24.14‡	2.91	15.38‡	19.59‡	1.09	
	D	12.45‡	9.97†	21.55‡	0.86	21.55‡	17.29‡	25.14‡	1.69	

* : $p < 0.05$

† : $p < 0.005$

‡ : $p < 0.001$

nificant differences along the Pleasure ($\chi^2(8) = 119.04, p < 0.001$), Arousal ($\chi^2(8) = 42.51, p < 0.001$) and Dominance ($\chi^2(8) = 78.07, p < 0.001$) dimensions, indicating that there were significant differences in the ratings for the affective labels. Again Friedman post-hoc tests were used to perform pair-wise comprising between the ratings for each of the affective labels, along each of the affective dimensions individually. The results of these tests are shown in table 6.6. It was found that while the standard deviations of the ratings were notably larger than those of the adult subjects (see figure 6.6), there were still important, and statistically significant differences for various labels, along the affective dimensions that differentiate the labels. For example, the *Happy*, *Excited*, *Angry* and *Annoyed* labels were rated as having high Dominance and were significantly higher than the ratings for the *Sad*, *Scared*, and *Surprised* labels, which all received low Dominance ratings. This is generally the same overall result as was found with the adult subjects. Similarly, these labels were also found to have significantly different ratings along the Pleasure dimension: *Angry*, *Sad*, *Annoyed* and *Scared* all received low Pleasure ratings, while *Happy*, *Excited* and *Surprised* all received high Pleasure ratings.

Thus, as with the adult subjects, the children subjects did appear to demonstrate an ability to robustly distinguish between the different affective labels and translate this to their use of the AffectButton and assigning facial gestures to the labels.

6.3.2 Discrimination Task Results

To provide a quick reminder, the discrimination task involved presenting subjects with neighbouring utterances along the stimulus continuum and asking subjects to rate them as either having the “same” affective meaning or a “different” affective meaning.

Referring to figure 6.1, and specifically the *differential* profile, the indication of CP is that as the AX pairs of neighbouring stimuli approach the categorical boundary, these are rated as “different” to a degree that is above chance levels (i.e.

Table 6.6: Results of the post-hoc Friedman pairwise comparisons for the children’s affective ratings for the affective labels in the Labelling Task. The table show the $\chi^2(1)$ results and indicate the associated p -value for each dimension of the AffectButton affect space independently.

Label		Ang	Ann	Hap	Exc	Sad	Scar	Sur	Calm	Rel
Angry	P	-								
	A									
	D									
Annoyed	P	0.14	-							
	A	2.91								
	D	7*								
Happy	P	27‡	23.15‡	-						
	A	0.05	1.64							
	D	0.33	4.48*							
Excited	P	26‡	15.38‡	1	-					
	A	0	2.91	0.06						
	D	3.85*	0	4.48*						
Sad	P	1.29	0.04	27‡	22.15‡	-				
	A	15.38‡	13.5‡	6*	11.64‡					
	D	17.29‡	5.14*	27‡	9.85‡					
Scared	P	2.29	0.14	27‡	26‡	0.57	-			
	A	2.91	0.62	1.09	0.18	3.85*				
	D	24.14‡	3.57*	13.37‡	3.85*	0.57				
Surprised	P	23.15‡	17.29‡	0.33	1.38	17.29‡	17.29‡	-		
	A	0.43	6*	2	1.8	15.38‡	4.48*			
	D	14.29‡	7*	19.59‡	5.54*	2.29	1.29			
Calm	P	7*	7*	8.33*	12.46‡	8.33‡	7*	5.14*	-	
	A	9.85‡	0.93	6.55*	4.17*	2.46	4.48*	12.57‡		
	D	13.37‡	1.29	19.59‡	3.85*	6.26*	2.29	2.29		
Relaxed	P	17.29‡	5.14*	4.48*	7.54*	19.59‡	11.57‡	3.57*	4.48*	-
	A	6.26*	1.38	2.13	6.55*	2.13	4.17*	12.46‡	0.36	
	D	14.29‡	3.57*	19.59‡	1.38	1.81	2.29	3.57*	1.81	

* : $p < 0.05$

† : $p < 0.005$

‡ : $p < 0.001$

tending toward a 100% “different” rating) while utterance pairs that lay within a categorical region received differential ratings that remain at chance level (i.e. tending toward a 50% “different” rating). As such, the differential profile follows an inverted “V” shape.

As such, χ^2 goodness-of-fit tests were performed to identify which of the ratings were *above* chance and which were at chance level, comparing each AX utterance pair against a flat, uniform distribution where 50% of the ratings would be classed as having the “same” affective meaning, and 50% classed as having a “different” affective meaning. These tests were performed for the results for results for Stimulus Sets 1 and 2 independently, as well as for each of the genders individually.

Two-way independent samples χ^2 tests were employed to check for statistically significant differences between the two genders in the distributions of their ratings for each of the utterance pairs presented.

6.3.2.1 Adult Subjects

Figure 6.7a shows bar graphs of the percentage of “different” ratings for each of the neighbouring utterance pairs in Stimulus Set 1, showing the ratings for all the subjects as well as for the two genders individually. Figure 6.7b shows the same for the utterance pairs in Stimulus Set 2. Upon visual inspection, it can be seen that the ratings appear to follow the inverted “V” profile, with this being more prominent for the utterances in Set 1 than in Set 2. It is also notable that there is a general skew in the ratings, with the highest ratings occurring for the comparison of Utter-3 vs. Utter-4 overall.

Table 6.7 shows the results of the χ^2 goodness-of-fit tests across the three different subject gender breakdowns (both, females and males), and for each Stimulus Set. With respect to Stimulus Set 1, AX pairs comparing Utter-2 vs. Utter-3 ($\chi^2(1, N = 28) = 3.572, p = 0.05$), and Utter-3 vs. Utter-4 ($\chi^2(1, N = 28) = 3.572, p = 0.05$) were found have ratings that were above chance levels. For the male subjects only Utter-2 vs. Utter-3 ($\chi^2(1, N = 11) = 4.455, p = 0.035$) were found to be above chance, while for the female subjects none of the ratings were

Table 6.7: χ^2 Goodness of fit tests for the adult subjects' comparison of neighbouring utterances in each of the Stimulus Sets.

Set	Utterance		Subjects					
	A	X	Both		Females		Males	
			Rating (%)	$\chi^2(1)$	Rating (%)	$\chi^2(1)$	Rating (%)	$\chi^2(1)$
1	Utter-0	Utter-1	46.667	0.310	44.444	0.529	50.000	0.000
	Utter-1	Utter-2	51.724	0.143	47.059	0.000	58.333	0.333
	Utter-2	Utter-3	65.517	3.572*	55.555	0.529	81.818	4.455*
	Utter-3	Utter-4	65.857	3.572*	70.588	2.882	63.636	0.818
	Utter-4	Utter-5	59.259	0.926	60.000	0.600	58.333	0.333
2	Utter-0	Utter-1	60.000	1.690	50.000	0.059	75.000	3.000
	Utter-1	Utter-2	72.414	7.000**	64.706	2.250	83.333	5.333*
	Utter-2	Utter-3	68.966	5.143*	61.111	1.471	81.818	4.455*
	Utter-3	Utter-4	93.103	20.571†	94.444	13.235**	90.909	7.364**
	Utter-4	Utter-5	70.000	5.828*	61.111	1.471	83.333	5.333*

*: $p < 0.05$

** : $p < 0.01$

†: $p < 0.005$

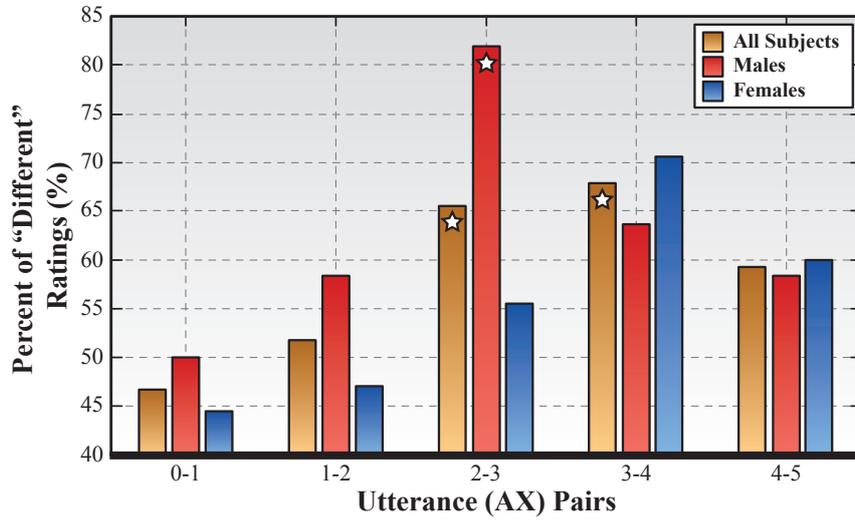
found to be above chance. For the utterances in Stimulus Set 2, the table shows that overall, only the ratings for Utter-0 vs. Utter-1 ($\chi^2(1, N = 28) = 1.690$, $p = 0.194$) were *not* above chance, with the same being true for the male subjects also ($\chi^2(1, N = 11) = 3.000$, $p = 0.083$). With respect to the female subjects, only Utter-3 vs. Utter-4 ($\chi^2(1, N = 17) = 13.235$, $p = 0.001$) *was* found to be above chance.

Finally, the two-way independent samples χ^2 test found no significant differences in the distribution of the ratings between the genders for any of the utterances pairs, for both Stimulus Sets. These results are summarised in table 6.8.

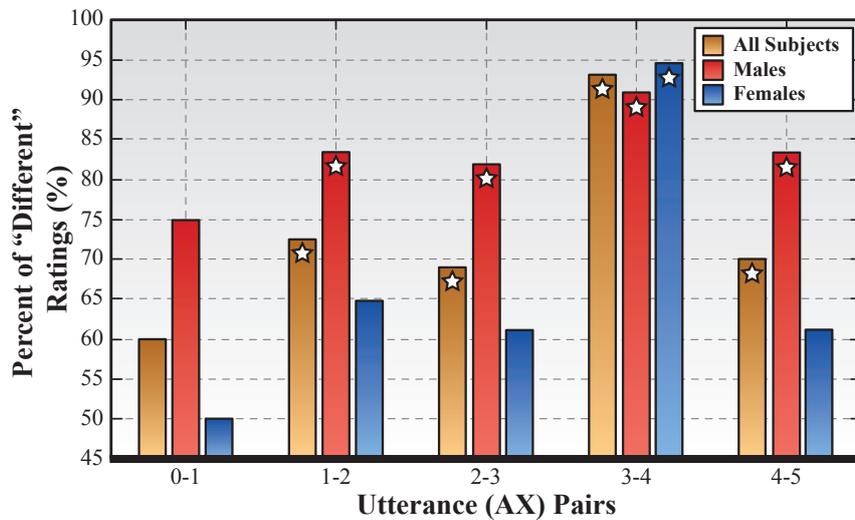
6.3.2.2 Children Subjects

Figure 6.8a shows bar graphs of the percentage of “different” ratings for each of the neighbouring utterance pairs in Stimulus Set 1, showing the ratings for all the subjects as well as for the two genders individually. Figure 6.8b shows the same for the utterance pairs in Stimulus Set 2. Upon visual inspection there appears to be no real trend in how the subjects have rated the utterances pairs. Furthermore, it appears that the vast majority of ratings are around chance level (50% rating).

Table 6.9 shows the results of the χ^2 goodness-of-fit tests for the utterances in Stimulus Sets 1 and 2 independently, across the three different subject gender



(a) Utterances Pairs in Stimulus Set 1.



(b) Utterances Pairs in Stimulus Set 2.

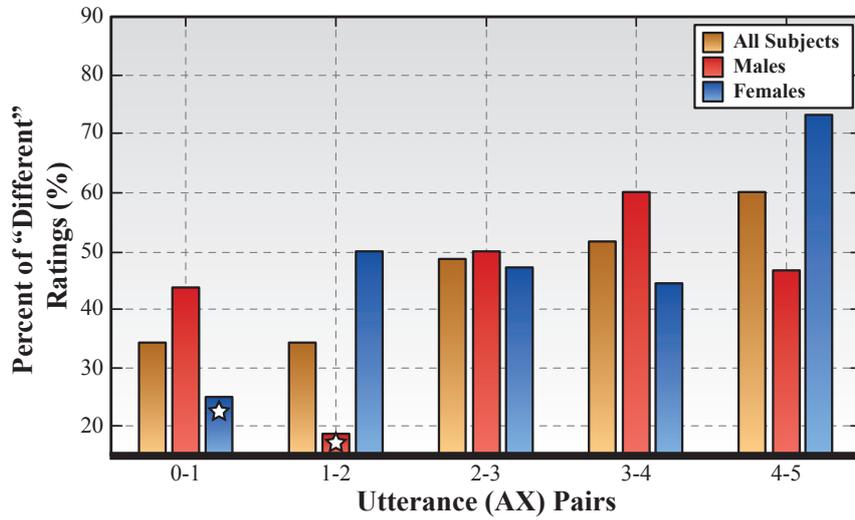
Figure 6.7: Bar graphs showing the percentage of “different” ratings given by the adults for the neighbouring utterance AX pairs for both Stimulus Sets. Bars marked with a star are ratings found to be significantly above chance at the 0.05 level. The ratings shown in this figure are summarised in table 6.7.

Table 6.8: Results of the two way independent samples χ^2 tests checking for significant differences between the genders in their ratings of each utterance pair. The table shows the χ^2 statistic and p value for both the adult and child subjects across the two Stimulus Sets.

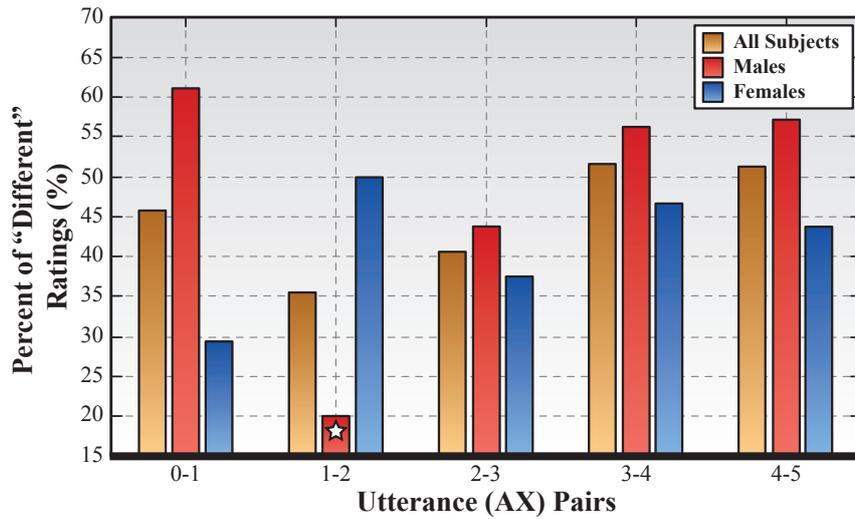
Set	Utterance		Subjects			
	A	X	Adults		Children	
			$\chi^2(1)$	p value	$\chi^2(1)$	p value
1	Utter-0	Utter-1	0.221	0.638	0.987	0.320
	Utter-1	Utter-2	0.191	0.662	1.697	0.193
	Utter-2	Utter-3	1.619	0.203	0.1922	0.166
	Utter-3	Utter-4	0.148	0.700	0.675	0.411
	Utter-4	Utter-5	0.008	0.930	1.025	0.311
2	Utter-0	Utter-1	1.454	0.228	0.512	0.474
	Utter-1	Utter-2	0.778	0.378	0.860	0.354
	Utter-2	Utter-3	0.958	0.328	0.261	0.609
	Utter-3	Utter-4	0.104	0.747	0.000	1.000
	Utter-4	Utter-5	1.0222	0.269	0.011	0.916

breakdowns (both, females and males). For Set 1, it was found that for all the subjects, none of the ratings were significantly different from a chance level distribution. With respect to the female subjects only the ratings for Utter-0 vs. Utter-1 ($\chi^2(1, N = 15) = 4.000, p = 0.046$) were significantly different to chance, with the majority of subjects rating the utterances as the “same”, while for the male subjects, only the ratings for Utter-1 vs. Utter-2 ($\chi^2(1, N = 12) = 6.250, p = 0.012$) were significantly different than chance, again with subjects rating these utterances as the “same”. Similarly, for the utterances in Stimulus Set 2, for all the subject combined, it was found all the ratings were not significantly different to a chance level distribution. The same was also true for the female subjects. For the male subjects, it was found that only the ratings for Utter-1 vs. Utter-2 were below chance level ($\chi^2(1, N = 12) = 0.540, p = 0.02$), with subjects rating these as the “same” in the majority.

Finally, the two-way independent samples χ^2 test found no significant differences in the distribution of the ratings between the genders for any of the utterances pairs, for both Stimulus Sets. These results are summarised in table 6.8.



(a) Utterances Pairs in Stimulus Set 1.



(b) Utterances Pairs in Stimulus Set 2.

Figure 6.8: Bar graphs showing the percentage of “different” ratings given by the children for the neighbouring utterance AX pairs for both Stimulus Sets. Bars marked with a star are ratings found to be significantly above chance at the 0.05 level. The ratings shown in this figure are summarised in table 6.9.

Table 6.9: χ^2 Goodness of fit tests for the child subjects' comparison of neighbouring utterances in each of the Stimulus Sets.

Set	Utterance		Subjects					
	A	X	Both		Females		Males	
			Rating (%)	$\chi^2(1)$	Rating (%)	$\chi^2(1)$	Rating (%)	$\chi^2(1)$
1	Utter-0	Utter-1	34.375	3.125	25.000	4.000*	43.750	0.250
	Utter-1	Utter-2	34.375	3.125	50.000	0.000	18.750	6.250*
	Utter-2	Utter-3	48.485	0.030	47.059	0.059	50.000	0.000
	Utter-3	Utter-4	51.515	0.030	44.444	0.222	60.000	0.600
	Utter-4	Utter-5	60.000	1.200	73.333	3.268	46.667	0.068
2	Utter-0	Utter-1	45.714	0.257	29.412	2.882	61.111	0.889
	Utter-1	Utter-2	35.484	2.613	50.000	0.000	20.000	0.540*
	Utter-2	Utter-3	40.625	1.125	37.500	1.000	43.750	0.250
	Utter-3	Utter-4	51.613	0.032	46.667	0.067	56.250	0.250
	Utter-4	Utter-5	51.351	0.027	43.750	0.250	57.143	0.429

*: $p < 0.05$

** : $p < 0.01$

†: $p < 0.005$

6.3.3 Identification Task Results

This section presents the results of the Identification Task, for both the adults and children independently. To provide a quick recap, referring to figure 6.1, during the Identification Task, stimuli that fall within a categorical region are usually giving similar ratings, with the different category regions having different overall ratings. This results in a non-linear step function in the ratings occurring over the categorical boundary.

Cronbach's α values were calculated for the pleasure ratings provided in order to assess the degree of agreement between subjects in their ratings for each of the stimuli. After this, a three-way (6x2x2) repeated measures ANOVA³ was performed for the Pleasure ratings, using the six Utterance Parameter configurations (within-subjects), two Stimulus Sets (within-subjects) and subject gender (between-subjects) as the three different factors. All significant main effects were followed up with post-hoc multi-comparison tests with Bonferoni corrections.

³An ANOVA was used here rather than a MANOVA as the AffectButton was modified such that the Pleasure and Dominance ratings were equal, and as the Arousal is a value calculated from these (chapter 3), there is little sense in performing tests for all three affective dimensions.

6.3.3.1 Adult Subject Results

The calculated Cronbach's α values indicate that overall the adults had a high degree of agreement in their use of the AffectButton during this task ($\alpha = 0.766$). Equally, the males ($\alpha = 0.757$) and females ($\alpha = 0.733$) also had equally high overall agreement as evidenced by their α levels.

The ANOVA found main effects due to both the Utterance Parameter configuration ($F(5, 125) = 64.269$, $MSE = 6.376$, $p < 0.0001$) and the Stimulus Set ($F(1, 25) = 9.082$, $MSE = 0.927$, $p = 0.006$). Also, a near significant effect was found due to subject gender ($F(1, 25) = 97.891$, $MSE = 1.416$, $p = 0.057$), however, no two-way or three-way interaction effects were found.

With respect to the main effect due to the Utterance Parameter configurations, the post-hoc multi-comparison tests revealed that there were significant differences between the majority of different Utterance Parameter configurations. Specifically, it was found that Utter-0 (mean = 0.748, 95% CI = [0.635 0.862]) received the highest rating and Utter-5 (mean = -0.115, 95% CI = [-0.229 - 0.002]) the lowest and that these were significantly different ($p < 0.01$). All the other utterances presented a negative gradient of ratings, regardless of the Stimulus Set. Table 6.10 presents the mean values, standard errors and 95% confidence intervals for each of the Utterance Parameter specifications. The tests also showed that the ratings for Utter-0 and Utter-1 were not found to be significantly different ($p = 1.0$), while both were found to be significantly different from all the other utterances ($p < 0.05$). Similarly, Utter-3 and Utter-4 were not found to be significantly different ($p = 1.0$) but too were significantly different from all the other utterances ($p < 0.05$). Finally, Utter-2 and Utter-5 were significantly different from all the other utterances ($p < 0.05$).

For the main effect due to the Stimulus Sets, the post-hoc tests revealed that the ratings for the utterances in Set 2 (mean = 0.388, 95% CI = [0.307 0.470]) received overall higher ratings than the utterances Set 1 (mean = 0.279, 95% CI = [0.203 0.355]), $p = 0.006$.

Figure 6.9a shows a plot of the ratings for each of the utterances overall, as

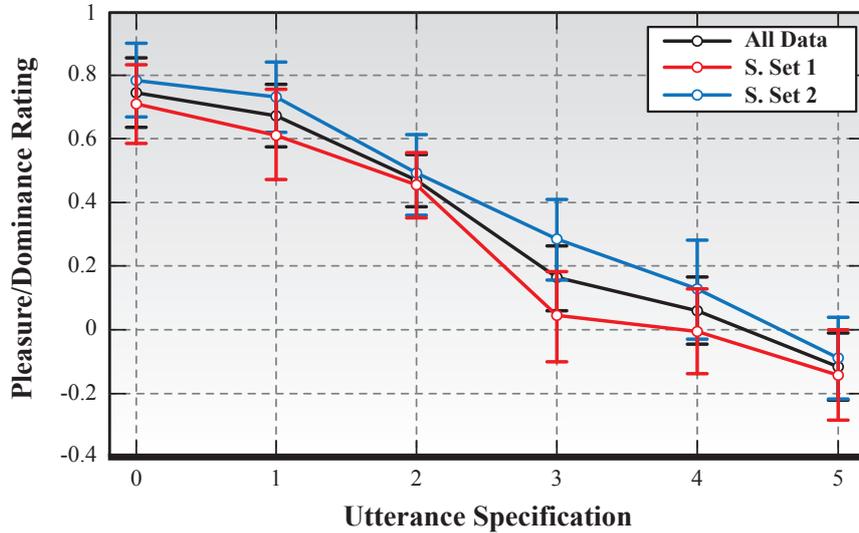
Table 6.10: Mean values, Standard Errors and 95% Confidence Intervals of the adult ratings for each of the Utterance Parameter configurations.

Utter	Mean	Standard Error	95% Confidence Interval	
			Lower Bound	Upper Bound
0	0.748	0.055	0.635	0.862
1	0.673	0.051	0.567	0.779
2	0.473	0.043	0.385	0.561
3	0.163	0.053	0.053	0.273
4	0.061	0.054	-0.051	0.173
5	-0.115	0.055	-0.229	-0.002

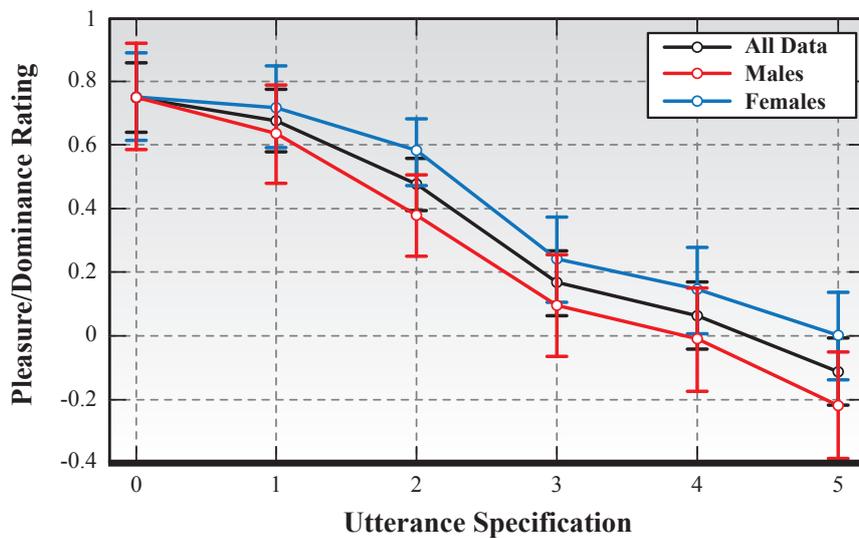
well as the ratings for the two Stimulus Sets. Also, while the effect due to gender was not found to be statically significant, it was found that the females (mean = 0.401, 95% CI = [0.312 0.490]) provided marginally higher ratings overall in comparison to the males (mean = 0.267, 95% CI = [0.160 0.374]). This is shown in figure 6.9b.

A visual inspection of figure 6.9a reveals an interesting observation in that the the ratings for utterances in Stimulus Set 1 appear to resemble a *step function* while those in Set 2 do not: rather they appear to follow a *linear function*. As shown in figure 6.1, non-linear rating profiles are a telling characteristic of CP.

To investigate this further, a one-way repeated measures ANOVA was performed using only the ratings for Stimulus Set 1 (see figure 6.9a, the red line), using the Utterance specification as the within-subjects factor. Again, a significant main effect was found for the Utterance specification ($F(5, 135) = 38.521$, $MSE = 3.737$, $p < 0.0001$). The post-hoc multi-comparison tests revealed that the six utterance specifications were grouped into two clusters. Utter-0 (mean = 0.719, 95% CI = [0.598 0.841]), Utter-1 (mean = 0.635, 95% CI = [0.491 0.780]) and Utter-2 (mean = 0.481, 95% CI = [0.375 0.586]) formed one cluster, with none of these having significant differences in their ratings. Similarly, Utter-3 (mean = 0.043, 95% CI = [-0.098 0.184]), Utter-4 (mean = -0.016, 95% CI = [-0.146 0.115]) and Utter-5 (mean = -0.126, 95% CI = [-0.275 0.024]) formed a second cluster, also with no significant differences in the ratings within this cluster. Furthermore, all the members of the two clusters were significantly different from all the members of the other cluster ($p < 0.001$).



(a) Comparing the Stimulus Sets. See figures D.3, D.4 and D.5 for the AffectButton faces associated with the mean values shown in this figure.



(b) Comparing the Genders.

Figure 6.9: Plots showing the mean values and 95% confidence intervals of the adult ratings of each Utterance Parameter specification across the Stimulus Sets (figure 6.9a), and subject genders (figure 6.9b). The descriptive statistics for these figures are summarised in tables D.2 and D.3.

Table 6.11: Mean values, Standard Errors and 95% Confidence Intervals of the child ratings for each of the Utterance Parameter configurations.

Utter	Mean	Standard Error	95% Confidence Interval	
			Lower Bound	Upper Bound
0	0.511	0.065	0.376	0.645
1	0.402	0.081	0.233	0.570
2	0.226	0.076	0.067	0.385
3	0.333	0.078	0.171	0.496
4	0.224	0.080	0.057	0.392
5	0.205	0.079	0.041	0.368

6.3.3.2 Child Subject Results

Overall, the children showed a notably lower level of agreement ($\alpha = 0.403$) in their ratings when compared to the adult subjects. When splitting the ratings by gender, the male subjects did show relatively high agreement in their ratings ($\alpha = 0.657$). The females however did not have a high degree of agreement ($\alpha = 0.024$).

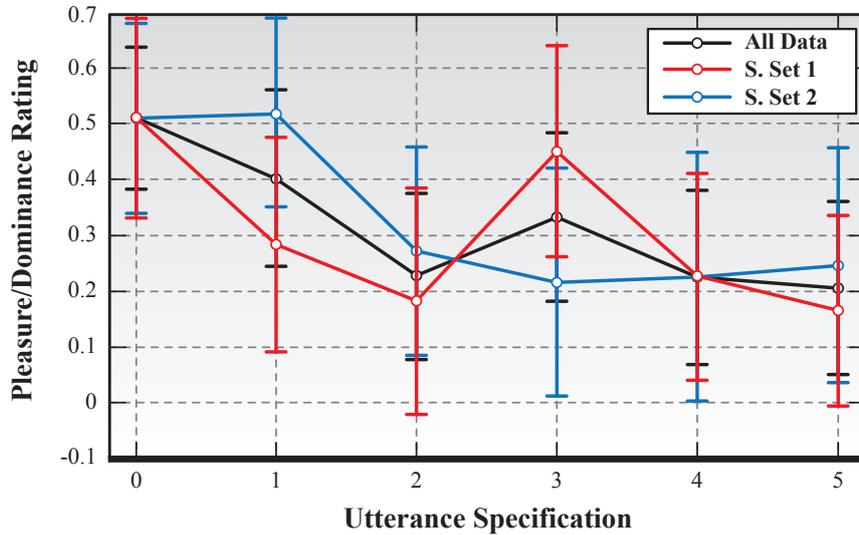
The ANOVA found only a significant main effect due to the Utterance Parameter configuration ($F(5, 105) = 2.730$, $MSE = 0.673$, $p = 0.023$). No other effects were identified. The post-hoc multi-comparison tests revealed that, unlike the results for the adult subjects, none of the different Utterance Parameter specifications were found to be different to a statistically significant degree. The mean values, standard errors and 95% confidence intervals are summarised in table 6.11.

To facilitate visual comparison with the results obtained from the adults subjects, figure 6.10 shows plots of the ratings for the different utterance specifications, across the two Stimulus Sets (figure 6.10a) and across the two genders (figure 6.10b).

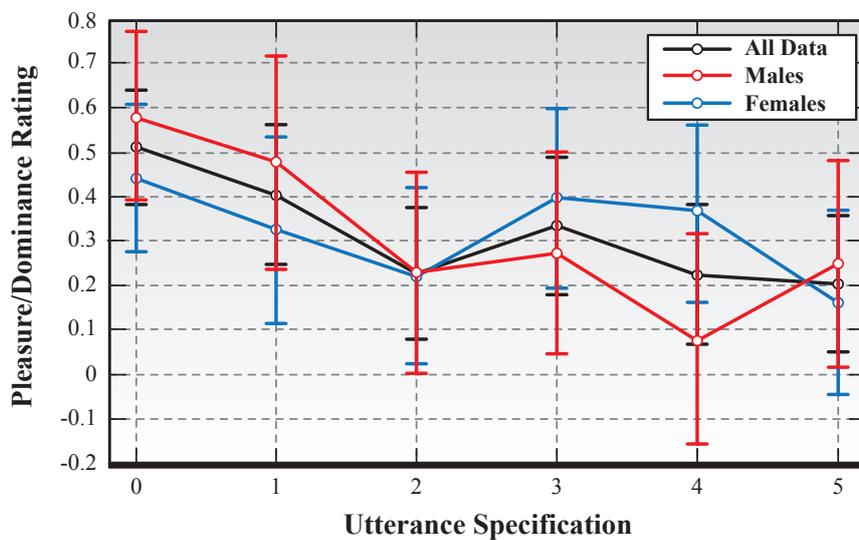
6.3.4 Summary of Results

As the results section has presented a dense population of results and statistical figures, this section serves to provide a overview summary of the important results that are to considered within the Discussion section.

The results of the Labelling Task show that both the adult and child subjects



(a) Comparing the Stimulus Sets. See figure D.6, D.7 and D.8 for the AffectButton faces associated with the mean values shown in this figure.



(b) Comparing the Genders.

Figure 6.10: Plots showing the mean values and 95% confidence intervals of the child ratings of each Utterance Parameter specification across the Stimulus Sets (figure 6.10a), and subject genders (figure 6.10b). The descriptive statistics for these figures are summarised in tables D.4 and D.5.

were able to clearly distinguish between the different affective labels by providing affective ratings for the labels that had, in the majority, statistically significant differences along the affective dimensions along which the affective labels are differentiated in a the AffectButton affect space. It was also found that there were no significant differences in the ratings between the genders, for either the adults or children. Overall, this provides strong evidence indicating that the subjects were indeed able to use the AffectButton in a robust and coherent manner, promoting confidence in the use of the AffectButton as the means of capturing the affective rating of utterances during the Identification Task.

With respect to the Discrimination Task, the differential ratings provided by the adult subjects tended to follow the inverted “V” profile that is associated with the presence of CP. This was more prominent with the ratings (of utterances being rated as “different) for utterances in Stimulus Set 1 than the utterances in Set 2. Furthermore, this is evidenced further by the finding that the ratings for the utterance pairs at either end of the continuum were found to be at chance level, while the pairs in the middle of the continuum tended to be above chance, which again follows the characteristics that are associated with CP. This was not found to be the case with the ratings provided by the children, whose results did not appear to follow the inverted “V” trend, and where in the vast majority, the differential ratings were found to be at chance level.

Finally, for the Identification Task, the affective ratings provided by the adults followed a clear negative gradient. Furthermore, the ratings for the utterances in Stimulus Set 1 were also found to have a statistically significant *step* function where there were two clusters of results, with all the members of one cluster receiving ratings that were significantly different to all the members in the other cluster. This is also a characterisation that is associated with CP. The ratings of the utterances in Stimulus Set 2 however did not follow this step function, but rather followed a *linear* trend which is not associated with the presence of CP. It was also found that for both Stimulus Sets 1 and 2 Utter-0 and Utter-5 were received ratings that were significantly different, showing that they did indeed

pot ray different affective states as evidenced via the facial expressions. With respect to the child subjects, these ratings in general followed no specific trend, and while there was an overall significant main effect due to the different utterance specifications identified, no significant differences were found between the different Utterance Parameter configurations via the post-hoc tests.

6.4 Discussion

This section provides a discussion of the results obtained during this experiment, and their relevance to the presence of Categorical Perception in subjects affectively rating NLUs. Following this, some methodological issues and drawbacks that may have influenced the outcome of the experiment are discussed. Finally, the findings of this chapter are discussed in a slightly broader perspective regarding HRI.

6.4.1 Results Discussion

It is clear from the results from the Discrimination and Identification Tasks that the children performed very differently during this experiment than the adults. This is notable through the lack of clear ratings in both tasks for the children: in the discrimination task the vast majority of ratings were all at chance level, and in the identification task, the low Cronbach's α values and general lack of any clear profile in affective ratings. However, the statistical analysis of the Labelling Task suggests that the children were able to use the AffectButton to assign facial expressions to affective labels and differentiate between these labels along key affective dimensions, in a coherent manner. This implies that either the children did not fully understand the tasks in the experiment, or that they found it difficult to coherently interpret the utterances on an affective level. In either case, it is clear that the subjects did not exhibit CP when rating the utterances during either the Discrimination or Identification Tasks. Rather, their ratings may be considered as more *random* and varied, which is in line with the findings in chapter 5.

The results from the adult subjects present a more interesting story however. The first observation that stands out is that the adults were clearly more coherent

in the ratings that they provided through each of the three tasks. The results of the Labelling Task show that the adults were indeed coherent in their use of the AffectButton, and were able to reliably associate different facial gestures to the different affective labels to a statistically significant degree. This lends strong support to the validity of the affective ratings captured by the AffectButton during the Identification Task.

With regard to the Discrimination and Identification Tasks, and their relation to CP, these two tasks should be considered simultaneously. During the Discrimination Task both the results for Stimulus Set 1 and Stimulus Set 2 tended to follow an inverted “V” shape, with the ratings located at the top of the V being statistically above chance while those at the bottom of the “V” were not significant above chance, though this was more prominent for the ratings for Stimulus Set 1. This supports both C_2 (that neighbouring stimuli near a prototype stimulus are rated as “different” to a degree that is not above chance) and C_3 (the neighbouring stimuli in the middle of the continuum were rated as “different” to a degree that was statistically above chance and formed an inverted “V” shape), and alone already suggests the presence of a categorical boundary in the region of the stimulus continuums where the peaks are highest. It is interesting to see that in the case of Stimulus Set 1 the profile of the affective ratings in the Identification Task followed a step function (between Utter-2 and Utter-3), which supports C_4 (that stimuli near a prototype stimulus have the same class membership and have a significantly different rating to the stimuli near the other prototype stimulus). Furthermore, this step coincides roughly with the peak in the corresponding results of the Discrimination Task (Utter-2 vs. Utter-3, and Utter-3 v.s Utter-4), which supports C_5 (that the peak of the inverted “V” in the differentiation profile and the step in the category membership profile occur at the same location in the continuum). When marrying these two sets of results together, there is strong evidence suggesting the presence of a categorical boundary along the Pleasure/Dominance dimension of the AffectButton affect space.

The Identification Task also found that there was a significant main effect due

to the difference in the Stimulus Sets, (which were differentiated by their Pitch Contour specification), and thus these results show (as the results in chapter 5 do also) that the Pitch Contour does appear to play a significant role in how subjects affectively interpret a NLU. This was also evident, though not specifically tested for during the Discrimination Task where there was a notable visual difference in the distribution of results between Stimulus Sets 1 and 2 (see figure 6.9).

There have also been some missed opportunities during the Discrimination Task, where some potentially insightful utterance pairs were not presented to subjects. For example, there were no comparisons made between corresponding utterances *across* the two Stimulus Sets (i.e. Utter-0 in Set 1 vs. Utter-0 in Set 2). Such a comparison would probe whether subjects perceived these as different.

During the Identification Task, the adults provided affective ratings for the two prototype utterances (Utter-0 corresponding to positive, and Utter-5 corresponding to negative) that were significantly different. This supports C_1 , confirming that the two prototype utterances did indeed represent two different categories. Using this along side the results of the Labelling Task, it would suggest that subjects did perceive these two extreme prototype utterances as different. This is an important confirmation as it does indeed show that the stimulus continuums did indeed represent two different affective classes at either end and thus given the results, there was indeed a categorical boundary covered by the Stimulus Set 1 continuum, and that the presence of the categorical boundary was apparently related to the Pitch Contour of the utterances.

Finally, while the main effect due to the subject gender in the Identification Task ANOVA was not found to be statistically significant (though it was close to being statistically significant), it is likely that with a large corpus of subjects this would become a significant main effect.

6.4.2 Methodological Remarks

While the results of the adult experiment indicate that CP is occurring, this experimental set up is not without methodological drawbacks. Gerrits and Schouten

(2004) have argued that CP findings depend upon the type of Discrimination task that has been employed. In the experiment presented here, the AX paradigm has been used, where subjects are presented with two stimuli and have to say whether they feel that they are the “same” or not. While this is a task with a low cognitive load, it tends to have a bias toward subjects providing more “different” ratings as there are not other pairs (presenting identical stimulus pairs) in the trail with which comparison may be made (Schouten et al., 2003; Gerrits and Schouten, 2004). As such, where neighbouring stimuli might be expected to fall within the same categorical region, these have a higher chance of being deemed as “different”, which reduces the inverted “V” shaped differential profile.

Other paradigms commonly used are the ABX and 4IAX comparison tasks. In the ABX task, subjects are presented with three stimuli, two of which are the same, and subjects must identify which of the first two stimuli (A or B) is the same as the last (X). The 4IAX task is a far more cognitively demanding task than either the AX or ABX tasks, where subjects are presented with two pairs per trail (e.g. AA-BA, AB-BB, etc.) and subjects must identify which of the two pairs contains the odd one out (for example, AA or BA). In the case of the ABX task, it is common to find a bias due to presentation order where the B and X stimulus are more likely to be identified as the same than A and X. This is theorised to be linked (in cases which use auditory stimuli) to the loading on auditory memory. The 4IAX task is a method that holds less overall bias, however has high cognitive loading, and requires that subjects listen to a total of four stimuli rather than 2 per trial, and as such may be cumbersome to explain to young children. This study has employed the AX as it was deemed to be the least demanding paradigm to use, which was an appealing factor when considering that experiment was to be conducted with young school children. This acknowledges that a different discrimination paradigm may have led to different results with a different interpretation.

Comment may also be made about the number of utterances that were used in each Stimulus Set (6 utterances). This is a relatively low number which resulted in

larger linear physical transitions between the two prototype stimuli in each set. As such, it may be likely that the stimuli would be perceived as more different during the Discrimination Task due to the larger differentials in the physical properties of the utterances. With respect to the Identification Task, this may also have provoked ratings that were different. However, while this may be the case, the results for Stimulus Set 1 still have provided compelling evidence supporting the presence of CP. That said, the results for Stimulus Set 2 could perhaps have tended more toward CP had there been more utterances in the continuum. A refinement to the experiment would have been to double the number of utterances in each stimulus set. However, there were concerns regarding the level of cognitive loading placed upon the young children, as well as the amount of time taken for each subject to complete the experiment.

Finally, the actual use of the AffectButton as a means for capturing affective ratings in this experiment may be subject to criticism. Given that it has been well established that humans exhibit categorical perception of facial expressions, it may be argued that this experimental setup has an inherent bias that would promote evidence supporting the presence of CP. In reality this is a potent criticism to make, and it is difficult to assess whether such a bias is indeed taking place, as well the magnitude. One would need to use a completely different measurement tool to gauge this, which in itself can lead to more criticisms - for instance whether a different tool does indeed provide a robust representation of an affective interpretation (it is argued here that using facial expressions does this). However, setting this aside, the findings of CP in this study are still relevant to the field of HRI as many of the current state-of-the-art social robots have expressive faces that are used to display facial expressions (e.g. Breazeal (2002); Delaunay et al. (2009, 2010); Kuratate et al. (2011)). In this light, robots that are capable of affective expression through multi-modal displays are more likely to be subject to CP due to the larger number of cues that are available to the human to decode the displays in a more refined manner.

6.4.3 Broader Discussion

The findings of CP (and the lack of) in this chapter have implications regarding both the generation of NLUs as well as the general use of NLUs during social HRI. With respect to the use of NLUs in HRI, the finding of CP suggests that subtle differences in the acoustic features of utterances do not necessarily translate to subtle differences in how these utterances are interpreted with respect to their affective meaning/colouring. Rather, the results show that utterances can be subject to a “magnet effect” whereby they are drawn to more coarse and prototypical affective interpretations (e.g. happy, angry, sad, act). This is a useful insight when it comes to attempting to predict how a given utterance may be interpreted by a subject, a capability that falls toward the modelling of a *Theory Of Mind* in social robots (see Scassellati (2002) for a discussion on this broader subject).

The results have shown that utterances were interpreted both in a coarse, binary manner, and in a manner with a higher degree of subtle sensitivity. This has implications regarding how affect should be represented in a robotic system, both with respect to the modelling of the robot’s internal affective state, but also to the modelling of the affective state of the person/agent that the robot is interacting with. Dimensional representations of affect have a potent collection of appealing characteristics. Firstly, they are able to cater for both coarse changes in affective states as well as more subtle changes, both of which are associated with how people interpret the affective states of they others that they are interacting with (Breazeal, 2002; Cowie and Cornelius, 2003; Schröder, 2003b). Furthermore, dimensional representations also lend themselves to use in the world of machine learning, an area that has a strong foothold in the state-of-the-art HRI.

Currently this experiment has presented utterances within a context-free manner, and found evidence showing CP of NLUs, though this evidence has not been as clear cut as CP found with respect to other types of stimulus (e.g. colour, phonetic sounds and facial gesture). However, real-world HRI is not context-free, rather all HRI contains implicit context, nor is it uni-modal, rather it is multi-modal. As such, a valuable extension to the experiment presented in this chapter

would be to investigate how the use of NLUs within a scenario with a more defined situational context may differ from the use in context-free settings, and whether the perceptual magnet effect may be more prominent in such situations. This is an issue that is addressed in chapter 8.

6.5 Summary

This chapter has presented the results of two experiments designed to investigate whether both adults and children affectively interpret NLUs in a categorical manner. Employing a methodology that followed the common practice found in psychological experiments studying CP, with some modifications to tailor the work toward the focus of HRI.

2 linear continuums of NLUs were created, each containing 6 NLUs, ranging from fast, high pitch utterances to slow low pitch utterances, with the only differences across the two continuums being the pitch contours of the utterances. Both adult and child subjects were asked to perform three tasks: a Labelling Task, a Discrimination Task and an Identification Task. The Labelling Task was performed to assess the validity of the AffectButton as a means of capturing affective ratings during this experiment. The results show that both the adults and children did appear to be able to use the tool in a coherent manner when assigning facial gestures to various affective labels. The Discrimination Task required subjects to listen to neighbouring utterance pairs in each continuum and rate them as either having a similar or different affective meaning. Here for both continuums the results of the adults followed the trends that are commonly associated with the presence of CP, while those of the children did not. Similarly in the Identification Task, where subjects were asked to listen to each utterance individually and use a simplified version of the AffectButton to assign a facial gesture to each utterance, the adults again had results that followed the trends associated with CP. However, the results given by the children did not. It was also found that the differences in the pitch contours between the two continuums did lead to significant differences in the affective ratings during the identification task.

The findings of this chapter have important implications, not only for the use and generation of NLUs in social robots, but also other areas regarding the representation and understanding of affect in social HRI. With regard to the use and generation of NLUs, it is clear that not all utterances that are subtly different will evoke subtle different affective interpretations in subjects: there interpretations have also shown to be rather coarse or binary and subject to a “magnet effect”. This effect pulls the interpretations to one of a few well established affective affective interpretations. With regard to the generation of NLUs, this is use to know as it means that one does not need to focus a great amount of effort toward producing subtly different utterances in order to evoke subtle different affective interpretations. It may be more fruitful to apply efforts toward identifying acoustic profiles of utterances that are highly representative of known and established affective states and simply adding noise in order to introduce variety in the utterances.

Chapter 7

Using Artificial Neural Networks to Automate NLU Production and Affective Charging

Summary of the key points:

- Feed forward Artificial Neural Networks were trained using the data collected from the experiments in chapters 5 and 6 in order to learn a mapping between peoples' affective interpretation of NLUs and the parameters of the NLU generation algorithm.
- Mappings learnt share similar characteristics to the acoustic correlates of emotional expressive in both the human voice and music.
- A human subject evaluation was performed with young children in order to assess the learnt mappings. NLUs were generated using the mappings and subjects asked to rate them using the AffectButton.
- Results show that even when NLUs have similar acoustic characteristics as those found in the human voice and music, subjects still attribute prototypical affective to utterances, but are not coherent in how they do this.

In the experiments presented chapters 5 and 6, the utterances used as stimuli have been hand crafted for the specific experiment in mind, each aimed at attempting to uncover how subjects respond to specific manipulations of the parameters characterising an utterance. While this is suitable for investigating the impact of each acoustic parameter on the affective interpretation, hand crafting utterances generally has very limited practical utility. Furthermore, the method for creating NLUs described in chapter 3 only provides what is essentially a *blueprint* for how utterances are described, characterised and synthesised. It does not provide a specification for how the different parameters characterising an utterance translate to different affective interpretations in the eyes of subjects. This chapter seeks to address this, and if possible, provide a remedy by investigating whether the generation of NLUs and specifying the values of utterance parameters may be *automated* in some manner using feed-forward Artificial Neural Networks (ANNs) to provide a means of generating and affectively charging NLUs, without the need to hand tailor utterances to convey a specific affective meaning.

The chapter begins with an overview of the different methods that have been used to create and affectively charge both gibberish speech and NLUs in the existing literature, as well as providing a brief discussion of how the insights gained in the field of speech synthesis in this regard have, in general, limited application to the field of NLUs. The field of Machine Learning is then introduced in order to provide a more formal problem statement that is to be addressed in this chapter using ANNs. This is followed by a description of how feed-forward ANNs have been used to try and uncover an affective mapping between the AffectButton PAD space, and the parameters of the NLU generation algorithm. A human subject evaluation of this learnt mapping is then presented, followed by a discussion of the networks that have been developed, and the evaluation results, concluding with a summary of the work presented.

7.1 Overview of NLU/gibberish speech Generation.

Within the world speech synthesis, emotional speech synthesis is now an issue that is receiving considerable attention. While there is a great volume of knowledge emerging from this field, overall, the state of this remains in early stages (see Schröder et al. (2010) for a comprehensive overview). Furthermore, while one might intuitively imagine that knowledge gained from the field regarding how to synthesise emotional speech may hold useful insights as NLUs may be affectively charged, in reality, many of these insights cannot be applied to the automated synthesis of emotional NLUs. The reason for this is simple. Speech synthesis is concerned with reproducing human speech, which is of course, more broadly speaking, means natural language. As such, the current state of the art in emotional speech synthesis has focused upon adapting the already well established underlying technologies in speech synthesis (e.g. format synthesis, diaphone concatenation and unit-selection (Schröder, 2001; Schröder et al., 2010)), of which the primarily goal remains the reproduction of language, where *what* is said and *how* it is said are facets that are difficult split apart due to the way that the existing technologies work. To clarify, given that NLUs - specifically in this body of work - are in essence abstract sounds in comparison to human speech, much of the work on emotional speech synthesis is generally incompatible with NLUs. With respect to gibberish speech however, this is not the case as gibberish speech relies heavily on the speech synthesis (TTS) technologies, and as a result, insights and developments in speech synthesis are very relevant and readily applicable, as evidence by the work of Breazeal (2002), Oudeyer (2003) and Yilmazyildiz et al. (2006, 2010).

While the lack of direct compatibility of developments within emotional speech synthesis to NLUs is an issue, there are some high level methods and ideas that have been adopted by authors generating both NLUs and gibberish speech, and as not to neglect this, the rest of this section will outline these.

Breazeal (2002), Oudeyer (2003) and Yilmazyildiz et al. (2010) describe algorithms and methods for generating utterances consisting of strings of gibberish text, which are input to a text-to-speech engine, the TTS engine/synthesiser settings used to *charge* the utterance are based upon static predefined values, or templates, which have been guided by the acoustic correlates reported in the psychology literature (e.g. Banse and Scherer (1996), Fernald (1989) and Burkhardt and Sendlmeier (2000)). Broadly speaking, these are examples of a *template* based approach to affective charging, where a small number of pre-defined, static, synthesiser settings are used to create utterances that convey a limited number of affective states. Jee et al. (2007, 2009, 2010) have used the same general approach with their music inspired NLUs, where they hand crafted a limited number of utterances to convey basic emotional states based upon observations on how emotion is communicated through music (e.g. Juslin and Laukka (2003)). The benefit of this approach is that the insights from related literature can be broadly drawn upon, and so little effort needs to be put into exploring how the settings of their respective synthesisers need to be manipulated to order to convey different affective states. This is particularly the case when TTS synthesisers are employed as much of this has already been established, explored and exploited in the field of speech synthesis. However, the main drawback is that the number of affective states that can be modelled is limited by the coverage of affective states that have been explored in the literature. Furthermore, this approach also tends to result in a small number of different synthesiser settings for each particular affective state.

Yilmazyildiz et al. (2006) and Yilmazyildiz et al. (2011) describe a slightly different method in which recordings of expressive speech portraying a limited number of discrete emotions are recorded from a voice actor and used either as a prosodic templates which are then mapped to recordings of neutral speech, and in the case where the expressive speech recording are of gibberish speech, are simply

used as the final utterances themselves. This tends to follow the *unit selection*¹ approach to speech synthesis. With respect to NLUs, Tuuri et al. (2011) have used a similar method in the development of their expressive sports wrist watch where they took voice actor portrayals, extracted the fundamental frequency profile from these and mapped these to MIDI notes in a music synthesiser, thus retaining the prosodic properties of the voice recordings, while removing the high level frequencies in the acoustic signals, from which the perception of words tends to emerge (Remez et al., 1981; Knoll et al., 2009). These approaches have two main benefits. Firstly, little computational effort is required, as a database of (affectively labeled) utterances is created beforehand, and used utterances only need to be selected from the database and played back during an interaction. Secondly, as voice actors have provided the expressive speech, the acoustic correlates for each emotional state are inherent in the recordings. However, the drawbacks to this method are that a database of voice recordings must be made, which is a cumbersome and time consuming process which results in a large database. This has limits on the practical use with respect to robots with embedded microprocessors and computers which have limited computational resources. Furthermore, the recordings can only be given a discrete label, which limits the number of different emotions that can ultimately be conveyed by these systems. However, given the knowledge that NLUs are subject to categorisation with respect to their affective interpretations, this is a drawback that may only be superficial.

As is evident, there have been more efforts in developing methods for generating and synthesising gibberish speech than there have been for creating NLUs that lend themselves to automated generation. It is likely that the main reason for this is that as gibberish speech tends to rely on TTS technology and the develops surrounding this, which inherently lends itself to more rapid and fruitful development for gibberish speech than NLUs.

¹Unit Selection (Iida et al., 2003) is where large collection of recordings of expressive speech are captured from voice actors and are split up into smaller *sound units*, which in turn are then concatenated to produce different utterances, and had proved to be one of the most fruitful methods in current speech synthesis. This approach, however, is limited by the number of available sound units, as well as the different emotions that have been portrayed by the voice actor(s), as well as the expressive tones of the various recordings. See Schröder et al. (2010) for a comprehensive overview of this technique with respect to emotional speech synthesis.

With respect to NLUs that consist of simple frequency and amplitude modulated sine waves, the main body of literature has been Komatsu (2005); Komatsu and Yamada (2007, 2008); Komatsu et al. (2010); Komatsu and Yamada (2011); Komatsu et al. (2011), however their focus has not been upon attempting to convey a variety of different affective states, rather they have focused upon how single sine waves can be used to manipulate how people perceive the attitude of the agent making the utterances. Furthermore, the utterances that they have used have been very simple, consisting of single sine wave signals with either a rising or falling pitch/frequency modulation, and as a result, there has been no need to develop a system for generating their utterances in an automated manner.

This body of research uses NLUs that are in essence the same *type* of acoustic signal (i.e. single sine waves that have a frequency and amplitude modulation) as used by Komatsu *et al.*, but extends beyond a single wave, to multiple, concatenated waves also. As such there are many more parameters required to characterise an utterance, and so in order to produce and characterise utterances in a standardised manner, a custom has been developed to do this (chapter 3). Due to the overall increased complexity, the method is capable of producing a rich variety of different NLUs. However, the specification of the utterance parameters has thus far only been achieved via human *hand craft*, which in the long term, is impractical. What is currently lacking is the ability to automatically specify these parameter values given that there is a desired affective interpretation - the mapping between affective interpretation and utterance parameters is unknown. It may be argued that the insights from the related psychological and musicology literature could be directly used, given the language inspired design of the NLU algorithm, however this is not clear and has not been explicitly tested for. Nor it is to be assumed that such a mapping is direct. Taking the insights regarding the acoustic correlates of the human voice and music and directly applying these to NLUs may not result in people having the desired interpretation. Chapters 5 and 6 have presented experiments where subjects were asked to affectively rate utterances, where the acoustic features of the utterances were systematically ma-

nipulated in such a way as to explore the utterance parameter space and probe the landscape of subjects' affective interpretations. The most logical next step is to investigate whether the data collected during these experiments may have a hidden general mapping between the affective interpretations of utterances and parameters used to generate the utterances. This is done using machine learning in this case.

7.2 Machine Learning

Machine Learning (ML) is a field within the world of computer science, and more specifically, *artificial intelligence* (AI), that concerns itself with developing techniques and algorithms that allow computers to *learn* from data by themselves. Mitchell (1997) provides a more formal definition for the notion of a computer that learns:

“A computer program is said to learn from experiment E with respect to some class of tasks T and performance measure P , if its performance at tasks in T , as measured by P , improves with experience E .” - Mitchell (1997).

For example, if a computer is learning to filter “SPAM” emails, then this process of filtering is the task, T . The performance metric, P , would be the number emails correctly labelled as “SPAM”, and finally, the computers' prior experience, E , would be having access to a collection of example emails which have been user labeled as either SPAM or not SPAM. This filtering task is an example of a computer that has learnt to perform a *classification* task through a *supervised learning* process.

There are many different ML techniques, each which is tailored toward specific problems, and each which have slightly different variants. These can range from very simple and classic techniques such as logistic or linear regression, to more tried and tested approaches such as Artificial Neural Networks, to the state-of-the-art and very exotic approaches such as AdaBoost and Echo-State Networks, to name a few. With this wide range of techniques also comes a wide range of problems that can be solved also, such as regression, classification, clustering,

feature learning, prediction and association rules, to name a few.

This section serves to provide a brief overview of the general approaches that are most commonly used in ML, and then to highlight the *supervised learning* approach and outline the two main tasks that this approach is used to address. Doing this allows us to put into perspective the ML problem that this chapter seeks to address given the training data that was collected, and sketch out a formal problem definition which can help in the selection of an appropriate learning approach and technique.

7.2.1 Approaches to Machine Learning

ML techniques tend to come in one of four different general flavours: *supervised learning*, *unsupervised learning*, *reinforcement learning* and *evolutionary learning*. In the case of the SPAM filter example, the approach used was a supervised learning approach in that each example email in the training data was paired with the correct response. Having the correct output data for each input data point is analogous to having a teacher sitting next to you and supervising your learning, thus ensuring that you are indeed learning the correct things.

In the case of unsupervised learning, example data is not presented with the correct responses. Instead, the algorithms attempt to identify similarities between different inputs such that they may be grouped or clustered together in some way, and ultimately be classified with some label (which the computer then decides on). The algorithm has to make sense of the data by itself, and thus is free to form its own internal, meaningful representations of the data that is present to it.

Reinforcement learning lays somewhere in-between supervised and unsupervised learning. The algorithm performs a task, and is then told whether the outcome is correct or incorrect, and in the case of the latter, it is not told how to amend the error. In the case of the former, while it is told that it has succeeded, it is not told which aspects of what it did were correct, and which were incorrect. The algorithm has to explore and try the various different possibilities with varying success until it works out how to solve the task at hand.

Finally, evolutionary learning is an approach that takes inspiration from the world of biology, where biological evolution is seen as a learning process also. Here, computers run algorithms that model analogies to the adaptations that biological organisms undergo in order to improve their survival rates and increase their chances of having offspring in the environment. This approach models different solutions as members of a population and seeks to “cross-breed” the most successful solutions (as measured against a given fitness metric) in order to produce new solutions. The hope with this approach is that an optimal, robust solution will emerge when this process is repeated many times on end.

The machine learning approach adopted in this thesis falls into the supervised learning category. The data that was collected in the experiments described in the two preceding chapters (chapters 5 and 6) contains both the parameter values of the different NLUs, as well as the affective interpretations as captured via the AffectButton. This training data contains data examples that are paired with “correct” responses (in the eyes of the subjects). As such, this form of *training* data makes a supervised learning approach very suitable.

7.2.2 Supervised Learning

Given the input data/correct response format of the training data that is presented to a supervised machine learning algorithm, supervised learning algorithms tend to tackle one of two types of problem: *classification* and *regression*. Both of which we shall in more detail in a moment.

The training data presented to the algorithms typically takes the following form: $\mathbf{T} = \{(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), (\mathbf{x}_3, \mathbf{y}_3), \dots, (\mathbf{x}_n, \mathbf{y}_n)\}$, where both \mathbf{x}_i and \mathbf{y}_i are vectors that can contain lots of pieces of data that can take either binary or continuous values. For example, in the case of this thesis, $\mathbf{x}_i = [x_1, x_2, \dots, x_m]$, where x_j is a given NLU parameter, and $\mathbf{y}_i = [p, a, d]^T$, where p , a and d are the respective Pleasure, Arousal and Dominance values for the given utterance with parameters \mathbf{x}_i .

It is the values in both \mathbf{x} and \mathbf{y} that play a large role in determining the

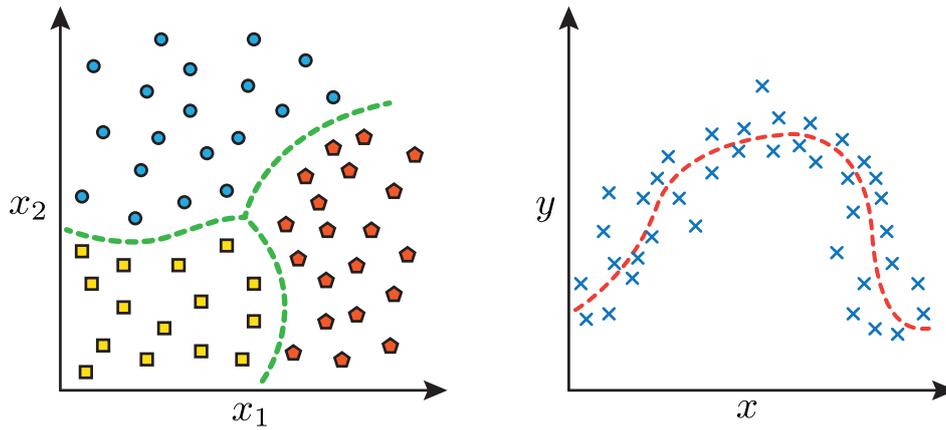
nature of the problem that needs to be solved by a given ML algorithm. When the output (\mathbf{y}) values take on discrete values (i.e. one and only one value within a given set), then this is a *classification* problem. Conversely, when the values of \mathbf{y} fall along a continuous range, predicting the values of \mathbf{y} given input data \mathbf{x} becomes a problem of *regression*.

Supervised learning, like the other four ML approaches, also holds the property of generalisation. As the training data that it has learnt from does not represent the entire input space, it is likely that examples that are not in the training set are presented to the algorithm after learning has finished. In these situations the trained algorithms are able to generalise and provide realistic output values, through either interpolation between data points, or extrapolation beyond the range of data points presented during training.

7.2.2.1 Classification

A classification problem consists of taking the input vectors of data points \mathbf{x}_i and predicting which of N classes (\mathbf{y}_i) they each belong to based upon the *features* of a given example (i.e. the contents of \mathbf{x}_i), where the classifier is trained using the exemplar data contained within the training set \mathbf{T} . Here also, the values in \mathbf{y} are discrete values which represent categories where \mathbf{x} is a member of only one class. This is not strictly a universal rule, as there are classifiers that perform *fuzzy logic* (Marsland, 2009), but these are far beyond the scope of this thesis.

The task of the learning algorithm in this case is to establish either one or multiple decision boundaries that are used to separate the input feature space into the different discrete regions, and then assign a category label to each region (see figure 7.1a). Depending of the type of values that are contained within \mathbf{x}_i these decision boundaries can either be linear or non-linear which allows for differing degrees of classifier complexity. Examples applications include Speech and Handwriting recognition as well we biological recognition such as tumour diagnosis.



(a) Example of a Classification problem where class membership is predicted based upon the features x_1 and x_2 and non-linear decision boundaries have been learnt.

(b) Example of a Regression problem where a mathematical function has been approximated by an algorithm allowing the output value y to be predicted by the input variable x .

Figure 7.1: Illustrative examples of a Classification problem and a Regression problem which may both be solved through Supervised Learning.

7.2.2.2 Regression

Regression problems are different in that rather than attempting to assign a discrete class membership to a given point in the input feature space, they attempt to predict and then assign one or more continuous values as outputs (\mathbf{y}) to the input point (\mathbf{x}). Essentially, regression algorithms attempt to fit a mathematical function describing a curve/surface so that it passes as close as possible to all the data points in \mathbf{T} . As such, regression is also known as a problem of *function approximation* or *interpolation*, working out the values between the data points that we know (see figure 7.1b).

The problem faced by the learning algorithm is to determine the *parameters* of the mathematical function that is being fit, as these depend on \mathbf{T} . For example, if you consider the equation of a straight line: $y = \theta_0 + \theta_1 x$, θ_0 and θ_1 are the parameters whose values need to be learnt by the algorithm, for example, a linear regression algorithm. Here also the linear shape of the function is determined by the variable x . If the input variable were x^2 , then the function shape would be non-linear and polynomial. Examples applications include House Price estimation, Robot Arm Torque Controllers and Stock Market Analysis.

7.2.3 Formal ML Problem Statement

The previous two chapters described experiments in which subjects affectively rated NLUs with known parameter values using the AffectButton measuring tool. As such, the data collected represents training data that consists of input/output value pairs. The goal of this chapter is to uncover a generalised *mapping* between these NLU parameters/affective ratings whereby it is possible to specify a desired affective rating, and through the use of an ML algorithm, obtain NLU parameters that are representative of this affective rating. More formally, a mathematical function needs to be approximated using supervised learning to solve a regression problem such that NLU parameter values may be predicted based upon a given PAD coordinate within the AffectButton affect space:

$$\mathbf{y} = h_{\theta}(\mathbf{x}) \quad (7.1)$$

where $h_{\theta}(\mathbf{x})$ is the approximated function (also known as the *hypothesis* function), \mathbf{x} is the desired input affective PAD coordinates that takes the form $[p, a, d]^T$ and \mathbf{y} is a vector consisting of the predicted NLU parameter values, taking the form $[x_1, x_2, \dots, x_m]$.

The function h_{θ} has a number of parameters, Θ , which characterise the mapping. The role of ML here is to identify the optimal values of Θ such that the *error* between $h_{\theta}(\mathbf{x})$ and \mathbf{y} is minimised for all data points in the training data set, $\mathbf{T} = \{(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), (\mathbf{x}_3, \mathbf{y}_3), \dots, (\mathbf{x}_n, \mathbf{y}_m)\}$:

$$J(\Theta) = \frac{1}{2m} \sum_{i=1}^m (h(\mathbf{x}^{(i)}) - \mathbf{y}^{(i)})^2 \quad (7.2)$$

where $J(\Theta)$ (also known as the *cost function*) is the total mean squared error across all the m training data points in \mathbf{T} , Θ are the parameters of h_{θ} , $h_{\theta}(\mathbf{x}^{(i)})$ is the predicted output of a given training data input point, and $\mathbf{y}^{(i)}$ is the desired output value.

7.3 ANN Design and Implementation

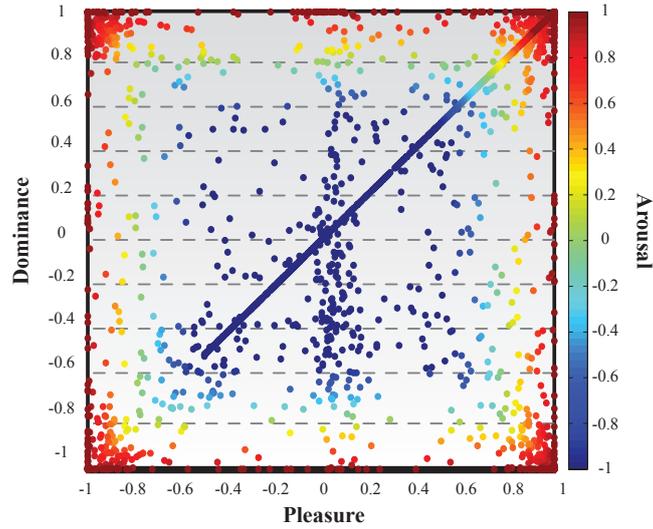
This section outlines the implementation of feed-forward ANNs and how they have been used to try and learn and generalise an functional mapping between the AffectButton affect space (as outlined about) , within which subject affective ratings are located, and the parameters of the NLU. First, the training data is described. Then the implementation of the ANNs is described, followed by a presentation of the the mappings that have been obtained.

7.3.1 Training Data

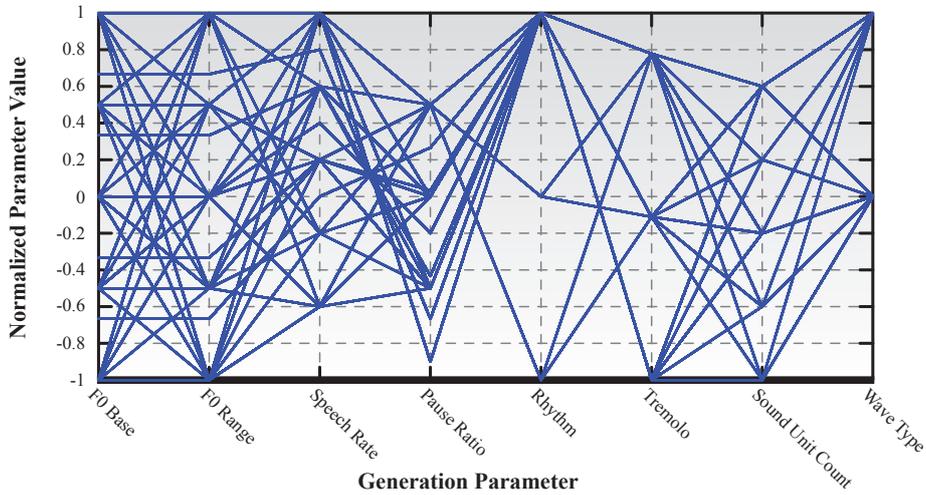
The data collected from the experiments described in chapters 5 and 6 were combined to form the training data set. Overall this set consisted of 2263 data points, with the majority of the data coming from experiments with children, rather than adults (see table 7.1). Figure 7.2 shows plots of both the affective ratings in the AffectButton PAD affect space, and the generation parameter values of the utterances to which these ratings correspond.

Input data consisted of affective ratings of utterances captured using the AffectButton, and as such consisted of vector holding the Pleasure, Arousal and Dominance values. As the AffectButton outputs values along each dimension in the range $[-1 1]$, no pre-processing was performed on the data before being into the ANNs. Figure 7.2a shows a scatter plot of all the input data. Note that there is a prominent diagonal line of data points in the figure, which is the data collected from the Categorical Perception experiments (chapter 6) and is the result of the constraining of the AffectButton during the Identification Task during these experiments. It is also notable, from a visual inspection, the distribution of data points throughout the affect space is uneven, with the majority of data points being clustered into the corners of the figure.

Output data consisted of the specific NLU generation parameters that were varied over the various experiments. These were the F0 Base, F0 Range, Tremolo, Speech Rate, Pause Ratio, Rhythm, Sound Unit Count, and Wave Type and the Pitch Contour. All values, except for the Pitch Contour, were normalised such



(a) Scatter plot of the input training data.



(b) Parallel plot of the output training data.

Figure 7.2: Plots of the Input and Output training data.

that, like the input data, their values fell into the range $[-1 1]$. The Pitch Contour was encoded as an integer that whose length was equal to the sound unit count, with the 5 contour shapes were each encoded as a single integer (see table 7.2) and being concatenated together.

This training data set presents a challenge for machine learning techniques in general. This is due to the high dimensionality of both the input and output spaces, and that the number of outputs is larger than the number of inputs. This means that the problem involves dimensional *expansion* rather than *reduction*. Furthermore, the utterance parameter space has been very sparsely sampled during the experiments (see chapter 5) providing few data points, with large euclidean

Table 7.1: Break down of training data set with respect to the original experiments within which the data was collected.

Experiment	N° of Data Samples	Percent (%)
Training Data Collection (Children)	1373	60.68
Categorical Perception (Adults)	430	19.0
Categorical Perception (Children)	460	20.32
Total	2263	100

Table 7.2: Schema for encoding the Contour shape of a given Sound Unit in the training data.

Contour Shape	Encoding
Flat	1
Rising	2
Falling	3
Rising-Falling	4
Falling-Rising	5

distances between them, which may well pose a problem from generalisation between training points. While this is the case, it can be seen from figure 7.2b that the output parameter space has been sampled rather evenly. There are however some regions of the output space that have been sampled more densely than others. For example, the Rhythm parameter has in most cases held the value of 1. This was done due to the inherent randomness that is introduced via the NLU generation algorithm when the value is lowered (see chapter 3, section 3.1). Similarly, the wave type parameter has only two values (1 for saw wave and 0 for sine wave) with most values being 0.

While the training data covers the nine different NLU parameters that were manipulated throughout the experiments, the Pitch Contour parameter as not included in the final training data set, and as a result, the ANNs were not employed to learn a mapping for this particular parameter. The reason for this is that as the experiments in chapters 5 and 6 have found inconsistent results regarding influence of the Pitch Contour upon affective ratings. Given the apparent elusiveness nature of this particular parameter, clearly more carefully designed research and experiments are required to fully understand the exact role of the Pitch Contours. As a result, the Pitch Contours of the utterances in the subject evaluations (section 7.4) and were again pre-specified in order to not confound the

results with respect to the other, mapped, utterance parameters.

7.3.2 Using Feed Forward ANNs

A *feed-forward* Artificial Neural Network (ANN) architecture was used to process the training data. ANNs are a very well documented method of machine learning that has been applied successfully in a variety of different fields (see Marsland (2009); Mitchell (1997); Bishop (2006) for extensive, detailed overviews), and as such, this chapter will not provide a detailed description of this machine learning technique and the underlying mechanisms, principles and algorithms regarding their operation and functionality.

However, with that said, the rationale for employing this particular type of machine learning tool, as opposed to the many others that also exists is as follows. Firstly, the application of ANNs has been extensively explored in many areas, and as such, the characteristics of their operation too is well documented, and thus they serve as an obvious and practical first stepping stone with respect to the goal of using machine learning techniques to automate affective NLU generation: ANNs are essentially a *classic* tried and tested machine learning technique. Secondly, ANNs provide both linear and non-linear *function approximation* functionality, which is in essence the problem that is being presented here: within the data there likely lays functional relationships between the three dimensions of the affect space and each dimension of the output space (i.e. each NLU parameter). Finally, with the exception of the Pitch Contour (which we have already outlined will be not addressed in this chapter), each NLU parameter holds a static value for each utterance, and thus only a static function mapping is required, so there is no need to employ extensions such as *recurrent* ANNs, which may well overcomplicate the problem at hand.

7.3.2.1 Network Topologies

Initially, two general network topologies were explored: a configuration based upon a single multi-layer perceptron with all inputs and outputs being handled

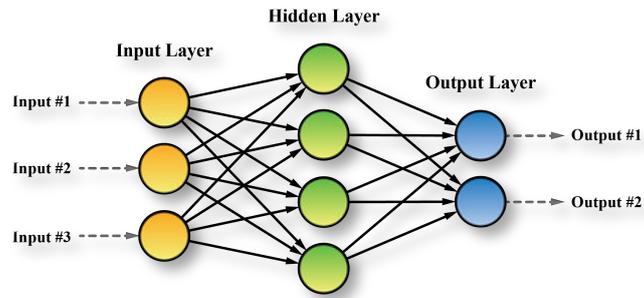
via a Single Multi-Layer Perceptron (S-MLP) network (figure 7.3a), and Multiple Multi-Layer Perception (M-MLP) network (figure 7.3b) where each ANN handled the same input values, but outputting a single output parameter (thus, this involves creating 8 individual networks, one of each output). This was done using the MatLab² neural network toolbox as it is a useful prototyping tool that affords extensive control and specification for the parameters that control how an ANN behaves (e.g. the transfer functions, learning rules, number of hidden nodes, learning rates, etc.).

The rationale for exploring both the S-MLP and M-MLP network configurations is related to the quality of the training data with respect to the mappings that may be learnt. In the case of an S-MLP topology, if the training data does not yield enough quality for a mapping to be learnt, or if a clear mapping does not exist, the quality of the mapping for the other parameters will likely be affected. In the case of the M-MLP topology, each output parameter is the sole focus of neurons the network and is unaffected by the quality of mapping for the other parameters. One can view this as an approach where the mapping for each individual is more robust, but comes at the expense of potentially overlooking potential relationships and interactions between the output parameters.

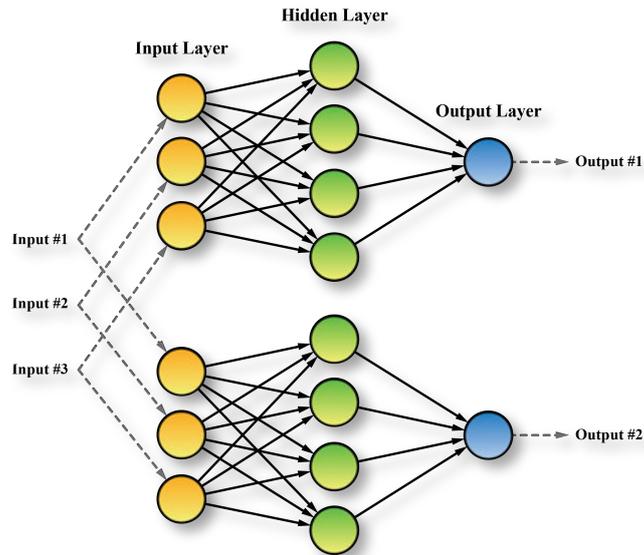
These two main types of network topology both had three layers: an input layer, a hidden layer (with a sigmoid transfer function) and an output layer (with a linear transfer function), see figure 7.4. The reason of this is that a network that has a hidden layer with a non-linear sigmoid transfer function and an output layer with a linear transfer function is capable of mapping any non-linear mathematical function, thus expanding beyond this serves little practical purpose.

The two network topologies were subject to a brief exploration (using the training data) with regard to the number of neurons that should be used to populate the hidden layer, and to get a feel for which of the two topologies should be adopted and used to learn the affective mapping, and in turn used for a subject evaluation. This process is not explained in this chapter as it diverts the focus of attention away for the main goal: to train ANNs to learn an affective mapping

²MatLab version 7.13 was used with version 7.02 of the Neural Network toolbox.



(a) Single MLP Network topology.



(b) Multiple MLP Network topology.

Figure 7.3: Multi Layer Perceptron topologies explored for mapping affective input values with output NLU algorithm parameters.

and then *evaluate* this mapping with human subjects. The question is not what specification of network is required, but rather do ANNs hold any potential for proving an affective mapping?

It is suffice to say that the results of this exploration revealed that the S-MLP topology (where all the inputs and outputs are handled by one network) held little promise for learning a mapping as the outputs or each parameter were found to be essentially *random*. The M-MLP topology (where there is a network for each individual output parameter), on the other hand, did show considerable promise for learning a mapping for each output parameter, showing more clean, smoother and clear non-linear mappings in general. Furthermore, it was found that networks with seven hidden neurons produced mappings that were deemed to complex and dynamic enough to represent such an affective mapping, rather

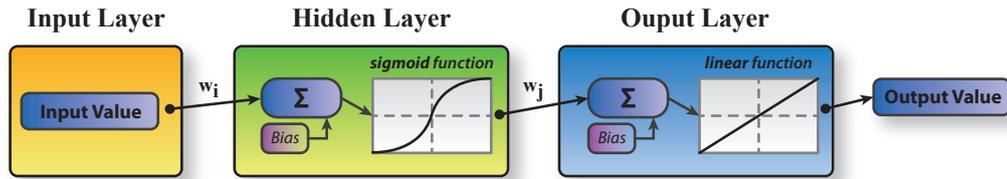


Figure 7.4: Example of a hidden layer with a sigmoid transfer function, and an output layer with a linear transfer function.

than producing mappings that were considered as too general and linear (which was the case for a small number of hidden neurons, e.g. 1 to 5), or mappings that were deemed to show characteristics of over-fitting to the data (which was the case with a larger number of hidden neurons, e.g. > 9).

7.3.3 Network Training

Given the insights outlined above, the M-MLP network topology, with seven hidden neurons, was employed to produce the mapping for each of the remaining eight utterance parameters: the Base Frequency, Frequency Range, Speech Rate, Pause Ratio, Rhythm, Tremolo, Wave Type and Sound Unit Count.

As all the ANNs did not always produce the same exact function mapping with the same training data, 50 networks were trained for each output parameter, and thus 400 networks were trained in total. For each PAD input value, the average output value of the 50 networks, for each of the eight output parameters, was taken as the final mapping values, thus producing a vector of 8 values. For each of the 400 networks the training data set was randomly split up into a training set (70% = 1540 data points), validation set (15% = 330 data points) and a test set (15% = 330 data points). These data sets (and their respective purposes and roles) are detailed below:

The Training Data set is used throughout the training process and is used to compute the error (cost function) of the current network configuration (i.e. the weights and biases of the neurons), which in turn is used to update these values and drive the network learning.

The Test Data set is used at the very end of the training process. This data is

kept aside and only shown to the network once training has been completed and allows for an assessment of how well the network is able to generalise to new data that it has not seen before.

The Validation Data set has a slightly more complicated role. It is used to check how well the network is learning *during* training after each *epoch* (training cycle). It is similar to the test data set in that it is able to assess the network's ability to generalise, however, it has a more important role also. The validation set is used to check whether the network is *over-fitting* to the training data. Normal training would result in the error for both the training and validation data sets to decrease as the number of epochs increases. However, when a network begins to over-fit to data, the error in the validation set begins to increase again while the error for the training data continues to decrease. The two error values diverge. It is at this stage that the training is then stopped as the network is no longer learning about the general trends in the data, but rather is beginning to learn about the noise that is in the training data set itself (Marsland, 2009). This is known as *over-fitting*. When the training and validation errors begin to diverge, and the validation error begins to increase, training is not stopped immediately. Rather, training is halted when the validation error has continued to increase over a given number of epochs (known as the number of *validation data failures*). Setting this to a low value halts training at the first sign of over-fitting, when it may be pre-mature. Increasing the value allows the training to continue while monitoring and ultimately better assesses whether over-fitting is indeed taking place.

Training was performed using the *Levenberg-Marquardt Backpropagation algorithm* (Hagan and Menhaj, 1994) with the *learning rate* held constant at 0.01. The number of validation data failures was set to 15, and the *target gradient* value was set to 0.001.

7.3.3.1 Training Results

There are a few ways in which a regression model may be evaluated. The most common is to observe how the error/cost function $J(\Theta)$ varies as model learns (i.e. across training epochs). As the model becomes more tuned to the data that it is learning from, the overall error decreases as the number of training cycles increases. When this occurs, it is a good indication that the model has indeed undergone an optimisation process and as a result is becoming a more accurate approximating to the underlying function within the training data. This decrease in the error is not always a linear process. Rather, depending on the learning algorithm that is used, as well as the error metric (in this case, the *mean squared error* (MSE)), the decrease in error is *non-linear*, where the overall MSE is quickly reduced in the early stages of the training, after which it begins to flatten out as the model becomes more refined.

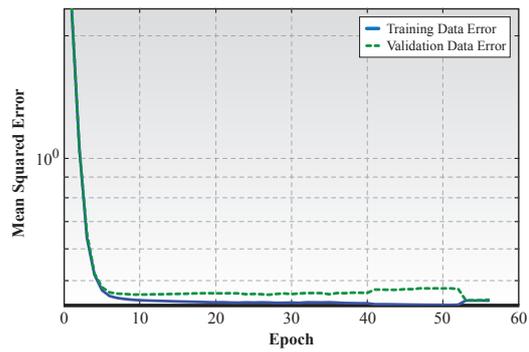
As outline above, the training process for the ANNs made use of three data sets. The training data, validation data, and then the test data. The training and validation data sets were used throughout the training process, and as such, their values after each training cycle may be monitored and plotted as a function of epochs. Figure 7.5 shows plots of the mean MSE in both the training and validation data sets for each of the eight NLU parameters. As there were 50 ANNs trained for each NLU parameter, the plots show the average MSE values across all 50 networks, for each epoch.

Each of the plots shows that indeed each of the 50 networks for each of the NLU parameters indeed did have a decrease in the MSE through the training process. The plots also reveal that the error for both the training and validation data sets had similar values throughout the training process, suggesting that the ANNs were indeed able to generalise also.

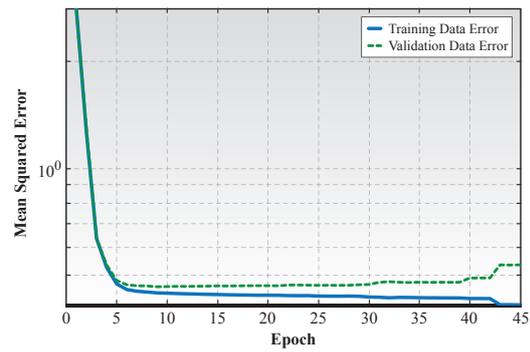
In addition to the error plots, table 7.3 shows the overall values of the training, test, and validation data for all 50 networks for each NLU parameter once training had finished, as well as the duration of training as measured in epochs, and the gradient values of error function. From this table it can be seen that the number

of epochs required to train each ANN, across the different NLU parameters varied notably. For example, the mean number of epochs for the Sound Unit Count and the Tremolo ANNs tended to be around double that for the Base Frequency and Frequency Range. This may suggest that the respective function mappings for each parameter varied considerably in complexity.

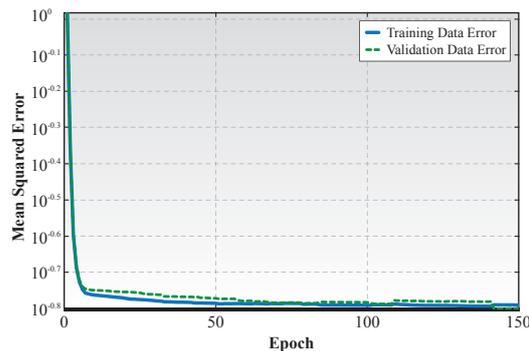
The table also shows that the performance with respect to the error in the three data sets was very comparable once training had completed, with this being true for each of the either NLU parameters. This again suggests that the networks have been able to approximate the underlying functions while retaining their ability to generalise (this, overfitting does not appear to have been a large problem).



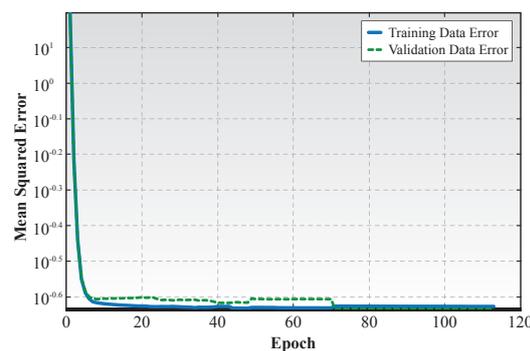
(a) Base Frequency



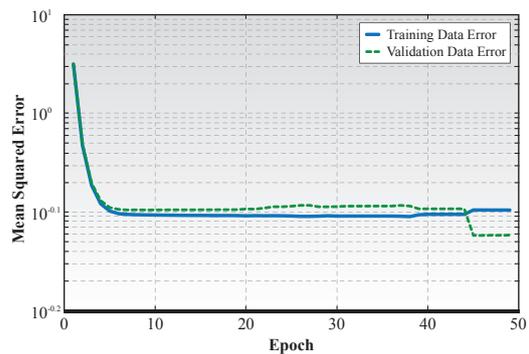
(b) Frequency Range



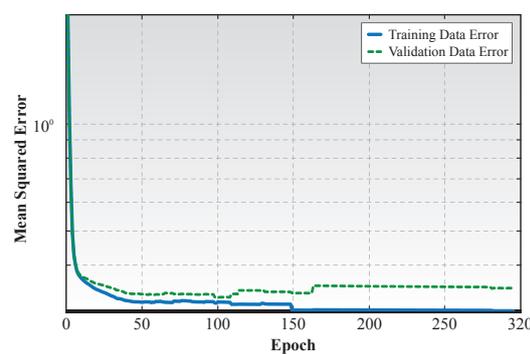
(c) Pause Ratio



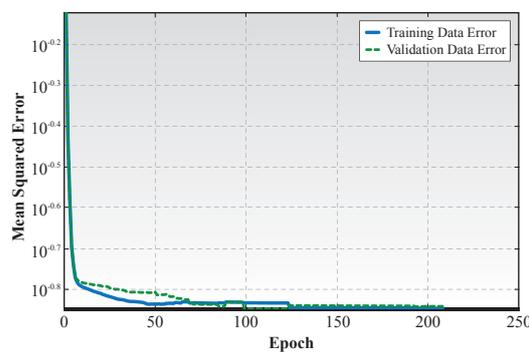
(d) Speech Rate



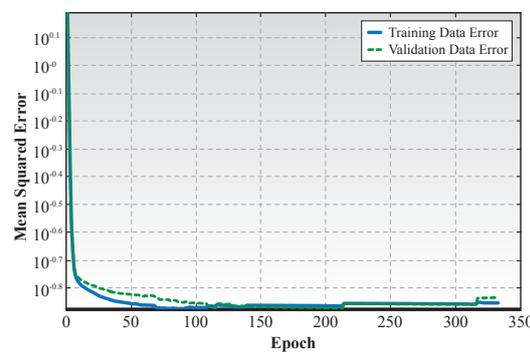
(e) Rhythm



(f) Tremolo



(g) Wave Type



(h) Sound Unit Count

Figure 7.5: Plots of the mean errors (based upon 50 ANNs) in the Training data and Validation data sets over the ANN training cycles (epochs), for each NLU parameter.

Table 7.3: Performance values for the eight trained ANNs. The table shows the number of Epochs performed during training, and the performance on the Training, Validation and Test data sets after training was completed, as well as the final gradient values.

Network	Values	Network Performance At the End of Training				
		Epochs	Train Performance	Validation Performance	Test Performance	Gradient Values
Base Frequency	<i>Mean</i>	26.44	0.4417	0.4689	0.4687	0.0135
	<i>Std</i>	9.008	0.007	0.0228	0.0210	0.0161
	<i>Min</i>	18	0.4233	0.4146	0.4298	5.76×10^{-5}
	<i>Max</i>	56	0.456	0.5255	0.5182	0.0564
Frequency Range	<i>Mean</i>	24.640	0.4382	0.4691	0.4669	0.0169
	<i>Std</i>	6.5489	0.0083	0.0254	0.0262	0.0235
	<i>Min</i>	18	0.4129	0.4124	0.4064	1.15×10^{-4}
	<i>Max</i>	45	0.4558	0.5353	0.5173	0.1280
Pause Ratio	<i>Mean</i>	48.540	0.1654	0.1734	0.1742	0.0075
	<i>Std</i>	30.125	0.0055	0.0102	0.0091	0.0125
	<i>Min</i>	18	0.1572	0.1510	0.1535	4.81×10^{-5}
	<i>Max</i>	150	0.1768	0.1950	0.1955	0.0664
Rhythm	<i>Mean</i>	25.620	0.0918	0.1064	0.1333	0.0065
	<i>Std</i>	8.0277	0.0077	0.0273	0.2354	0.0096
	<i>Min</i>	18	0.0721	0.0395	0.0448	1.02×10^{-4}
	<i>Max</i>	49	0.1094	0.1827	1.7519	0.0480
Speech Rate	<i>Mean</i>	27.600	0.2357	0.2503	0.2481	0.0071
	<i>Std</i>	15.4906	0.0044	0.0126	0.0152	0.0110
	<i>Min</i>	17	0.2253	0.2179	0.2175	7.35×10^{-5}
	<i>Max</i>	133	0.2453	0.2848	0.2780	0.0508
Tremolo	<i>Mean</i>	71	0.3176	0.3362	0.3441	0.0094
	<i>Std</i>	46.9881	0.0247	0.0284	0.0380	0.0152
	<i>Min</i>	25	0.2829	0.2757	0.2736	2.30×10^{-5}
	<i>Max</i>	295	0.3648	0.4292	0.4512	0.0784
Wave Type	<i>Mean</i>	48.200	0.1489	0.1581	0.1589	0.0065
	<i>Std</i>	31.6937	0.0068	0.0117	0.0117	0.0096
	<i>Min</i>	17	0.1383	0.1334	0.1337	2.05×10^{-5}
	<i>Max</i>	208	0.1645	0.1848	0.1953	0.0446
Sound Unit Count	<i>Mean</i>	84.1400	0.1369	0.1482	0.1539	0.0066
	<i>Std</i>	59.2043	0.0100	0.0156	0.0175	0.0152
	<i>Min</i>	28	0.1256	0.1046	0.1222	2.58×10^{-5}
	<i>Max</i>	333	0.1619	0.1725	0.1967	0.0896

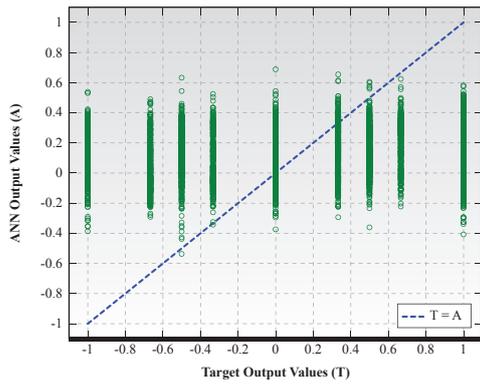
Table 7.4: Correlation Coefficients (ρ) between the target output values and the ANN output values (based on 50 ANNs), for each NLU parameter.

Parameter	Correlation (ρ)
Base Frequency	0.2114
Frequency Range	0.1992
Pause Ratio	0.3248
Speech Rate	0.1191
Rhythm	0.2303
Tremolo	0.5322
Wave Type	0.3841
Sound Unit Count	0.5596

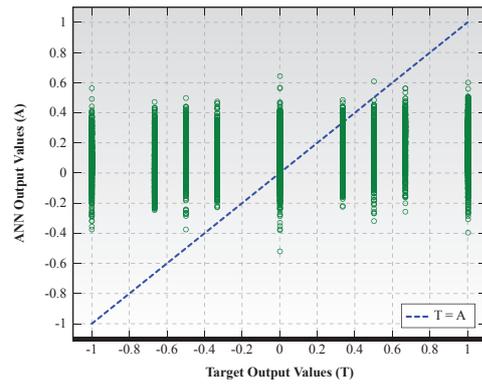
7.3.3.2 Function Approximation Accuracy

While monitoring the error in a networks performance can provide insights as to whether the network has been learning to approximate a function, doing this provides limited insights as to the final accuracy of the mapping that is being learnt. A more fruitful approach to this end is the compare the expected (or *target*) output values for a given input value against the value output by the network. This is shown in figure 7.6 for each of the NLU parameters. A perfect approximation would result in the ANN output values being identical to the target values, and thus all the data points in the plot would fall along the diagonal blue line. The plots however show that this is not the case for each of the ANNs trained for each of the NLU parameters. Rather, the plots show that the approximations made by the networks overall are far from perfect and hold considerable inaccuracy, as well as not covering the full possible range of output values.

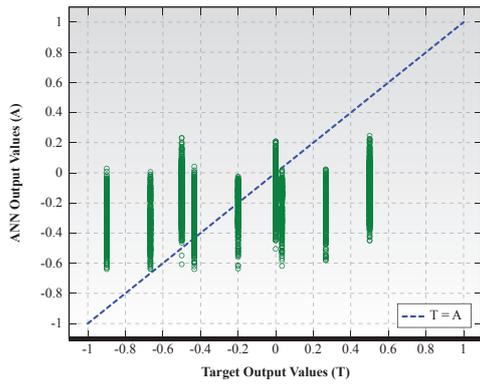
This is also evidenced further by the low correlation coefficients between the ANN output values and the target values (see table 7.4). However, while this is the case, this method of assessing the final network does not account for the complexity of the function that needs to be approximated, which in this case is known to be complex due to conflicting affective ratings for NLUs with similar properties. As such, it is perhaps not overly surprising that the approximation accuracy is by no means perfect.



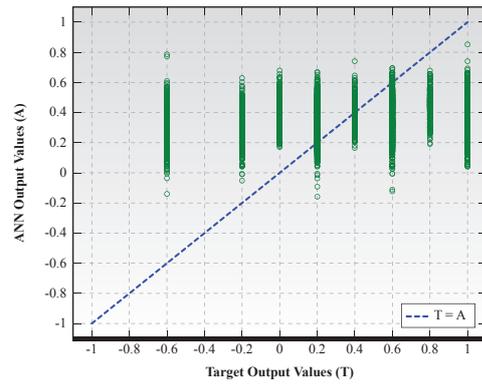
(a) Base Frequency



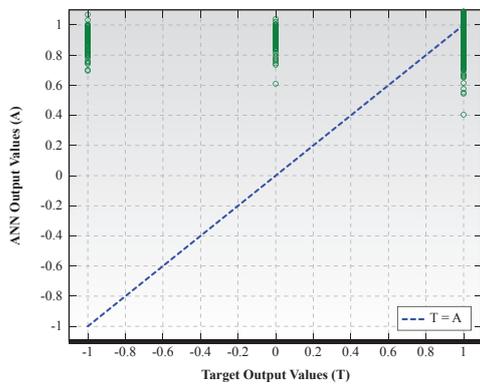
(b) Frequency Range



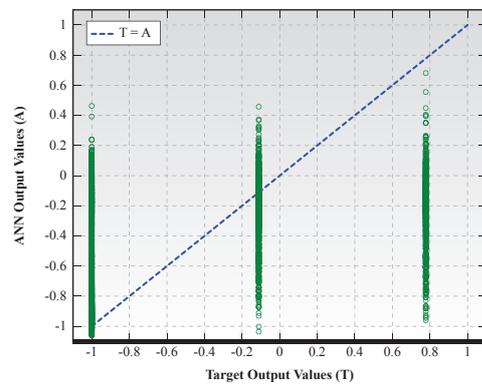
(c) Pause Ratio



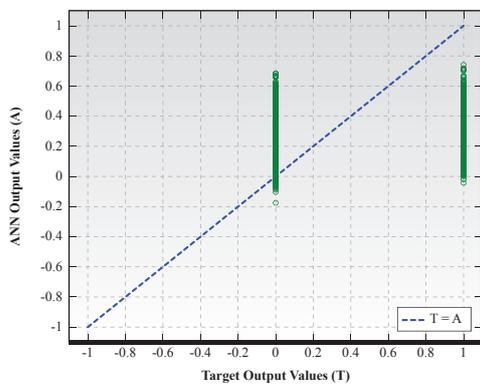
(d) Speech Rate



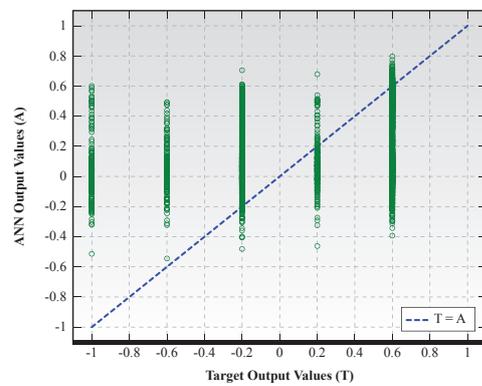
(e) Rhythm



(f) Tremolo



(g) Wave Type



(h) Sound Unit Count

Figure 7.6: Plots of the ANN output values (of 50 ANNs) vs the target output values for each of the NLU parameters. See table 7.4 for the Correlation Coefficients.

Table 7.5: Mean, Standard Deviations, Minimum and Maximum values for the ANN Mappings for each of the utterance parameters.

Parameter	Mapping Values			
	Mean	Std	Min	Max
Base Frequency	0.0982	0.0922	-0.1156	0.4174
Frequency Range	-0.1156	0.0874	-0.1005	0.3555
Speech Rate	0.3074	0.0807	0.1564	0.5362
Pause Ratio	-0.0847	-0.0847	-0.3199	0.0435
Rhythm	0.9246	0.0334	0.7993	0.9806
Tremolo	-0.2527	0.1304	-0.9694	0.0081
Sound Unit Count	0.0593	0.0999	-0.1472	0.5949
Wave Type	0.3353	0.0666	0.0454	0.4952

Table 7.6: Partial Linear Correlation Coefficients (ρ) between the affect space dimensions and the generation parameter values output from the ANN.

Parameter	Affect Space Dimension		
	Pleasure	Arousal	Dominance
Base Frequency	0.164	0.381	0.592
Frequency Range	0.482	0.400	0.715
Speech Rate	0.609	-0.325	0.650
Pause Ratio	-0.486	0.095	-0.688
Rhythm	0.226	0.124	0.155
Tremolo	0.104	0.186	-0.171
Sound Unit Count	0.088	-0.278	0.319
Wave Type	-0.172	0.507	-0.201

7.3.4 Affective Mapping Analysis

Figure 7.7 displays surface plots of the averaged mapping values between the affect space (along the Pleasure and Dominance dimensions) and the individual parameter spaces. Upon initial visual inspection these plots show correlations between a variety of the parameters. It is also worthy to draw attention to the value range of the parameter mappings, which tend to be small rather than covering the full range of $[-1, 1]$ which the training data set did do. Table 7.5 shows the mean, standard deviation, minimum and maximum values of the mappings for each of the NLU parameters.

Table 7.6 shows the Liner Partial Correlation Coefficients (ρ) between each dimension in the affect space (accounting for the other two dimensions) and each acoustic parameter. These correlations reveal that there are certain notable trends within the mappings. For example, the correlation coefficients indicate that the

Speech Rate parameter is positively correlated with both the Pleasure ($\rho = 0.609$) and Dominance ($\rho = 0.650$) dimensions. Also, the Base Frequency and Frequency Range parameters are also found to be positively correlated with all the affect space dimensions, with a varying degree. What is interesting to see is that some of the of these values appear to concur with findings that have been reported in the related psychological and musicology literature. For example, Banse and Scherer (1996) confirmed predications made by Scherer (1986) regarding the acoustic correlates of the human voice across different emotions states, finding that the fundamental frequency (Base Frequency parameter) of the human voice increases with *happy*, while decreasing with *sad* and *anger*, with the range of the fundamental frequency (Frequency Range parameter) following a similar pattern. Also, with respect to the Speech Rate, it was found that with a *happy* emotion, speech rate went up (meaning a flagger Speech Rate and a Smaller pause Ratio), while going down with a *sad* emotional state.

As well as calculating the correlations between the individual input and output values, the correlation coefficients between each of the output values were also calculated using Spearman's ρ . These coefficients are shown in table 7.8. These coefficients reveal valuable insights as to how the different outputs co-vary. For example, a significant ($p < 0.01$), positive correlation between the Base Frequency and Frequency Range is identified ($\rho = 0.806$). Similarly, a significant ($\rho = -0.813$), negative correlation is identified between the Speech Rate and Pause Ratio parameters. The Rhythm parameter was also found to be with both the Speech Rate ($\rho = 0.463$, $p < 0.01$) and Pause Ratio ($\rho = -0.626$, $p < 0.01$). The corrections between the Speech Rate, Pause Ratio and Rhythm parameters are all intuitive in that each of these parameters directly influences the temporal characteristics of an utterances. In order to produce a *fast* utterance, the Speech Rate should be high, with a small Pause Ratio (with the opposite being true for producing *slow* utterances).

From these correlations, more general characteristics (that are specific to this mapping) may be extracted regarding the features of utterances and their respec-

Table 7.7: General description of utterance characteristics in different regions of the AffectButton PAD affect space.

Affect Space			Label	Utterance Description
P	A	D		
1.0	1.0	1.0	Excited	Fast, high pitch utterances, with short pauses a high frequency range and a clear <i>beat</i> and a medium tremolo.
0.5	0.11	0.5	Happy	Same as Excited, but overall to a lesser degree, however the rhythm is more renounced.
-1	1.0	1.0	Angry	Medium pitch and frequency range and a high tremolo. Utterances are slow, with medium pauses and medium rhythm.
-0.5	0.11	0.5	Annoyed	Similar to Angry, but slower utterances, little rhythm, high tremolo and low base frequency and frequency range.
1.0	1.0	-1.0	Surprised	Medium base frequency and range, with medium fast speech rate and rhythm, but a low pause ratio. Tremolo is high.
0.5	0.11	-0.5	Content	Similar to surprised, but lower frequencies rhythm, but it slightly faster speech and slightly higher pause ratio. Tremolo remains high..
-1.0	1.0	-1.0	Scared	Medium base frequency and range. Slow speech rate and high pause ratio, and a medium rhythm value. Tremolo is high.
-0.5	0.11	-0.5	Sad	Low base frequency and range. Slow utterances with a high pause ratio, but with high rhythm. Tremolo is low.
0	-1	0	Relaxed	Medium base frequency and range, and medium speech rate and pause ratio. Rhythm is high and tremolo is low.

tive (and supposed) affective meaning. Table 7.7 outlines the utterance characteristics for each of the nine affective prototype locations within the AffectButton PAD space.

Figures 7.7f, 7.7g and 7.7h stand apart from the other plots in figure 7.7 as in these there is a clearly visible diagonal peak/trough in the mappings. These diagonals coincide (with respect to the affect space) with the data collected from the Categorical Perception experiments and are due to the modification of the AffectButton, where the input PAD space was limited such that the Pleasure and Dominance ratings were equal to each other, and their values limited to fall within the range $[-0.5, 1]$. Furthermore, in the utterances used in this experiment, the

values for the Tremolo, Sound Unit Count and Wave Type were all fixed and as such have been learnt by the ANNs, which is also something that impacts the subject evaluation of the mappings.

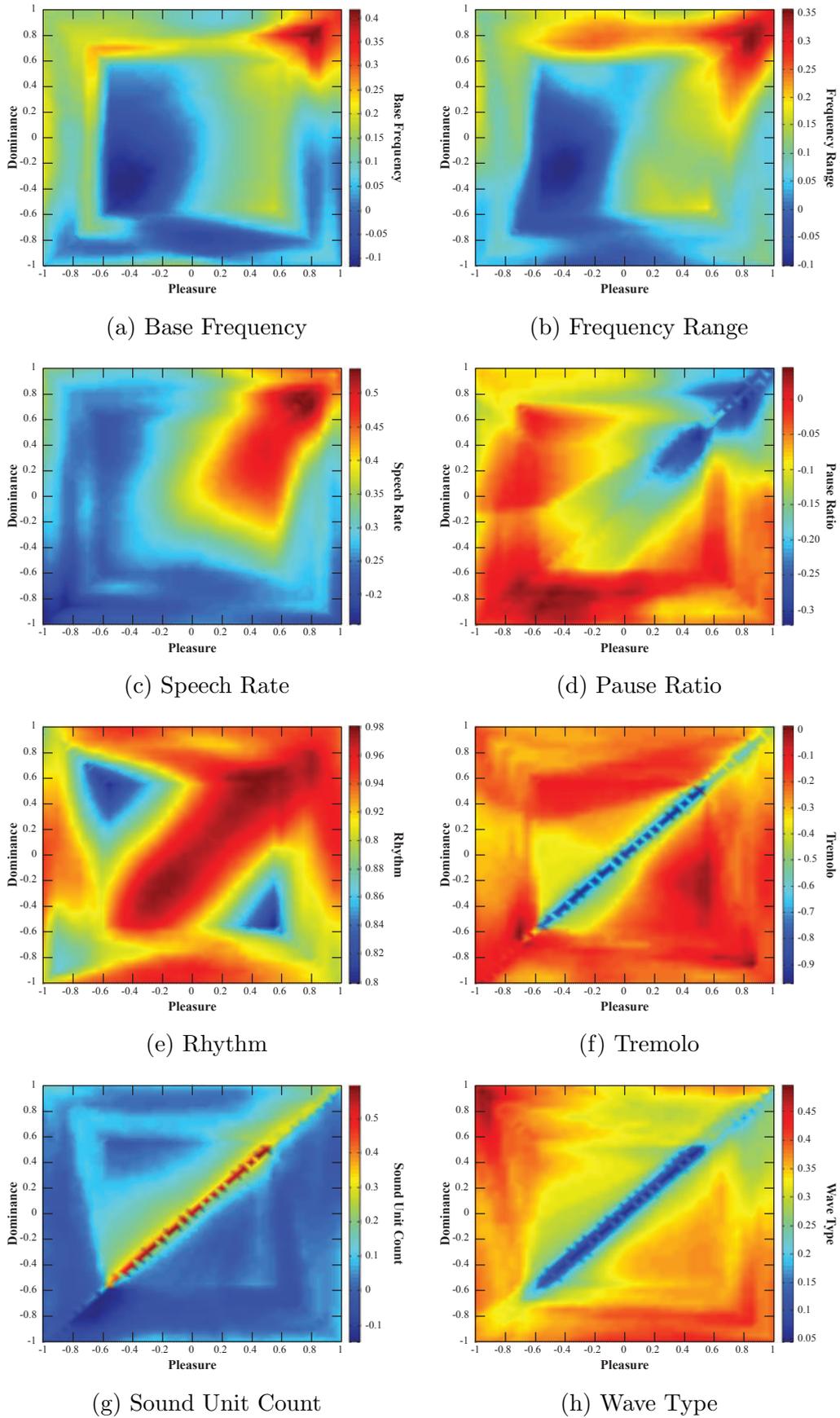


Figure 7.7: Plots of average mappings learnt by the independent ANNs for each of the 8 parameters.

Table 7.8: Spearman's ρ Correlation Coefficients for the correlation between the different ANN output values of the utterance parameters.

	Base Freq	Freq Range	Speech Rate	Pause Ratio	Rhythm	Tremolo	Sound Unit Count	Wave Type
Base Freq	1.0							
Freq Range	0.806**	1.0						
Speech Rate	0.463**	0.602**	1.0					
Pause Ratio	-0.402**	-0.545**	-0.813**	1.0				
Rhythm	0.106**	0.091**	0.463**	-0.626**	1.0			
Tremolo	-0.021	-0.006	-0.202**	0.439**	-0.565**	1.0		
Sound Unit Count	0.084**	0.032	0.343**	-0.502**	0.324**	-0.463**	1.0	
Wave Type	0.058*	-0.060*	-0.480**	0.523**	-0.486**	0.408**	-0.316**	1.0

* : $p < 0.05$ ** : $p < 0.01$

7.4 Subject Evaluation

Up to this point, this chapter has been concerned with the training of ANNs to learn an affective mapping between input coordinates within an affect space, and output values for parameters that characterise NLUs, the results of which have been described above. This section details an experiment in which these mappings are evaluated and tested in order to assess their validity. In essence, this evaluation seeks to uncover whether the mappings indeed produce NLUs which subjects are able to accurately and coherently decode, and having the desired affective interpretation. If an NLU n is generated using the input coordinate (p_i, a_i, d_i) , that subjects indeed provide an affect rating (p_r, a_r, d_r) that is similar to (p_i, a_i, d_i) . If this is not the case, then it suggests that either the mapping is incorrect, or that subjects in general find it cumbersome to attribute affective ratings to NLUs coherently, when presented outside of an interaction context and in an context-free scenario.

7.4.1 Experimental Setup

The experiment was setup with a local primary school where young children (aged 7-8 yrs) were recruited to partake. The children were presented with NLUs generated (and played through the robot's onboard speakers) using parameters output by the ANNs and asked to rate these using the AffectButton by assigning facial expressions to each utterance. If the mappings that have been learnt by the ANNs are indeed *correct* (i.e. subjects are sensitive to the different acoustic cues in the utterances and are able to decode the affective meaning accurately), then one would expect to see a strong correlation between the PAD values used to generate utterances and the corresponding PAD ratings associated with those utterances

As with the experiment in chapter 6, since the AffectButton was used to capture affective ratings, subjects completed three different tasks: a Matching Task, an Identification Task, and finally a Labelling Task. These are explained in sections 7.4.1.2, 7.4.1.3 and 7.4.1.4 respectively.

7.4.1.1 Stimulus Production

13 PAD values from the AffectButton input space (figure 7.8) were selected and input to trained ANNs to obtain mapped values for the Base Frequency, Frequency Range, Speech Rate Pause Ratio and Tremolo parameters (table 7.9). Given the small range of the mapping values for many of the parameters, for this experiment, the mapping values were normalised such that they fell into, and fully covered the numeric range of $[-1\ 1]$, thus also covering the full working range of the generation algorithm. Had this not been done then it would be likely that many of the NLUs would have sounded very similar if not identical. The Sound Unit Count, Rhythm, Wave Type and Pitch Contour values were all pre-specified and held constant.

Two different pitch contours were predefined. This was done as no mapping was learnt for this parameter and allowing this to be randomised for each of the utterances generated would introduce an random element that would not be accounted for when presenting these utterances to subjects. The difference between the two pitch contour specifications was that the contour shape of the first and last sound unit were identical, both set to *rising* in the case of the first specification, and both set to *falling* in the case of the second specification (see figure 7.10). This was done in order to test whether the different pitch contours would evoke different affective ratings from the subjects.

The rhythm parameter was held constant at a value of 1 for all utterances as this parameter, when below a value of 1, controls the amount of random variance with which the duration of each sound unit is specified. Thus, if the value is not equal to 1, Then it cannot be guaranteed that each NLU would be identical with respect to the duration of each individual sound unit. the sound unit count was held constant with a value of 5, keeping to the aim of studying longer, nor complex utterances, rather than short, single burst *iconic* utterances. Finally, only sine waves were used as the carrier signals, as has been the case in the majority of the work in this thesis.

For each of the in PAD input values, the ANN output values and pre-specified parameter values were input to the NLU generation algorithm and the outputs

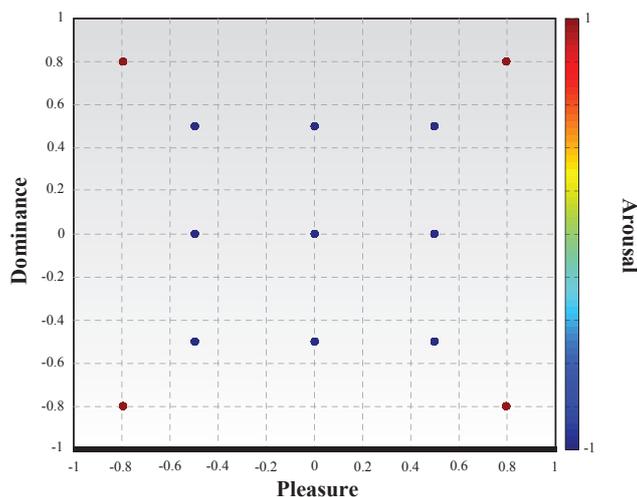


Figure 7.8: AffectButton PAD values used as inputs to the trained ANNs used to obtain the output utterance parameter values used to generate the experimental NLUs.

recorded. This was done for each of the two Pitch Contour specifications (CS1 and CS2), thus producing a total of 26 experimental utterances.

7.4.1.2 Matching Task

The matching task was used as a means to help subjects explore the range of the facial gestures that the AffectButton is able to produce. Subjects were first guided through the buttons input space by the experimenter. Once complete, subjects were presented with an image of a prototypical AffectButton face and asked to match the face of the onscreen AffectButton to the face in the image (see figure 7.11a). Eight of the nine prototypical AffectButton faces were used (see figure E.1) and presented to subjects in a random order. If it appeared that subjects were struggling to match a face, the experimenter, provided a suggestion as to the region of the AffectButton input space that would produce the facial gesture to be matched.

7.4.1.3 Identification Task

Once acquainted with the AffectButton via the matching task, subjects were then asked to listen to the experimental NLUs and assign a facial gesture to match how they thought that the robot felt. Subjects were first presented with five test

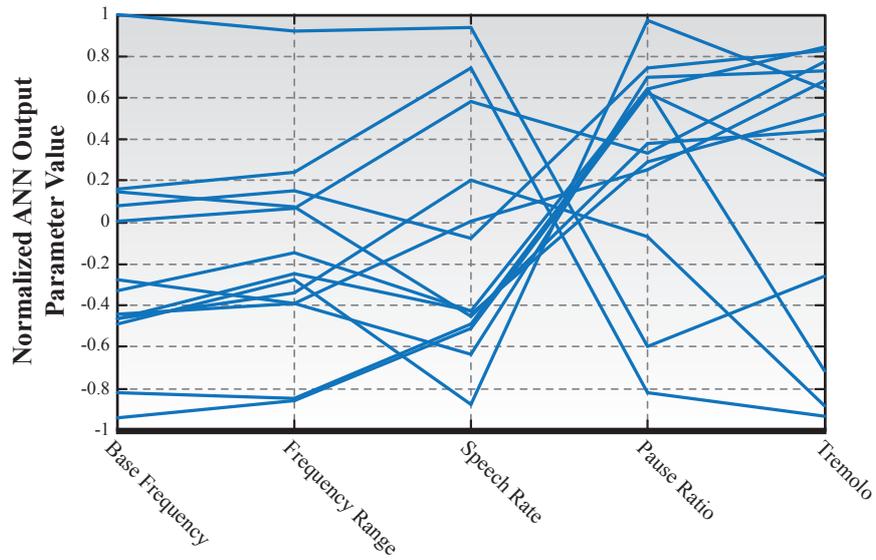
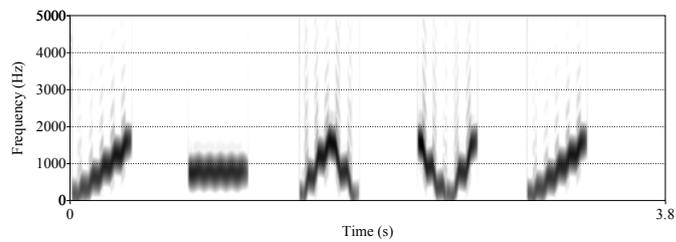
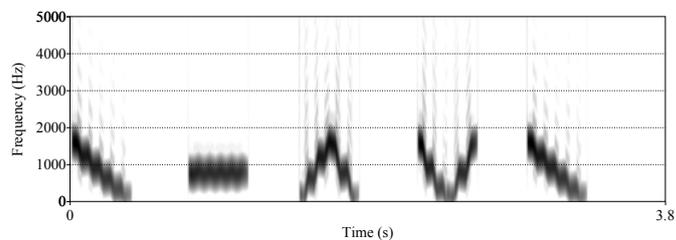


Figure 7.9: Normalised ANN output values for the utterance parameter obtained from the PAD inputs.



(a) Contour Specification #1 (CS1).



(b) Contour Specification #2 (CS2).

Figure 7.10: Spectrograms of NLUs with the two different pitch contour specifications.

Table 7.9: Obtained Mappings between the PAD values input to the ANNs and the output parameter values, scaled to fit the working range of the NLU generation algorithm.

NLU	PAD Values			Parameter Values				
	P	A	D	B. Freq	F. Range	S. Rate	P. Ratio	Trem
1	-0.8	0.111	-0.8	753.9	861.5	1.33	0.934	0.25
2	-0.8	0.111	0.8	1072.97	1036.8	2.37	0.66	0.17
3	-0.5	-1.0	-0.5	528.35	570.31	2.22	0.78	-0.28
4	-0.5	-1.0	0	590.89	574.70	2.27	0.77	0.09
5	-0.5	-1.0	0.5	775.78	803.84	1.92	0.81	0.29
6	0	-1.0	-0.5	837.17	925.42	2.43	0.61	0.20
7	0	-1.0	0	767.08	828.70	4.01	0.44	-0.34
8	0	-1.0	0.5	864.10	807.17	3.52	0.59	0.27
9	0.5	-1.0	-0.5	1040.87	1077.11	3.31	0.83	0.33
10	0.5	-1.0	0	1001.85	1034.02	4.96	0.63	0.30
11	0.5	-1.0	0.5	1080.22	1120.10	5.36	0.08	-0.36
12	0.8	0.111	-0.8	767.54	873.31	2.44	0.78	0.33
13	0.8	0.111	0.8	1500	1460.77	5.83	0.19	-0.10

utterances, before being presented with the 26 experimental NLU. The order of the test utterances was held constant for all the subjects, while the order of the experimental utterances was randomised. Subjects were allowed to hear each utterance as many times as they wished, but were not permitted to hear any previous utterances.

7.4.1.4 Labelling Task

Finally, in order to assess the subjects' ability to use the AffectButton, subjects undertook a labelling task, whereby they were asked to assign a facial gesture to match an affective label (see figure 7.11c). The labels that were used were: *Excited*, *Happy*, *Angry*, *Surprised*, *Sad*, *Scared* and *Calm* and were presented in a random order. As with the labelling task used in the Categorical Perception experiment in chapter 6, this choice of labels was motivated by the prototype faces that are hardcoded in the AffectButton and the overlap with the theory of basic emotions (Plutchik, 1994), with the assumption that children would be familiar with the affective labels and have an good understanding of there meaning.

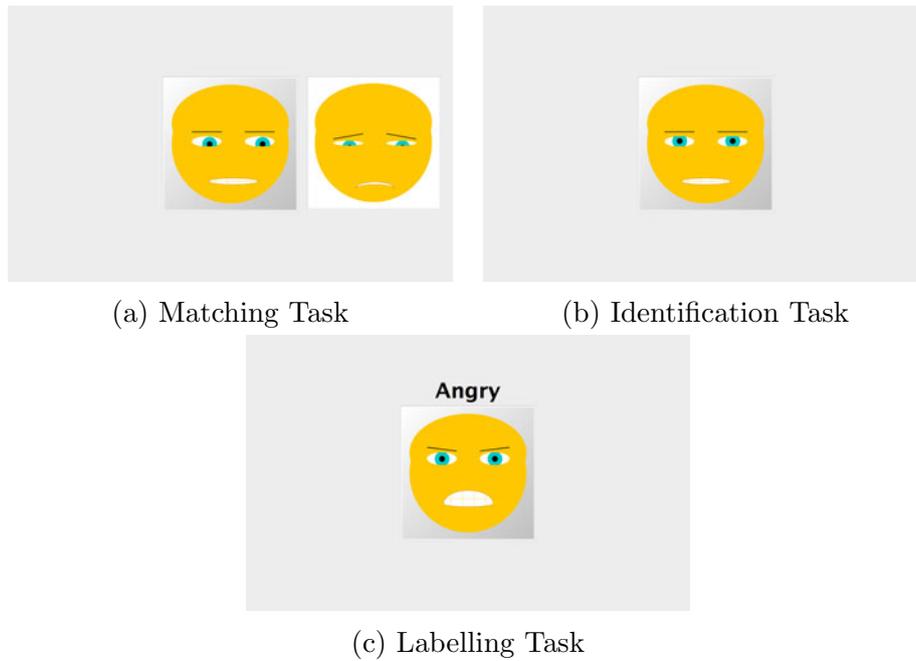


Figure 7.11: Images of the subjects' laptop for each of the three tasks performed during the experiment.

7.4.1.5 Experimental Procedure

Subjects were recruited via two year 3 (7-8 yrs) classes in local primary school, and were taken out of class time individually to partake in the experiment. The experiment itself was conducted in a spare classroom in the school, and spread over two days of the week (one day for each class). The children were seated in front of a laptop computer, with the Nao robot standing on the table facing the child. A second laptop was operated by the experimenter, and was used to manage the experiment and orchestrate the data flow between the robot and the subjects laptop, collecting all the data in a single location. Figure 7.12 shows an image of the experimental set up.

The Nao was programmed such that when touching it on the head, it cycled through the different stimuli in each task. This was controlled by the experimenter. The robot was also programmed to exhibit some natural behaviour (e.g. gazing, shifting weight, moving fingers and arms) as a means to avoid presenting the robot as a *static* object, but rather promote the idea of the robot being “alive”.

When each child entered the room, they were seated in front of the robot. The



Figure 7.12: Image of the experimental setup.

experimenter explained that the robot, called “Pop”, was a little different from humans as he didn’t speak in english, but rather, he spoke with sounds, however, we were unable to understand what Pop was saying when he spoken, and how he felt when he was speaking. The children were told that they needed to tell the experimenter how they thought Pop was feeling when he said something. Following this, the children were asked whether they had used a laptop computer before, and were shown how to use the trackpad, and how to control the AffectButton face³. Once acquainted with the laptop itself and the use of the trackpad, the children performed the Matching, Identification and Labelling Tasks, in that order.

No reward was given after the experiment, however at the end of each day, the class was given a little demonstration of the Nao, and were free to ask any questions about the robot.

7.4.2 Results

In total, 25 children (aged 7-8 years old) partook in the experiment: 12 boys and 13 girls. The average time taken to complete the experiment was 12 minutes (std = 3.5 mins).

As the matching task served as a means to help subjects become acquainted with the AffectButton, with the experimenter, in some cases, providing suggestions to the subjects, the data collected regarding the PAD values provided by

³The movement of the mouse cursor was limited to within the area of the AffectButton in order to make the task simpler and less confusing for the children.

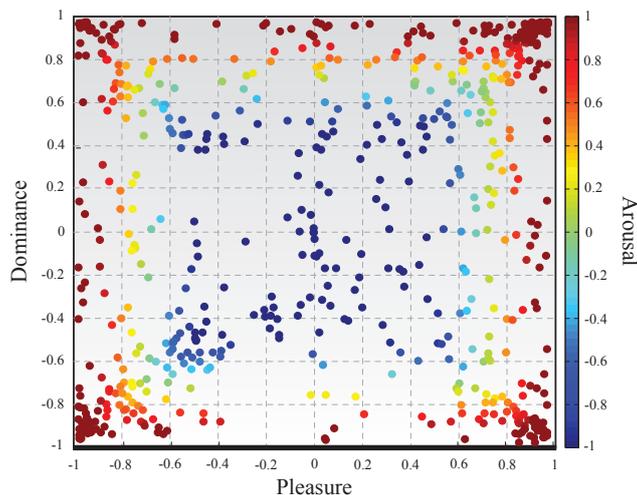


Figure 7.13: Plot of the PAD ratings for the 26 NLUs presented to subjects during the Identification Task.

the subjects for each face are not subject to analysis.

The data for the Identification and Labelling Tasks were also found to be non-normal. Due to this (as with the experiments in chapter 5), traditional tests such as the (M)ANOVA are not applicable as a means of performing analysis as the assumption of data normality is violated. As such, the data collected has been subject to non-parametric tests as an alternative.

7.4.2.1 Identification Task Results

Figure 7.13 shows a plot of all the subject PAD ratings obtained for the experimental NLUs. A visual inspection of the data points indicates, as with the ratings observed in the experiment in chapter 5, that the subjects tended to provide extreme ratings: ratings are drawn to the corners of the affect space. We again appear to see characteristics of categorical perception (chapter 6)

Cronbach's α was used as the measure of agreement between subjects in their ratings. The α values were calculated for the ratings grouped by the two genders (males, female and both) as well as the ratings for the two different pitch contour shapes (CS1, CS2 and both). The values are shown in table 7.10 and show that subjects had, in general weak agreement along the Pleasure and Dominance dimensions, while having notably high agreement along the Arousal dimension. This shows that while subjects had little agreement in their selection along the

Table 7.10: Cronbach’s α values indicating the degree of agreement between subjects in their ratings of the NLUs. α values are shown for ratings grouped by gender, and for each of the pitch contour specifications.

Subjects	Contour	Affect Space Dimension		
		Pleasure	Arousal	Dominance
All	Both	0.378	0.808	0.433
	CS1	0.218	0.693	0.271
	CS2	0.341	0.681	0.367
Females	Both	0.536	0.793	0.429
	CS1	0.197	0.668	0.312
	CS1	0.581	0.703	0.174
Males	Both	0.079	0.818	0.476
	CS1	0.288	0.724	0.087
	CS2	-0.339	0.658	0.491

Table 7.11: Pearson Correlation Coefficients (ρ) between the input and observed PAD values. None of the ρ values are statistically significant at the 0.05 level.

		Subject Ratings		
		P	A	D
Input PAD Values	P	-0.034	-0.023	0.031
	A	0.042	0.044	0.025
	D	-0.078	-0.004	-0.024

Pleasure and Dominance dimensions (i.e. along the horizontal and vertical axis of the AffectButton input space), they did exhibit strong agreement in their rating with respect to the *radial* distance from the horizontal and vertical origin (i.e. the center co-ordinates of the button). This too supports the notion of categorical perception being present.

Table 7.11 shows the Pearson Correlation Coefficients (ρ) between the PAD values used as inputs to the ANNs, and the subject ratings for each of the generated NLUs. All the values are near zero and show that, overall, subjects did not *coherently* provide ratings that were *similar* to the original PAD values used to generate the NLUs, nor did they *coherently* provide ratings that were *dissimilar* to the original PAD values. Rather, the values show that there is little relation between the input PAD values used to generate the NLUs, and the ratings for these NLUs.

The findings of the Cronbach’s α and the Pearson Correlation Coefficients are also echoed in table 7.12, which shows the partial linear correlation coefficients

Table 7.12: Partial Linear Correlation Coefficients (ρ) between the parameters of the experimental NLUs and their PAD affective ratings.

Parameter	Affect Space Dimension		
	Pleasure	Arousal	Dominance
Base Frequency	-0.045	0.025	0.039
Frequency Range	-0.032	0.024	0.051
Speech Rate	-0.068	-0.004	0.028
Pause Ratio	0.078	0.002	-0.027
Tremolo	0.051	0.023	-0.002

between each NLU parameter, and the PAD (along each dimension, accounting for the remaining dimensions) obtained from the subjects. The table shows that there is essentially no correlation between any of the 5 varied parameters (Base Frequency, Frequency Range, Speech Rate, Pause Ratio and Tremolo) and the affective ratings, as all the ρ values are low.

Friedman tests were used to perform a non-parametric one-way repeated measures ANOVA, testing for differences in the ratings of the NLUs. This was done for each affective dimension independently, with the data being isolated by subject gender (males, females and both) as well as the pitch contour shapes (CS1, CS2 and both). Figure 7.14 shows plots of the mean values and standard deviations of the ratings for each NLU specification, for both pitch contour specifications. The table shows that there was significant differences between any of the the ratings for the different NLUs along any of the affective dimensions. This was true when grouping the data by subject gender and the pitch contour specification. These results are summarised in table 7.13.

Friedman tests were also employed to compare the ratings across the two different pitch contour specifications (PC1 and PC2). As with the tests checking for differences due to the different utterance specifications, no significant differences were found due to the different pitch contours, along any of the three affective dimensions, regardless of the subject gender. These results are summarised in table 7.14.

Kruskal-Wallis (KW) tests were used to check for differences in ratings between the two genders, along each affective dimension individually. These tests found no significant differences in the affective rating provided by the two genders along

Table 7.13: Results of the Friedman tests, testing for significant differences in the affective ratings for each of the NLUs. Ratings are grouped by gender as well as the contour specification. The table shows the degrees of freedom for each test, the χ^2 values and the associated p values.

Subjects	Contour	<i>d.f.</i>	Affect Space Dimension					
			Pleasure		Arousal		Dominance	
			χ^2	p	χ^2	p	χ^2	p
All	Both	25	26.27	0.393	20.09	0.742	22.6	0.601
	CS1	12	10.09	0.608	11.47	0.489	11.34	0.500
	CS2	12	14.34	0.280	9.31	0.677	10.13	0.605
Females	Both	25	26.19	0.398	24.65	0.482	27.98	0.309
	CS1	12	13.73	0.318	7.56	0.819	12.85	0.380
	CS2	12	15.64	0.208	16.05	0.189	10.65	0.559
Males	Both	25	34.11	0.106	24.83	0.472	21.83	0.645
	CS1	12	14.77	0.254	15.57	0.212	10.1	0.608
	CS2	12	15.27	0.227	10.11	0.606	9.33	0.675

Table 7.14: Results of the Friedman tests, testing for significant differences in the ratings across the two different Pitch Contour Specifications with ratings grouped by subject gender. The table shows the degrees of freedom for each test, the χ^2 values and the associated p values.

Subjects	<i>d.f.</i>	Affect Space Dimension					
		Pleasure		Arousal		Dominance	
		χ^2	p	χ^2	p	χ^2	p
Both	1	1.58	0.209	0.18	0.671	0.02	0.898
Females	1	0.14	0.706	0.1	0.751	1.9	0.168
Males	1	1.91	0.167	0.81	0.369	2.26	0.133

the Pleasure ($\chi^2(1) = 1.4$, $p = 0.236$), Arousal ($\chi^2(1) = 3.12$, $p = 0.075$) or Dominance dimensions ($\chi^2(1) = 0.11$, $p = 0.735$).

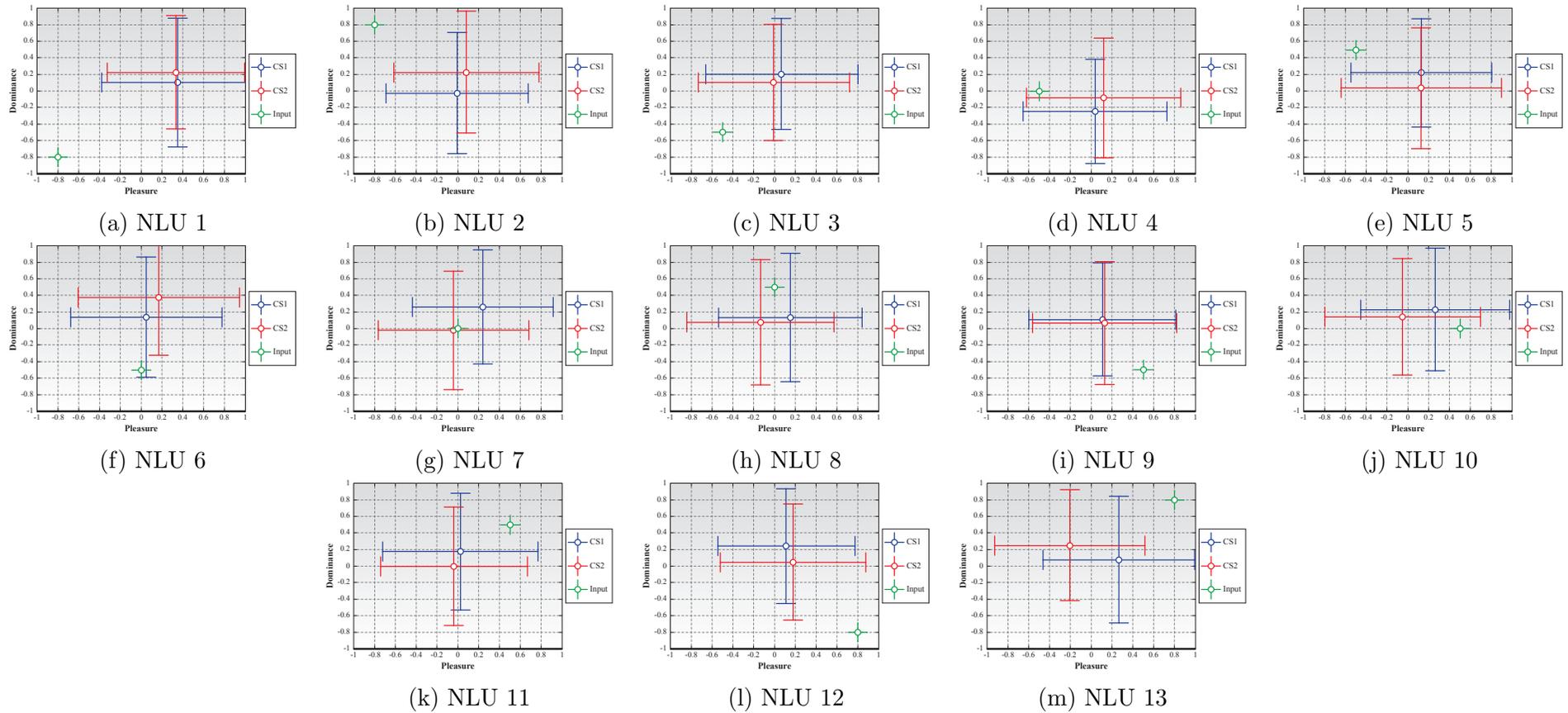


Figure 7.14: Mean and Standard Deviations of the ratings for each NLU specification. Blue data shows the ratings for CS1 and red points for CS2. The green point shows the original PAD input value used to generate the NLU parameters. The descriptive statistics for these plots are summarised in table E.1.

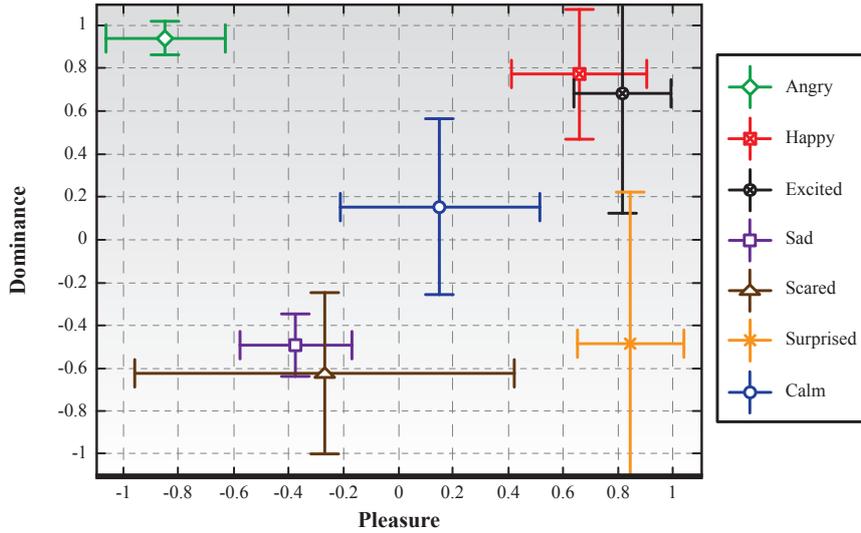


Figure 7.15: Plots of the mean and standard deviations of the ratings for each affective label in the Labelling Task (see table 7.16 for a summary of the descriptive statistics). Figure E.2 shows the AffectButton facial gestures for the mean ratings for each of the affective labels.

7.4.2.2 Labelling Task Results

Cronbach’s α values were calculated for the label ratings, along each of the affective dimensions, with data grouped by subject gender. The α values are summarised in table 7.15. Overall, the values are low, and in the case of the Pleasure dimension, negative. This indicates that there was little agreement in how subjects used the AffectButton.

Figure 7.15 shows a plot of the mean and standard deviations of the ratings for each affective label. A visual inspection suggests that while there are low Cronbach’s α values, subjects provided a broad variety of ratings and appear to have been able to differentiate between some of the labels along both the Pleasure and Dominance dimensions. To further investigate this, Friedman tests was used to check for significant differences in the ratings for each of the affective labels.

Table 7.15: Cronbach’s α values calculated for the results obtained during the Labelling Task, across the subjects genders.

Subjects	Affect Space Dimension		
	Pleasure	Arousal	Dominance
Both	-0.281	0.363	0.216
Females	-0.836	0.365	0.397
Males	-0.310	0.271	0.026

Table 7.16: Mean and standard deviations of the ratings for each of the affective labels in the Labelling Task, along each of the affect space dimensions.

Label	Affect Space Dimension					
	Pleasure		Arousal		Dominance	
	Mean	Std	Mean	Std	Mean	Std
Angry	-0.850	0.218	0.968	0.097	0.940	0.080
Happy	0.658	0.246	0.582	0.559	0.774	0.302
Excited	0.815	0.177	0.817	0.405	0.682	0.559
Sad	-0.375	0.204	-0.690	0.458	-0.491	0.147
Scared	-0.269	0.690	0.510	0.713	-0.623	0.376
Surprised	0.845	0.193	0.950	0.107	-0.486	0.710
Calm	0.150	0.363	-0.601	0.502	0.155	0.409

The tests found that there were significant differences in ratings along the Pleasure ($\chi^2(6) = 118.59, p < 0.001$), Arousal ($\chi^2(6) = 103.63, p = 0.001$) and Dominance ($\chi^2(6) = 99.41, p < 0.001$) dimensions.

Pairwise Friedman tests were performed as post-hoc tests in order to identify which labels were significantly different in their ratings, with this being done for each of the three affective dimensions individually. The results of these tests ($\chi^2(1)$ values) are summarised in table 7.17. The tests found that in the majority of cases there were significant differences in the ratings for each affective label. However, it was also found that certain *similar* labels did not have significantly different ratings, thus revealing a degree of *confusion* between the labels. For example, the ratings for *Happy* and *Excited* were only significantly different along the Pleasure dimension ($\chi^2(1) = 4.84, p < 0.05$), with *Excited* having a higher rating than *Happy*. Also, the ratings for *Scared* received a higher Arousal rating than *Sad* ($\chi^2(1) = 16.67, p < 0.001$). Finally, there was no significant difference found between the ratings along the Dominance dimension between the *Excited* and *Surprised* labels. The means values and standard deviations of the ratings for each different affective label are summarised in table 7.16.

Finally, Kruskal-Wallis tests were employed to check for differences in how the two genders rated the affective labels. The tests found no significant differences along either the Pleasure ($\chi^2(1) = 0.56, p = 0.456$), Arousal ($\chi^2(1) = 0.08, p = 0.781$) or Dominance ($\chi^2(1) = 0.06, p = 0.812$) dimensions.

Table 7.17: Results of the pairwise Friedman test comparing the ratings between of the affective labels in the Labelling Task. The table shows the $\chi^2(1)$ values with an indication of the degree of statistical significance along each affect space dimension.

Label		Ang	Hap	Exc	Sad	Sca	Sur	Calm
Angry	P							
	A	-						
	D							
Happy	P	25‡						
	A	11.27‡	-					
	D	8.17†						
Excited	P	25‡	4.84*					
	A	0.11	2	-				
	D	2.67	0.36					
Sad	P	17.64‡	25‡	25‡				
	A	25‡	17.64‡	25‡	-			
	D	25‡	25‡	17.64‡				
Scared	P	11.56‡	14.44‡	17.64‡	0.36			
	A	6.25*	0.53	2.57	16.67‡	-		
	D	25‡	25‡	17.64‡	3.24			
Surprised	P	25‡	14.44‡	3.24	25‡	14.44‡		
	A	3	5.56*	0	25	5.4*	-	
	D	23‡	11.56‡	13.5‡	4.84*	0.04		
Calm	P	25‡	21.16‡	21.16‡	17.64‡	3.24	25‡	
	A	25‡	21.16‡	21.16‡	0.22	19.17‡	25‡	-
	D	25‡	17.64‡	9†	14.44‡	14.44‡	9†	

* : $p < 0.05$

† : $p < 0.005$

‡ : $p < 0.001$

7.4.3 Summary of Results

To provide a brief summary of the results, the results of the Matching task were not subject to statistical analysis as the purpose of this task was primarily to aid subjects in the process of exploring the AffectButton input space.

For the Identification Task, it was found that there was no real relationship between the PAD values used to generate and affectively charge NLUs, and the ratings that subjects gave these utterances. This was furthered as there was also no relation found between the values of the utterance parameters and the subjects' ratings. While there were 26 different utterances rated, each with notably different parameter values, were no differences in how subjects rated these utterances - in essence, subjects showed little coherence in their affective interpretations of the

utterances. Finally, it was also found that there was no difference in how the two genders rated the utterances, nor was there a difference due to the different pitch contours.

The results of the Labelling Task found that subjects indeed appeared to be coherent in how they assigned facial gestures to different affective labels. More importantly, the results demonstrate that the children were able to discriminate between the different affective labels along affective dimensions that differentiate the affective labels. This is taken as strong evidence that the children were indeed able to use the AffectButton tool correctly.

7.5 Discussion

As this chapter can naturally be split into two sections: the design and implementation of the ANNs, and the subject evaluations, the discussion presented in this section will address these two sections individually.

7.5.1 ANN Design and Implementation

This chapter has presented the implementation of simple feed-forward ANNs to uncover potential correlations between the parameters and affect space dimensions based upon a training data set. While there are similarities between the mappings learnt by the ANNs and other works such as Laukka et al. (2005); Scherer (2003); Burkhardt and Sendlmeier (2000), there are problems that stand in the way of being able to make direct comparisons. For instance, the review works of Scherer (1986, 2003), Banse and Scherer (1996) have a focus around emotional labels rather than affective dimensions. Similarly the large scale review of affective expression in music and speech by Juslin and Laukka (2003) presents a similar problem. Given that these are works providing a review of fields of research, this highlights the more general issues that revolve around the use of different representations of emotions and affect and how these differing representations can make meaningful comparisons cumbersome. As such, the work presented in this thesis indeed also falls victim to the same shortcoming by the use of a highly

specific and novel affective measuring tool.

The end solution that has been employed has used a single ANN for each utterance parameter, rather than using a single network to handle all the inputs and outputs in one model. This solution was informed through brief explorations where it was found that a single network did not provide adequately robust mappings when compared to a collection of multiple networks. This may well be the case as the initial training data presented a fundamental problem where, if a single network were used, it would need to perform dimension expansion, rather than reduction. By splitting the training data into smaller sets, each tailored to a specific utterance parameter, and subsequently training individual networks on these sets, the problem suddenly becomes one of dimension reduction, a task at which ANNs have been shown to perform well (Marsland, 2009). Though it can be argued that doing this came at the expense of potentially overlooking relationships and interactions between the output parameters, the similarity between the obtained mappings and the findings of the related literature would suggest that it has not been a problem. Rather, it appears to have been a fruitful choice to make.

The use of feed-forward ANNs also has the characteristic of function approximation. The ANNs provide a single surface mapping between the inputs and outputs, which inherently always produces the same output value for a given input value. With respect to generating NLUs, what this entails is that, for certain parameters at least (e.g. the Base Frequency, Frequency Range, Speech Rate and Pause Ratio), the networks will produce very similar (if not identical) characteristics within utterances, which runs the risk of having repetitive utterances for a given location in the input affect space. An easy way to overcome this is to add noise to the outputs and/or inputs, thus introducing random variability and reducing this repetitiveness.

While this issue of a static mapping is not necessarily a negative characteristic as it simplifies the initial investigation at hand, it does raise the question of how to avoid having repetitive utterances, which can be a problem in HRI as it runs

a risk of the robot being perceived as pre-programmed which tends to have an adverse effect on how people view the robot (Belpaeme et al., 2012; Ros Espinoza et al., 2011). There are two possible ways to overcome this beyond simply adding noise to the ANN outputs, and they are heavily intertwined. Randomising the number of sound units within an utterance is potentially one solution, as is pseudo-randomising the Pitch Contour. The intertwined nature comes from the fact that as the number of Sound Units in an utterance changes, so must the Pitch Contour. This chapter has purposely avoided the mapping of the Pitch Contour and used a fixed value of the number of Sound Units in an utterance as this research has far as been unable to gain a coherent understanding upon how this feature of an utterance actually impacts how peoples' affective interpretations. As such, further research addressing this is clearly required.

7.5.2 Subject Evaluation

The results of the Identification Task indicate that the subjects did not provide affective ratings that were similar to the original PAD input values used to generate the values for the NLU parameters in a coherent manner. This is evidenced by lack of correlation between the original PAD values and the obtained PAD ratings. Furthermore, there is also no correlation between the affective ratings provided by the subjects and the generation parameters, showing that subjects did not appear to associate particular generation parameters with different facial gestures/affective states. This is to be expected as the trained ANNs have provided a correlated mapping between the PAD affect space the the parameter values, and if there is no relationship between the input PAD values and the subjects' PAD ratings, there will also be no relationship for the utterance parameters either. More specifically, this may be characterised as subjects' not being coherent in their selection of the horizontal and vertical location of the mouse cursor in the AffectButton, but they did exhibit coherence with respect to the radial distance from the centre of the button.

Furthermore, the broad range of ratings for each utterance suggest that sub-

jects were very varied in their affective interpretations, perhaps to the degree that the ratings were *psuedo-random*, as the subjects did exhibit coherence in the use of the AffectButton along the Arousal dimension. However, the plot of the raw data (figure 7.13) shows clusters of data points in the areas where the prototypical affective facial expressions are located, indicating that while there is no real coherence between subjects, they did appear to have interpretations that were aligned with particular affective prototypes (e.g. happy, sad, angry, surprised): there appears to be a magnet effect/categorical perception again. This is in line with the results of chapters 5 and 6.

The results of the Labelling Task show that subjects did appear to exhibit a strong coherence in their use of the AffectButton. For example, it is clearly shown that subjects were indeed able to distinguish, to a statistically significant degree, between the *Angry*, *Sad* and *Scared* labels and the *Happy*, *Excited*, *Calm* and *Surprised* labels along the Pleasure dimension. Similarly, subjects also appear to have been able to distinguish between the *Angry*, *Happy* and *Excited* labels, and the *Sad*, *Scared* and *Surprised* labels along the Dominance and Arousal dimensions. In both these cases, it is clear that subjects were able to discriminate between the various labels along the dimensions that fundamentally differentiate the labels. These findings are more supportive of the notion that subjects were able to use the AffectButton in a manner whereby they were able to distinguish between the different basic emotions labels as well as the different facial expressions that are associated with these labels. This counters the notion that the lack of clear results in the Identification Task are due to an inability of subjects being able to use the AffectButton robustly.

Considering the results of both the Identification and Labelling tasks, the lack of overall coherence between subjects in their affective interpretations of NLUs, and the lack of correlation between the parameter values and the affective ratings draws out the question of why this is the case, given that the ANNs have provided mappings that have notable similarities with the findings that are reported in the psychology and musicology literature. Why is this the case, and are there

any ingredients missing currently that are hindering the ability for subjects to affectively interpret NLUs coherently?

On one hand, given that the ANN mappings have notable similarities with findings from related literature, it appears that the findings regarding the acoustic features of the human voice as well as music and how these relate to affective interpretation do not translate well, (if at all) to NLUs, though it does indicate the people are indeed sensitive, to some degree, to the acoustic features of NLUs. On the other hand it is notable that all the experiments thus far have presented utterances to subjects in a context-free manner: there has been no real task or interaction with a clear context within which the NLUs have been used by the robot/presented to the subjects, which could well be an factor that influences (as it does in animation) how NLUs are interpreted. Currently, with the data collected in this evaluation, it is not possible ascertain exactly where the problem lies. What can be said with confidence is that when utterances are presented in a context-free manner, and the utterances have acoustic characteristics that are similar to those reported in related literature, people do not appear to associate different types of sounds with a given affective meaning.

7.5.3 Methodological Remarks

This section addressees some potentially important methodological drawbacks that may have impacted the final subject evaluations that were performed and the results obtained.

7.5.3.1 ANNs versus other ML Techniques

The world of ML is now a large field of research with many active members exploring a wide variety of different ML techniques and algorithms and their applications to many different real world problems. As such, there are many different flavours and types of ML solutions that could have been used to tackle the regression problem that was faced in this chapter. While the justification for using ANNs, over these other techniques follows in the next subsection, this

section provides a brief outline of some of the other possible approaches (presented in particular order) that may have been used, and are recommend as alternative, and potentially for fruitful approaches for the future.

AdaBoost Regression is an approach to the regression problem that embraces the AdaBoost meta-algorithm. Put plainly, this algorithm acts as a means of assessing the outputs of many simple regression models (which are essentially *weak* in their performance) and combining these in a weighted format which results in a more accurate and complex regression and function approximation. While this approach requires that many different regression models be created and trained, these models are all simple and relatively computationally inexpensive which make this approach practically feasible from a real-time computing standpoint. This is something that is important for HRI.

Echo-State Networks are a branch of what is known a *reservoir computing*. These are recurrent neural networks with many sparsely connected nodes in the *reservoir*. Rather than having to learn the weights of the neurone connections within the reservoir, the last of the training algorithm is to learn the weights of the connections to the output layer, of which there are far fewer connections than connections within the reservoir. A key benefit to this method is that it is able to handle very non-linear problems and thus maybe used to model very non-linear dynamical systems.

Gaussian Process Regression is a regression method that has its roots in statistics. In this approach, the function that needs to be approximated can be modelled using a number of different gaussian distributions. With this approach, each predicted output based upon an input value has an associated probably which indicates how sure the model is of this prediction. Furthermore, very complex functions may also be approximated given that there is no limit set upon how many different gaussian distributions may ultimately be used.

Support Vector Machines Regression is yet another alternative approach to the regression problem. In this approach, the training data is mapped into a high-dimensional hyper-plane in which the complexity of the regression problem is then

reduced. A regression model is trained to the data as it sits in this hyper-plane. Regression is then practically performed by taking an input data point, identifying the data point's location in the hyper-plane, predicting the output value, and taking this value back to the original representation of the training data. The benefit of this approach is that the functions that may be mapped may be very complex indeed in their original dimensions, but once in a hyper-plane representation, only a simple regression model is needed in order to perform function mapping and prediction tasks.

This is only a brief outline of a limited number of different ML techniques that may be used to address regression problems. The following section serves to outline why in light of these more sophisticated and complex methods, feed-forward ANNs have been adopted in this chapter.

7.5.3.2 ANN Implementation

One issue to highlight is in the choice of machine learning technique used to learn the mapping. In this case feed-forward ANNs have been employed, with the justification that they meet the main requirement of the problem being presented: the approximation to non-linear relationships between the input and output variables - a regression problem. However, the field of machine learning now has a rich history, and a very large and active community that has developed and explored a large number of different techniques and algorithms, some of which are outlined above. One might ask, given this rich choice of different algorithms that all achieve similar outcomes, albeit via different methods, why choose one of the most basic approaches? Furthermore, why has there not been more extensive, deeper exploration of the setup, training and ultimate behaviour of the ANNs that have been used?

To address the first point, while feed-forward ANNs may be limiting in comparison to more sophisticated machine learning methods (e.g. radial basis functions Marsland (2009) or Gaussian Mixture Models Murphy (2012)), they are very well documented and met the requirements of the problem at hand, and thus it was

deemed as unwise to employ more sophisticated techniques that could potentially overcomplicate things. Furthermore, the purpose of this chapter (and this work more generally) is not to perform a deep exploration of how machine learning techniques can be used to generate and affectively charge NLUs, and which techniques may be more suited than others. Rather, this particular chapter serves as an initial stepping stone toward these more specific and applied questions, which themselves form a whole new direction of research. Given this, it is also more rational and productive in the long term to begin with a simple approach to the problem, and in doing so, any future research need only address the more sophisticated methods of approaching the problem of affective mapping.

The second issue of the limited exploration of the setup of the ANNs ultimately used, many of the same reasons as decried above apply. However, the main reason for this is not to draw the focus of attention away from the purpose of this chapter with respect to the general thesis presented here. This body of research is not about machine learning. It is concerned with investigating, more broadly, how NLUs may be utilised by a robot during real world social HRI. Within this scope, the use of ANNs has a primary function of providing a means to try and produce an affective mapping between the acoustic features of an utterance, and how utterances generated that are affectively charged using this mapping are subsequently interpreted. This was something that was missing with respect to the NLU generation algorithm. Suffice to say, this main goal of the chapter has been achieved. An affective mapping has been learnt using a collection of simple feed-forward ANNs and training data collected by specifically designed experiments. Moreover, the mapping that has been learnt has characteristics that are similar to the acoustic correlates that are reported in both the psychological literature and the music literature. Perhaps the most important insight that has been gained, via the subject evaluation, is that people (specifically, children) do not appear to coherently associate different affective states with the NLUs, while this has been the case with natural language, even when there are few contextual cues, as shown by Le Sourn-Bissaoui et al. (2013). Thus, the conclusion to draw here is that the

interpretations associated with the affective chartings of the human voice do not transfer well to their use in NLUs.

7.5.3.3 Subject Evaluation Methodology

Perhaps the first criticism to make regards the number of subjects who partook. When compared with the experiment used to collect the majority of the training data, nearly double the number of children partook. Had more children been used than perhaps a ten could have emerged from the data. A similar line of argument also applies to adults, who have not been evaluated here.

Also, one may make comment over the fact that only two different Pitch Contours were used. As the previous experiments have shown that pitch contour can matter, but not in all cases, it could be that the particular Pitch Contour specifications used could be those that did not make a difference. The Pitch Contour is problematic and elusive facet of an NLUs in a broader sense also, and the full extent of their influence is not yet understood. This is why they have been omitted from the machine learning process, and had their values pre-specified, rather than randomised for the evaluation. The knock on effect to this is that the number of sound units for each utterances was also held at a fixed value, which limits the evaluation for both shorter and longer utterances, and the overall generalisation of the insights gained.

Another criticism that can be made regards the limited number of NLUs samples from the affect space. In total, only 13 different PAD values were used to generate different utterance parameter specifications. The rationale for this is essentially two fold. Firstly, the PAD affect space locations that were selected were very near the nine prototype facial expressions in the AffectButton, and also represented, generally speaking, the extreme values of each of the mappings. Secondly, given the lessons learnt regarding experimental design from the experiment in chapter 5 (where the complex presentation of utterances resulted in a cumbersome statistical analysis), it was deemed suitable to have all subjects rate the same stimuli in order to retain the repeated measures experiential design.

7.6 Summary

This chapter has presented an exploration of the potential for automating the generation and synthesis of NLUs by employing Artificial Neural Networks to learn a mapping between desired affective interpretations and the parameter specifications required to achieve the desired affective interpretation.

Training data from the experiments in chapters 5 and 6 have been used to train the networks and has resulted in mappings between the acoustic parameters of the NLUs and the input affect space that share a number of similarities with the acoustic correlates of both the human voice and music with respect to affective expression. While in the experiments in which training data was collected, subjects tended to show limited coherence in their affective interpretations of the utterances, the fact that the mappings are similar to those in the related literature indicates that people are indeed sensitive, to some degree, to the acoustic features of the utterances, and do mildly associate different acoustic features of NLUs with different affective meanings.

A human subject evaluation was carried out with local school children in order to gain an assessment of whether the learnt affective mappings indeed did evoke affective interpretations in people in a more coherent manner. Using the same form of experimental method as in chapters 5 and 6 the results of the evaluation reveal that even when NLUs have acoustic features and correlates that are similar to those found in both the human voice and music, these do not translate to the same affective interpretations in people, or even to an increased coherence between subjects in this regard.

There are two potential factors that can lead to this. Firstly, that the notion that subjects can indeed interpret, coherently, NLUs as having distinct affective meanings is perhaps wrong, and as such the findings from the related literature do not translate to their use in NLUs in the same way. Secondly, it is also noted that all the experiments thus far have not included any concrete form of situational context, which when reflecting on the world on animation appears to play an important role in scaffolding and directing how people interpret NLUs.

The current experimental arrangement has not been able to differentiate these two possibilities, and as such, the latter factor is subject to investigation in the next chapter.

Chapter 8

The Influence of Situational Context upon the Interpretation of NLUs

Summary of the key points:

- The influence of a valenced physical interaction with the robot over how the affective meaning of NLUs is explored via an online experiment.
- Videos showing the robot either being subject to an action, making an NLU, or making the NLU in reaction to the action were shown to adults, and were rated along a valence scale.
- Results show that the affective meaning of an NLU is overridden by the valence of the action when the NLU is made in reaction to the action and that NLUs do not bias the perceived valence of the action.
- When the affective interpretations of the interaction and the NLUs are aligned, this evokes more extreme affective interpretations from people.

All the experiments presented thus far in this thesis have lacked any concrete situational context¹, with varying levels of coherence and agreement in affective interpretation between subjects. For example, chapters 5, 6 and 7 all found that children showed little coherence in their ability to assign facial expressions to their affective interpretations of NLUs. While presenting utterances without situational context provides a means of obtaining *base* levels of interpretation, which are not confounded, this is not a true representative of real world HRI, which is inherently embedded in a situational context. The work presented in this chapter explores how the inclusion of a situational context may influence the interpretation of NLUs.

The experiment presented in this chapter seeks to address two related questions: can the situational context override the interpretation of an utterance, and can an utterance override the interpretation of the situational context? These two questions are relevant for the use of NLUs in HRI as situational context is an inherent facet of an interaction and likely plays an influential role in how events in the environment are perceived and interpreted by people (or agents). With respect to the use of NLUs, it may be that the inclusion of context has the effect of *biasing* a subjects interpretation of what may otherwise be an ambiguous utterance. Conversely, it may also be possible that the inclusion of an utterance provides a means of biasing what may be an ambiguous situation to having a particular interpretation. It is likely that both cases are true, with the influence of an utterance being determined by the ambiguity of the context, and visa-versa. However, at this stage, given the difficulty in obtaining constant coherence and agreement between subjects in their affective interpretations of NLUs, it is reasonable to presume that, generally speaking, NLUs hold more ambiguity than context.

With respect to HRI, there have been few efforts to investigate how context and NLUs influence each other. Komatsu et al. (2010) investigated the ability of

¹Here, *Situational Context* relates to an interaction scenario in which the use of NLUs has been embedded where there are clear cues that relate to and guide the mood of the overall interaction as it is unfolding. For example, a game of chess, where there are clearly interpretable outcomes (such as a good or bad move, or winning or losing the game) which each have a valence (Castellano et al., 2013).

their artificial subtle expressions (ASEs) to influence and sway the decision making behaviour of a subject whilst playing a simple treasure hunt computer game where a robot was able to provide clues as to where the treasure was hidden. ASEs made by the robot were intended to convey the confidence the robot had in the accuracy of its clue. Their results suggest that when the robot made a sound with a *decreasing* frequency modulation, subjects rejected the clue provided far more frequently than when the frequency modulation was held constant (flat). This case demonstrates that the use of a simple utterance by the robot was enough to significantly alter the behaviour of the subjects, and that subjects were indeed sensitive to the acoustic features of their utterances. With respect to the two questions posed at the start of this chapter, this study provides evidence that utterances can be used to alter the interpretation of the situational context, and in turn through this, the behaviour of the subjects.

The ASEs used in the work by Komatsu et al. (2010) are, however, simplistic in comparison to the NLUs used in this thesis. ASEs consisted of short (500 *ms*) single tones with either a decreasing (400 *Hz* to 250 *Hz*) or flat, constant (400 *Hz*) frequency modulation, whereas the NLUs here are more complex, consisting of multiple tones concatenated, each with a different modulation, and have a similar temporal duration (i.e. they were all of similar length in time). Furthermore, the context of the game in which ASEs were presented was held constant, thus the influence of the context was not addressed. This chapter presents an experiment in which both the situational context and the NLUs presented in this context were varied, allowing both variables to be studied simultaneously.

8.1 Experimental Setup

The experiment set out to test the following hypotheses:

- H_1 : An NLU overrides the interpretation of a situational context.
- H_2 : The situational context overrides the interpretation of an NLU.

Situational context, in these hypotheses are external actions that happen to the robot (such as the robot being slapped), and have an implied affective interpretation. The NLUs are also presumed to have an affective interpretation (albeit more coarse), and are presumed to be generally more ambiguous than those of the contexts in this respect. The nature of the interaction between the context and NLUs is quantified through the affective interpretations provided by subjects.

To test these hypotheses, an online survey was set up using an online crowdsourcing service, where subjects were asked to affectively rate videos of a robot across five conditions:

- C_{NLU}^P : The robot emitting a *positive* sounding NLU.
- C_{NLU}^N : The robot emitting a *negative* sounding NLU.
- C_{Action} : The robot being subject to an action from a human (e.g. the robot being slapped on the head) with no response.
- C_{Action}^P : The robot being subject to an action from a human and emitting the NLU from C_{NLU}^P in reaction.
- C_{Action}^N : The robot being subject to an action from a human and emitting the NLU from C_{NLU}^N in reaction.

Conditions C_{NLU}^P and C_{NLU}^N provide the base interpretation for the respective NLUs without any situational context; a sanity check to see whether the NLUs are indeed interpreted differently. Similarly, condition C_{Action} obtains the base interpretation for the actions, again necessary before we can measure the relative influence of NLUs on actions. Conditions C_{Action}^P and C_{Action}^N assess the interaction between the NLU and the context when combined.

Relating these conditions to the hypotheses (see figure 8.1), if H_1 is true, it would be expected that, for example, the ratings for C_{Action}^P and C_{Action} would be significantly different while C_{Action}^P and C_{NLU}^P would have similar ratings - the NLU has been able to *pull* the rating of the action away from the original interpretation. Conversely, if H_2 were true, the opposite would be expected: the

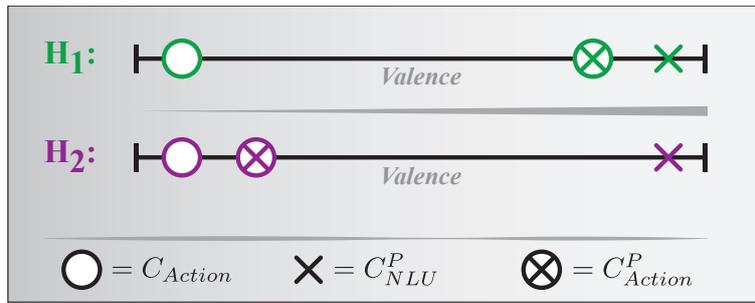


Figure 8.1: Example of how the video conditions C_{Action} , C_{NLU}^P and C_{Action}^P may hypothetically be rated, for each of the two Hypotheses.

ratings for C_{Action}^P and C_{NLU}^P would be significantly different while C_{Action} and C_{Action}^P would have similar ratings - the situational context has been able to *pull* the rating of an utterance away from the original interpretation.

Ratings were given along a 9-point Likert scale representing *valence*, where 1 corresponded to “very negative” and 9 to “very positive”. Subjects were asked to watch each video and provide a rating of how they thought the robot *felt* based upon the action that happened in the video, or the sound that the robot made in the video. As this was an online experiment, it was decided that using a measuring tool based on dimension representations of affect, such as the Self Assessment Manikin (Lang and Bradley, 1994) would introduce a level of complexity that would confound the results. The dimensions of the SAM are cumbersome to explain to naive subjects in person, let alone subjects who are recruited via online crowd-sourcing.

Five different action scenarios were selected and videos for the conditions C_{Action} , C_{Action}^P and C_{Action}^N produced, thus producing 15 videos in total. The five different actions provided a gradient of affective interpretations across the Likert scale (this is described in more detail in the following section). These videos are referred to as *Action Videos*. Similarly, five different NLUs were produced using the NLU generation method (chapter 3), and videos were recorded of the robot emitting the utterance. These are referred to as the *NLU Videos*. Two of these NLUs were used as the NLUs in conditions C_{Action}^P and C_{Action}^N as they were deemed to be most representative for positive and negative affective states respectively. The process of selecting the actions and NLUs, and producing the



Figure 8.2: Setup of the professional audio equipment used to capture the audio. The video recorder is located at the right hand side of the image (just out of sight).

videos for the five different conditions is detailed in the following section.

8.1.1 Stimulus Production

All videos (including the NLU videos) were recorded using both a digital video recorder, and professional audio equipment. The audio equipment consisted of a hyper-cardioid microphone directed at the speakers on the robot’s head to capture high quality audio of the utterances made by the robot, and an omnidirectional microphone to capture the more ambient sounds in the environment (e.g. motor activity and sounds caused by the actions of the human). Figure 8.2 shows a picture of this video and audio recording arrangement.

All audio captured via the microphones was recorded in stereo, and was then subject to the following post processing performed using the open source Audacity sound editing software². Each stereo channel was independently normalized to -1 dB. Background noise was removed from each channel using the noise removal algorithm within Audacity. The two channels were then merged into a single mono channel, with this mono channel being normalized again to -1 dB before being exported.

²Audacity may be downloaded from: www.audacity.sourceforge.net

8.1.1.1 Action Stimulus

11 Different action videos were recorded with the aim of identifying a subset of these that would represent a gradient from very negative actions, mildly negative actions, natural actions, mildly positive actions and very positive actions. Each video was no longer than 6 seconds long, with no speech being made by either the human or the robot. Videos were made of the robot being subject to the following actions:

- A slap to the side of the head.
- A poke in the chest with a board marker pen.
- A poke in the forehead with a board marker pen.
- A flick with the fingers to the forehead.
- Clicking fingers in front of the robot's eyes.
- Covering the robot's eyes with a hand.
- Waving a hand in front of the robot's eyes.
- Tickling the robot under the chin.
- Tickling the robot in the ribs.
- The robot being stroked on the head.
- The robot receiving a kiss on the head.

8.1.1.2 NLU Stimulus

8 NLUs were recorded using the NLU generation algorithm (see chapter 3) and were produced with a variety of acoustic parameter configurations (see table F.1 and figure F.1). As with the experiment in chapter 4, the acoustic parameters were not subject to explicit, systematic alterations, but rather were intended to elicit affective interpretations that would cover as broad a range of the Likert scale as possible. In this light, two utterances from the Categorical Perception

(CP) study were also included as it was thought that these would represent the extremes of the valence dimension (as they did in chapter 6). All utterances were normalized and recorded onto a mono track, such that each speaker onboard the Nao produced the same sound. Each utterance was then played through the robot, with the robot remaining stationary thus to have no background noise from motor activity, and only the audio being recorded.

A short video of the robot standing making small, neutral movements³ was made, again with the audio being recorded separately, and the audio recordings of the utterances were added to the audio track of the video to produce identical videos with different utterances. The total length of each NLU video was 4 seconds.

8.1.1.3 Pilot Study & NLU and Action Selection

A small pilot study was conducted to gain insight as to how the 11 different scenarios and 8 NLUs might be interpreted, and to guide the final selection of videos and utterances for the experiment. Subjects were recruited through the university, and were asked to watch each video individually and rate how they thought the robot felt based upon what happened in the video, or by the utterance that the robot made. The same 9-point Likert scale was employed to capture ratings. Videos were made available online, and subjects provided their ratings electronically by filling in and returning a spreadsheet to the author via email.

In total, 15 subjects responded. 10 were male (mean age = 28.4, std = 3.31), and 5 were female (mean age = 28.2, std = 4.32). Cronbach's α was used as a measure of internal agreement between subjects. For the Action Videos an α value of 0.9546 was obtained, while for the NLU Videos an α value of 0.8773 was obtained. Both values indicate strong agreement between subjects, with more agreement in the ratings of the actions videos than the utterances.

Figure 8.3 shows a bar graph of the mean values and standard deviations of the ratings for each of the 11 Action Videos. The results reveal that the Slap was

³Neutral movements were included in order to present the robot as *alive*, rather than a static object.

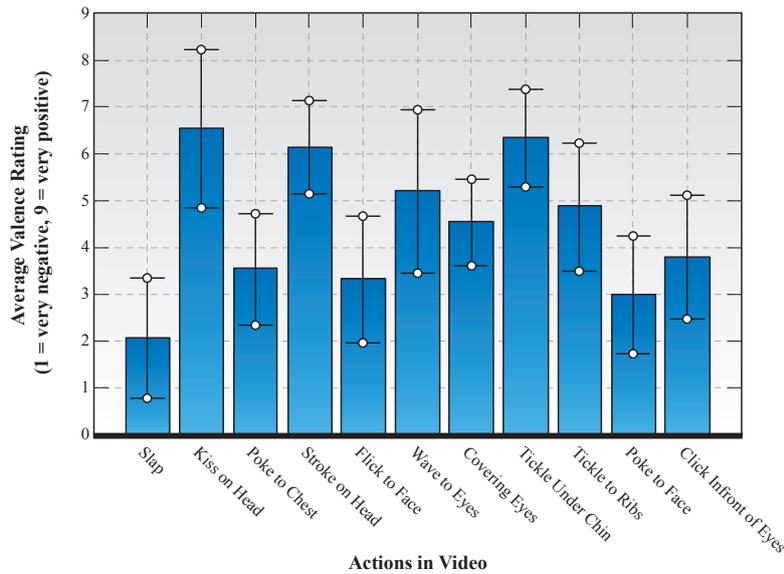


Figure 8.3: Bar graph showing the mean and standard deviations for the ratings of the Action Videos in the pilot study. These ratings are summarised in table F.3.

rated as the most negative action (mean = 2.0667, std = 1.2799) while the Kiss on the robot’s head was rated as the most positive action (mean = 6.533, std = 1.6847). The robot being stroked on the head (mean = 6.1333, std = 0.9904) and being tickled under the chin (mean = 6.3333, std = 1.0465) also received similar mean ratings to the Kiss on the Head. The video showing someone waving their hand in front of the robots eyes (mean = 5.2000, std = 1.7403), someone covering the Nao’s eyes (mean = 4.5333, std = 0.9155) and tickling the robot on either side of the Torso (mean = 4.8667, std = 1.3558) were rated approximately in the middle of the scale, and thus may be considered as neutral. The videos showing a poke to the chest (mean = 3.5333, std = 1.1872), a flick to the head (mean = 3.3333, std = 1.3452), poke to the face (mean = 3.0000, std = 1.2536) and finger clicking in front of the robot’s eyes (mean = 3.8000, std = 1.3202) all received mean ratings between 2 and 4 and as a result can be considered as mildly negative when compared with the slap action. Overall, it can be seen that the videos provide a gradient across ratings, while not covering the full range of the scale, with respect to the positive region of the scale in particular.

Figure 8.4 shows a bar graph of the mean values and standard deviations of the ratings for the NLU videos. A notable feature of these ratings is the small range

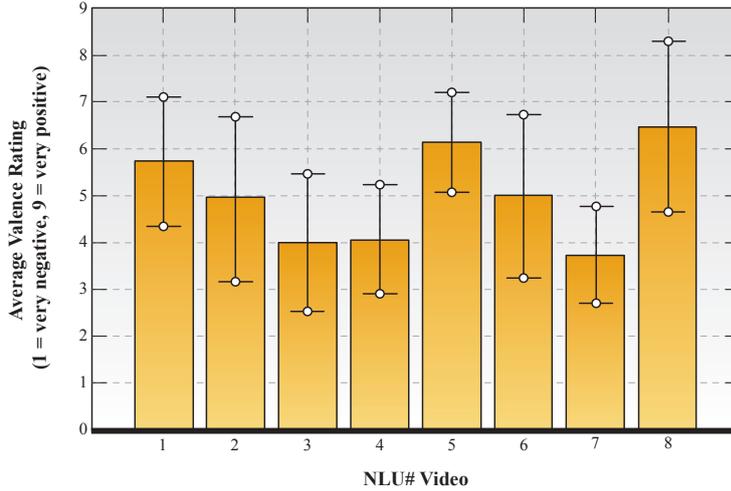


Figure 8.4: Bar graph showing the mean and standard deviations for the ratings of the NLU videos in the pilot study. These ratings are summarised in table F.4.

of responses within the scale. The lowest rating was for NLU#7 (mean = 3.7333, STD = 1.3870) and the highest rating was for NLU#8 (mean = 6.4667, STD = 1.8074). This, as with the results in previous chapters, illustrates the general ambiguity of NLUs when presented without any context. In these results it is also interesting to note that the two utterances taken from the CP study were found to have similar ratings, while in the CP study they were rated as significantly different when rated using the AffectButton.

8.1.1.4 Final Video Production

The five action videos selected were the Slap, Kiss to the Head, Flick to the Face, Stroke to the head and covering of the robot's eyes (see figure 8.5). As well as being guided by the results of the pilot study, this choice was also made as these actions were all applied to the head of the robot. 5 of the 8 NLUs were selected from the pilot study (NLUs #1,2,3,7 and 8) as these most represented a gradient of valence ratings in the pilot study. NLUs #7 and 8 were selected to be used as the reaction NLUs in conditions C_{Action}^N and C_{Action}^P respectively given that they had the two most extreme ratings. The rationale for including 5 NLU videos in the final survey was an attempt to disguise the repetition of NLUs #7 and 8 in the action/NLU combination videos. This repetition may have lead to subjects recognising the NLUs and in turn recalling their affective ratings and introducing

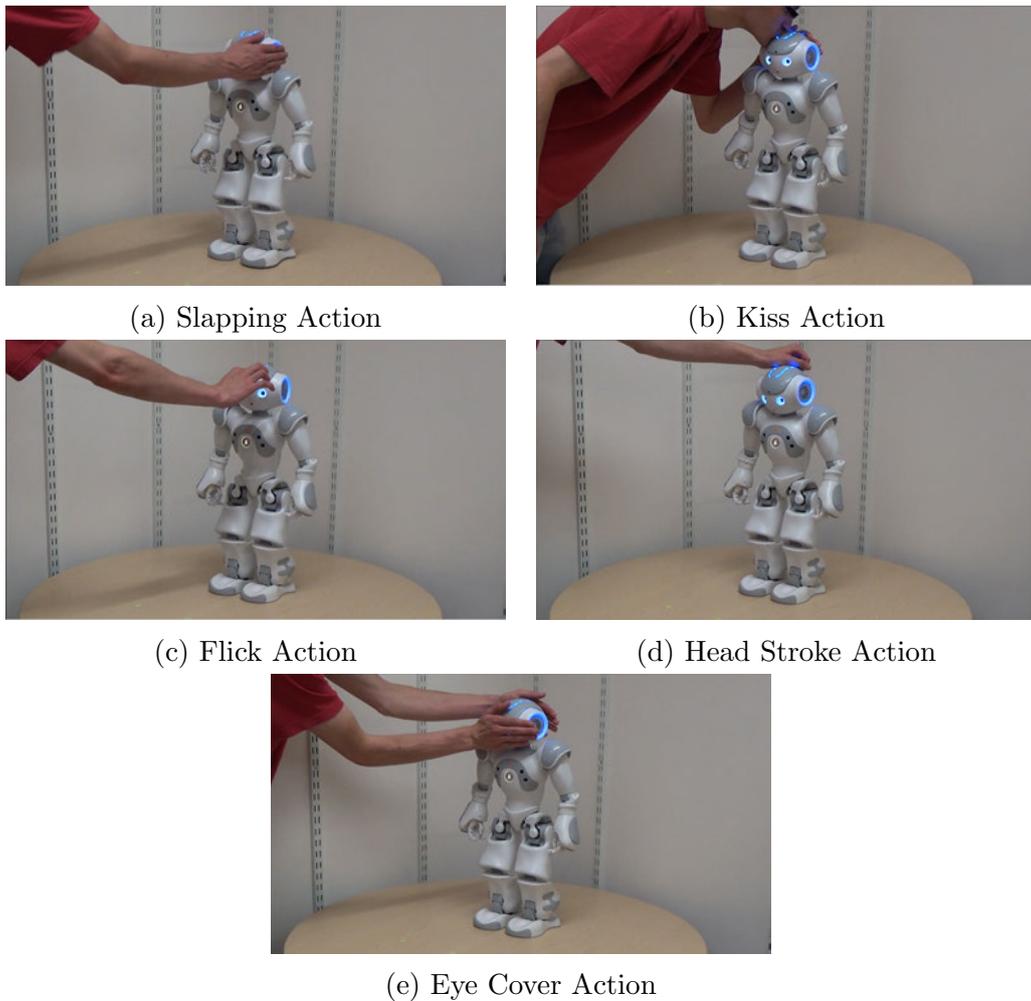


Figure 8.5: Images from the videos showing the five different action scenarios used in the final study, selected via the pilot study.

a bias.

In order to produce the videos for the combined action/NLU video conditions (C_{Action}^P and C_{Action}^N), two videos were made for each selected action, one with NLU#7 being added to the audio track of the video, and one with NLU#8 being added. Thus in total, 10 videos were produced. By producing the videos in this manner, the start of the two NLUs were synchronised exactly, and the visual of the video being identical between the two conditions also, thus ensuring that it was only the NLUs that were different between these two video conditions.

8.1.2 Experimental Procedure

Using crowd sourcing methods to conduct online studies have recently been shown to be a fast and fruitful means to gather information on HRI on a large scale (*e.g.*

Breazeal et al. (2013) and Chernova et al. (2011)), and as such this experiment was conducted using an online crowd sourcing service, *CrowdFlower*⁴. Recruitment of subjects was limited to the USA to provide a limit on and focus to the subject demographic, and subjects were rewarded 0.3\$ USD for their participation.

At the start of the experiment, subjects were asked to provide their age and gender. Each video was then presented with all 20 videos in one of three orders⁵ (it was not possible to completely randomise the order of the questions using the CrowdFlower service). For each video, subjects were asked to provide an affective rating indicating how they thought that the robot *felt* based upon what happened to the robot in the video, or by the sound that the robot made. Each video also had a *validation* question, which asked a specific question about the content of the video. This was done to confirm whether a subject indeed had watched the video and not provided a random rating. Questions queried details such as the colour of the robot’s eyes, the colour of the human’s t-shirt, whether or not the robot made a sound in the video, and what action happened in the video. All the validation questions were forced choice with either two or three options. Finally, at the end of the experiment, subjects were asked whether they had seen the robot before, to determine whether they were familiar with the Nao platform in some way.

Overall, subjects were asked to rate 5 NLU Videos, 5 Action Videos, 10 Action/NLU combination videos, 2 personal questions, and the question regarding their familiarity with the robot. Thus, there were 23 questions in the survey in total. The average time taken to complete the experiment was no more than 6 minutes.

8.2 Results

In total, 324 people responded via the CrowdFlower service, however, the data for 21 subjects was omitted as their accuracy on the validation questions (mean

⁴CrowdFlower can be accessed from www.crowdfLOWER.com.

⁵Orders are presented in table F.2

= 93.96%, std = 11.02%) fell below 80%. 303 Subjects were thus used in this analysis. Of these subjects, 87 were male (mean age = 32.24, std = 10.27) and 216 were female (mean age = 37.15, std = 11.39). 99 Subjects answered the first order of questions, 101 subjects for the second order and 103 for the third order. 151 subjects reported that they had seen the robot before (53 males, mean age = 32.24, std = 11.12, and 98 females, mean age = 37.71, std = 11.12).

Cronbach's α was used as a measure of internal agreement between subjects. For the videos showing only the NLU conditions (C_{NLU}^P and C_{NLU}^N), the α value for the ratings was 0.973, and for the videos showing only the action (C_{Action}) the α value for the ratings was 0.997. The ratings for the videos showing the Action/NLU combinations (C_{Action}^P and C_{Action}^N) both had α values of 0.998. Collapsing all the ratings together an α value of 0.9973 was obtained. All these α values are high, indicating a strong level of internal agreement between all subjects across all of the conditions.

In order to access the interpretation of the NLUs, the ratings for the NLU videos (C_{NLU}^P and C_{NLU}^N) were subject to a 3-way repeated measures ANOVA (5x2x2), where the factors were the video shown (within subjects), subject gender (between subjects) and robot familiarity (between subjects). The Action Videos were subject to the same ANOVA design.

To test the impact of the Action/NLU combinations (C_{Action}^P and C_{Action}^N) of the five different actions, the same 3-way repeated measures ANOVA (5x2x2) design was used. The video factor consisted of the videos showing each of the five different conditions. The two other factors were again subject gender and robot familiarity. All results were also subject to post-hoc multi comparison tests using the Scheffé method with Bonferroni corrections used to identify the relative ratings between the various factors and video conditions.

The remainder of this section presents the results of the analysis for the NLU Videos alone (section 8.2.1), the Action Videos alone (section 8.2.2), and then the analysis of each of the five video conditions for each of the five different action scenarios (sections 8.2.1 to 8.2.2).

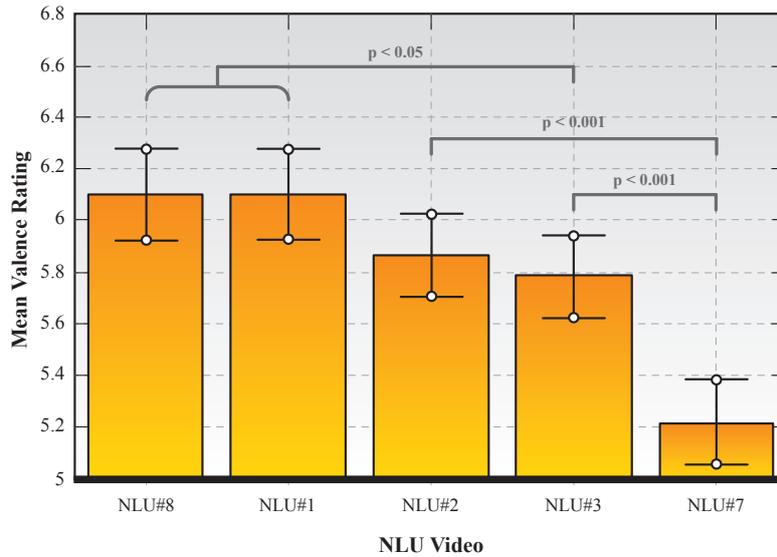


Figure 8.6: Bar graph showing the Mean values and 95% Confidence Interval for the valence ratings for each of the 5 NLU videos. These ratings are summarised in table F.5.

8.2.1 NLU Videos

The ANOVA identified a significant main effect due to the video shown ($F(4, 1196) = 32.024$, $MSE = 31.206$, $p \leq 0.001$). No main effects were found due to the subject gender or the robot familiarity, and no interaction effects were found.

The post-hoc tests showed that NLU#7 (mean = 5.129, 95% CI = [5.053 5.385]) was rated significantly lower than all the other NLUs ($p \leq 0.001$). NLU#1 (mean = 6.100, 95% CI = [5.923 6.277]) and NLU#8 (mean = 6.102, 95% CI = [5.926 6.278]) jointly had the highest rating⁶ and received significantly higher rating than NLU#3 (mean = 5.781, 95% CI = [5.621 5.942]), $p \leq 0.05$. NLU#2 (mean = 5.866, 95% CI = [5.703 6.029]) was found to only have a statistically higher ($p \leq 0.001$) rating than NLU#8. These results are shown in figure 8.6, and show that indeed NLUs#7 and 8 represented the extremes with respect to the affective ratings.

⁶Thus there was no significant difference between these two NLUs.

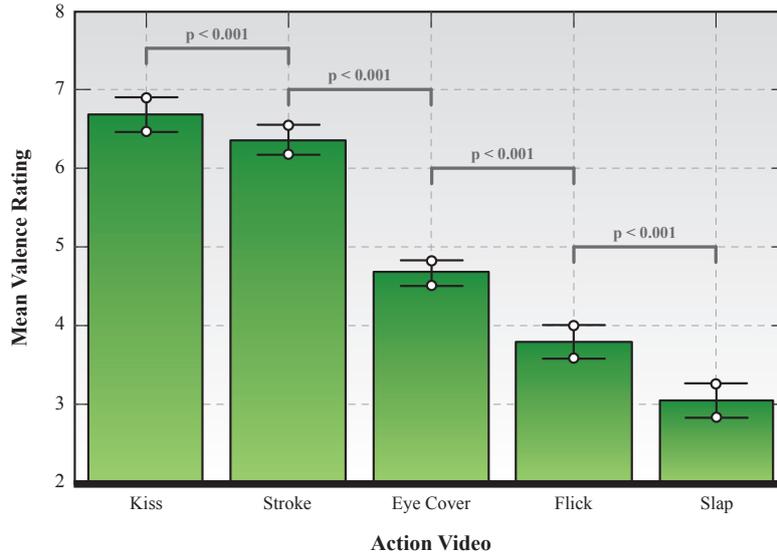
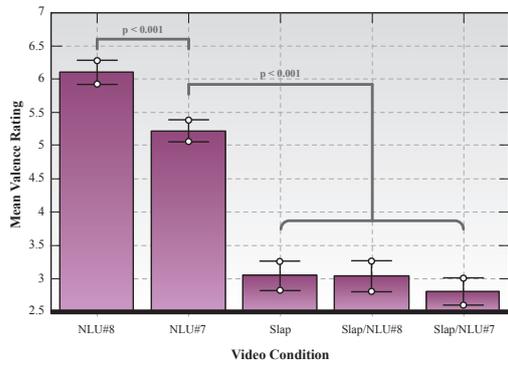


Figure 8.7: Bar graph showing the Mean values and 95% Confidence Interval for the valence ratings for each of the 5 Action videos. These ratings are summarised in table F.6.

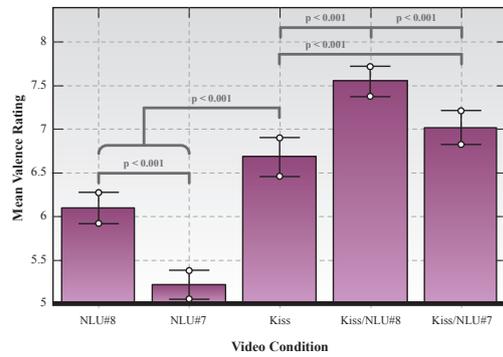
8.2.2 Action Only Videos

For the Action Videos a significant main effect was found due to the action that was shown in the video ($F(4, 1196) = 251.833$, $MSE = 601.029$, $p \leq 0.001$), with no other main effects or interaction effects found.

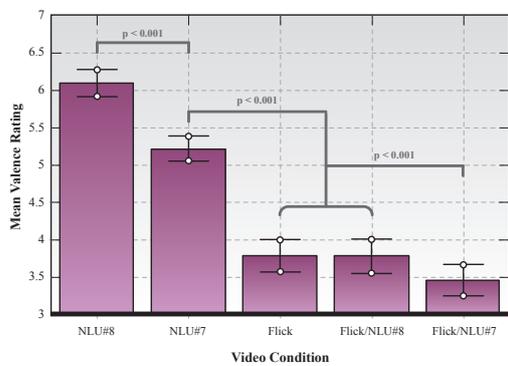
The post-hoc tests revealed that the videos indeed represented a gradient of affective interpretations across the 5 different actions, and that the ratings for all actions were significantly different ($p \leq 0.001$). The video showing the robot being slapped received the lowest rating (mean = 3.045, 95% CI = [2.828 3.262]), followed by the flicking action (mean = 3.787, 95% CI = [3.578 3.995]). The covering of the robot's eyes action (mean = 4.669, 95% CI = [4.517 4.822]) represented the middle action, being interpreted as relatively neutral. The video of the robot being stroked on the head received the second highest rating (mean = 6.361, 95% CI = [6.172 6.549]), while the video showing the robot being kissed on the head received the highest rating (mean = 6.687, 95% CI = [6.469 6.905]). These results are shown in figure 8.7.



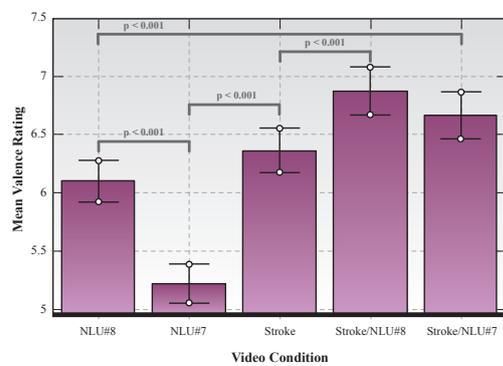
(a) Slapping Action



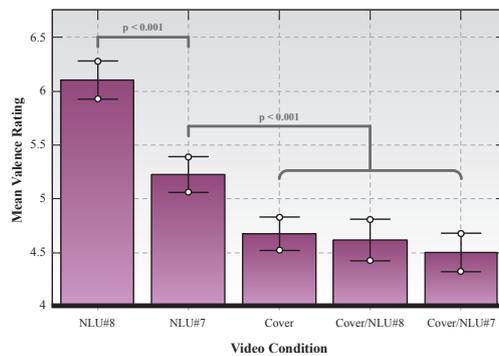
(b) Kissing Action



(c) Flicking Action

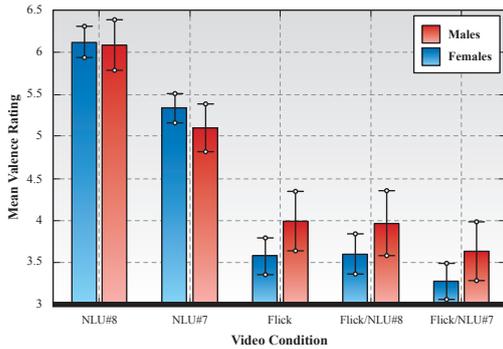


(d) Stroking Action

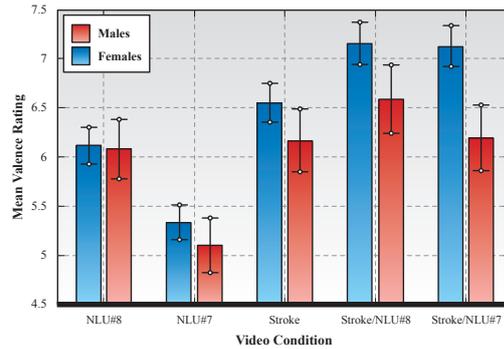


(e) Cover Action

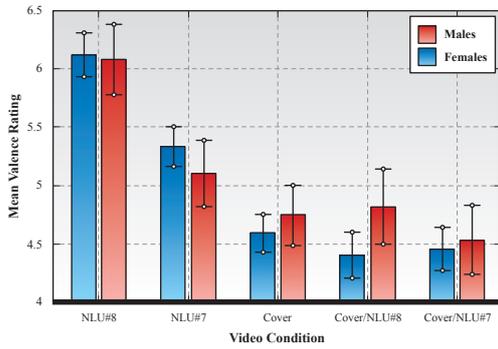
Figure 8.8: Results of the 3-way repeated measures 5x2x2 ANOVA showing the Mean ratings for the videos as well as the 95% Confidence Interval for each of the 5 video conditions. across the 5 action scenarios. Primary statistically significant differences are shown, and the other significant differences may be inferred from those already displayed. These ratings are summarized in table F.7



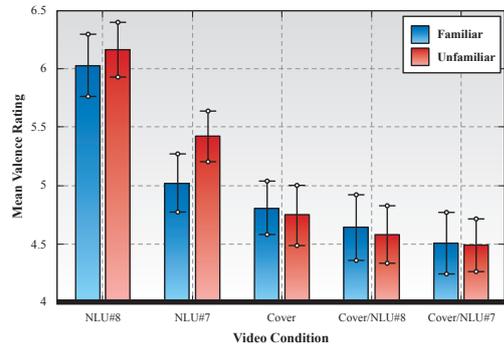
(a) Flick Action: Plot of the interaction between the subject gender and the video shown (see table F.8).



(b) Stroke Action: Plot of the interaction between subject gender and the video shown (see table F.9).



(c) Cover Action: Plot of the interaction between subject gender and the video shown (see table F.10).



(d) Cover Action: Plot of the interaction between the robot familiarity and the video shown (see table F.11).

Figure 8.9: Plots of the interaction effects identified through the 3-way ANOVAs. Each plot shows the Mean values and 95% CI for the ratings across each of the five video conditions, with each line either representing the subject gender or robot familiarity factors.

8.2.3 Action/NLU combination Videos

This section presents the results of the 5x2x2 ANOVA analysis for each of the five action scenarios individually. Referring to figure 8.8, the C_{NLU}^P and C_{NLU}^N conditions are present in all of the scenarios, and thus these two conditions have the same mean ratings and 95% confidence intervals throughout while the ratings for conditions C_{Action} , C_{Action}^P and C_{Action}^N for each action scenario are different.

8.2.3.1 Slap Action Scenario

For the slapping action, only a main effect due to the video shown was found ($F(4, 1196) = 301.774$, $MSE = 546.218$, $p \leq 0.001$). No other main effects or interaction effects were found.

The post-hoc tests showed that the videos showing condition C_{NLU}^P (mean = 6.102, 95% CI = [5.926 6.278]) and C_{NLU}^N (mean = 5.129, 95% CI = [5.053 5.385]) had a significantly different rating ($p \leq 0.001$), and that the videos showing condition C_{Action} (mean = 3.045, 95% CI = [2.828 3.262]), C_{Action}^P (mean = 3.042, 95% CI = [2.810 3.275]) and C_{Action}^N (mean = 2.805, 95% CI = [2.604 3.007]) had no significant differences in their ratings while they were all rated significantly lower than the video showing C_{NLU}^N , $p \leq 0.001$ (figure 8.8a).

8.2.3.2 Kiss Action Scenario

For the kissing action, the ANOVA showed that there was again only a main effect due to the video condition ($F(4, 1196) = 139.445$, $MSE = 191.377$, $p \leq 0.001$) and no interaction effects.

The post-hoc tests revealed that all of the videos had significantly different ratings ($p \leq 0.001$). The video showing the C_{Action}^P condition received the highest rating (mean = 7.550, 95% CI = [7.378 7.721]), followed by C_{Action}^N (mean = 7.017, 95% CI = [6.822 7.212]). The C_{Action} condition (mean = 6.687, 95% CI = [6.469 6.905]) was rated higher than C_{NLU}^P (mean = 6.102, 95% CI = [5.926 6.278]) while the C_{NLU}^N condition received the lowest rating (mean = 5.129, 95% CI = [5.053 5.385]). These results are shown in figure 8.8b.

8.2.3.3 Flicking Action Scenario

It was found that, in the flicking scenario, there was a significant main effect due to the video condition that was shown ($F(4, 1196) = 184.597$, $MSE = 309.671$, $p \leq 0.001$), as well as an interaction effect between the subject gender and the video condition ($F(4, 1196) = 3.021$, $MSE = 5.069$, $p \leq 0.05$).

With respect to the main effect due to the video shown, the post-hoc tests indicated that the videos showing the action (C_{Action}) and action/NLU combinations (C_{Action}^P and C_{Action}^N) were rated as significantly lower than the video showing C_{NLU}^N ($p \leq 0.001$). There was no difference in the rating between C_{Action} (mean = 3.787, 95% CI = [3.578 3.995]) and the C_{Action}^P (mean = 3.784, 95% CI = [3.556 4.012]). The C_{Action}^N video (mean = 3.456, 95% CI = [3.251 3.661]) was rated significantly lower than all other videos ($p \leq 0.001$). These results are shown in figure 8.8c.

For the interaction effect, the means and 95% confidence intervals indicate that for the two NLU videos, females provided marginally higher ratings, while for the flicking action and combination videos the females provided lower ratings. However, post-hoc independent samples t-tests found that there were no significant differences between the genders for either C_{NLU}^P ($t(301) = 0.109$, $p = 0.914$), C_{NLU}^N ($t(301) = 1.021$, $p = 0.308$), C_{Action} ($t(301) = -1.767$, $p = 0.078$), C_{Action}^P ($t(301) = -1.553$, $p = 0.121$), or C_{Action}^N ($t(301) = 1.772$, $p = .077$). These results are shown in figure 8.9a.

8.2.3.4 Stroke Action Scenario

For the stroking action, the ANOVA found that there were main effects due to the video shown ($F(4, 1196) = 66.843$, $MSE = 98.964$, $p \leq 0.001$) and the subject gender ($F(1, 299) = 11.411$, $MSE = 55.833$, $p \leq 0.001$), as well as an interaction effect between these two factors ($F(4, 1196) = 4.707$, $MSE = 6.968$, $p \leq 0.01$).

With respect to the main effect due to the video condition, the post-hoc tests revealed that the C_{Action}^P had the highest rating and was significantly different ($p \leq 0.001$) to the ratings for C_{NLU}^P (mean = 6.102, 95% CI = [5.926 6.278]) and the video of C_{Action} (mean = 6.361, 95% CI = [6.172 6.549]), while no significant

difference was found between the C_{Action}^P and C_{Action}^N (mean = 6.663, 95% CI = [6.464 6.862]) or the C_{Action} and C_{NLU}^N videos. The C_{NLU}^N video had the lowest rating (mean = 5.129, 95% CI = [5.053 5.385]) and was significantly different from all the other videos ($p \leq 0.001$). This result is shown in figure 8.8d.

The tests also revealed that overall the mean rating for the female subjects (mean = 6.459, 95% CI = [6.326 6.592]) was significantly higher ($p \leq 0.01$) than the ratings provided by the male subjects (mean = 6.027, 95% CI = [5.813 6.241]).

The interaction effect between the videos and subject gender was identified as female subjects providing higher ratings than the males, with the difference between the ratings being larger for the action/NLU combinations than the other videos. Post-hoc independent samples t-tests found that for C_{Action}^P the female (mean = 7.159, 95% CI = [6.944 7.374]) subjects provided significantly higher ratings than the males (mean = 6.588, 95% CI = [6.242 6.933]), $t(301) = 2.801$, $p = 0.005$. Similarly for C_{Action}^N the female (mean = 7.129, 95% CI = [6.920 7.339]) subjects again provided significantly higher mean ratings than the males (mean = 6.196, 95% CI = [5.859 6.534]), $t(310) = 4.789$, $p < 0.001$. These results are shown in figure 8.9b.

8.2.3.5 Eye Cover Action Scenario

It was found that the video condition again had a significant main effect ($F(4, 1196) = 78.989$, $MSE = 105.718$, $p \leq 0.001$), and that there were interaction effects between the subject gender and the video condition ($F(4, 1196) = 2.571$, $MSE = 3.441$, $p \leq 0.05$), and the robot familiarity and the video condition ($F(4, 1196) = 2.871$, $MSE = 3.842$, $p \leq 0.05$).

For the main effect due to the videos condition, the post-hoc tests revealed that C_{Action} video (mean = 4.669, 95% CI = [4.517 4.822]), C_{Action}^P (mean = 4.611, 95% CI = [4.422 4.799]) and C_{Action}^N (mean = 4.497, 95% CI = [4.322 4.672]) videos had no significant difference, but all were rated as significantly lower than the C_{NLU}^N video (mean = 5.129, 95% CI = [5.053 5.385]), $p \leq 0.001$ (see figure 8.8e).

The interaction effects were identified as the female subjects providing higher

ratings than the males for the NLU videos, while providing lower ratings for the action and action/NLU combination videos. The post-hoc t-tests found that there was a significant difference in the ratings between the genders for C_{Action}^P ($t(301) = -2.144, p = 0.033$), with males (mean = 4.819, 95% CI = [4.498 5.139]) providing a higher mean rating than the females (mean = 4.403, 95% CI = [4.204 4.602]).

Similarly, the subjects who were familiar with the robot provided higher ratings for the NLU videos than subjects unfamiliar with the robot, with the opposite being true for the action and action/NLU combination videos. The post-hoc t-tests found that there was only a significant difference for C_{NLU}^N ($t(301) = -2.395, p = 0.017$), with the subjects familiar with the robot (mean = 5.421, 95% CI = [5.202 5.639]) giving higher ratings than the subjects unfamiliar (mean = 5.017, 95% CI = [4.768 5.267]) with the robot. These results are shown in figures 8.9c and 8.9d.

8.2.4 Summary of Results

It was found the five different actions shown in the videos indeed did represent a gradient of different affective ratings along a valence Likert scale. The action where the robot was slapped was rated as being the least positive, followed by the video showing the robot being flicked in the forehead. The middle, or neutral action was the scenario in which the robot's eyes were covered. Subjects rated the video showing the robot being stroked on the head as mildly positive, while the video of the robot being kissed on the head received the highest valence rating, being most positive.

With respect to the videos showing the robot making one of five NLUs with no action, these too were found to represent a gradient of valence, though the overall range of ratings was notably smaller than that of the action videos. The two NLUs that were rated as being least positive and most positive were then used in the action/NLU videos.

In the videos showing the combination of the actions and the two NLUs, it was found that in all cases, subjects rated these as having the same degree of valence

as the particular action alone, or a more extreme valence (i.e. more positive or negative).

8.2.4.1 Slapping Action

For the videos showing the robot being slapped, both the videos in which the robot made the *positive* and *negative* utterances were rated as being equally negative as the video showing only the robot being slapped. The NLU only videos were both found to be rated as more positive than the action only videos and both the NLU/action videos.

8.2.4.2 Kissing Action

The combination videos were found to have slightly more extreme, positive ratings than the video showing the robot just being kissed on the head. It was found that the combination of the positive NLU and the action were rated as more positive than the video of the negative NLU and the action, while the negative NLU and action video were still rated as more positive than the action only video. The positive NLU video received the same rating as the action only video, but was rated as significantly less positive than the NLU/action combination videos. The negative NLU video was found to have the lowest overall rating.

8.2.4.3 Flicking Action

It was found that the positive NLU and action combination were rated as being equally negative, while the video showing the negative NLU and action were rated as more negative. Both the NLU only videos were rated as being more positive than the action only video and the two NLU/action combination videos.

8.2.4.4 Stroking Action

Similarly to the Kissing action scenario, the combination of the positive NLU and the stroking action were rated as more positive than the video showing only the action as well as the combination of the negative NLU and the action. No difference was found in the ratings between the negative NLU and action combination

and the action only video. It was also found that there was a difference between the two genders in how they rated the two action/NLU combination videos, where the female subjects rated these videos as more positive than the male subjects. Again, the positive NLU video received the same rating as the action only video, but was significantly less positive than the NLU/action combination videos. The negative NLU video was found to have the lowest overall rating.

8.2.4.5 Eye Cover Action

Both the NLU/action combinations were found to have the same relatively neutral rating as the video showing only the action, while the videos of only the NLUs were both rated as more positive. As with the flicking action and the slapping action, both the NLU only videos were rated as being more positive than the action only video and the two NLU/action combination videos.

8.3 Discussion

This section provides a discussion of the experiment and results obtained from a variety of perspectives such as regarding the Hypotheses outlined at the beginning of the chapter, the drawbacks of the methodology, limitations regarding the interpretation of the results and the implications of these results with respect to the practical use of NLUs during social HRI.

8.3.1 Main Effects

The high Cronbach's α values indicate that overall the subjects were in agreement in their interpretations of the videos presented, and in their use of the Likert scale when interpreting and rating the videos. As a result, this promotes confidence regarding the validity of the results obtained via this measuring scale.

With respect to H_1 (the hypothesis that the interpretation of an NLU overrides that of the action), none of the results provide any support for this hypothesis. No significant effects were found, across any of the actions, where the interpretation of either the C_{Action}^P or C_{Action}^N condition was significantly different to the C_{Action}

condition *and closer* to the rating of either the C_{NLU}^P or C_{NLU}^N conditions. Thus, it can be concluded that the NLUs did not have an effect whereby they pulled the interpretation of an Action/NLU combination toward the base interpretation of the NLU.

With respect to H_2 (the hypothesis that the interpretation of an action overrides that of an NLU), all the results provide clear evidence supporting this hypothesis. In all of the action scenarios, the videos showing the action/NLU combinations (C_{Action}^P and C_{Action}^N) received ratings that were either not significantly different from the action only video (C_{Action}), or they were significantly different and more *extreme* than the action video rating. Coupling this with the finding that in all but one of the action conditions (the stroking action) the action videos were found to be rated significantly different and more extreme than the NLU only videos. It is clear that the actions provided a strong biasing effect that overrode the base interpretation of a given NLU.

These results did also yield a further insight: the results for the kissing action, stroking action and flicking action all indicate that the alignment in valence between the NLU and the action can actually *enhance* a given rating of an action, pushing it to a more *extreme* interpretation (i.e. C_{Action}^P was significantly more extreme than C_{Action} whereas C_{Action}^N was not). While this is a rather intuitive result, given that the NLU provided a reaction to the action, it also demonstrates that while the action provided a dominant bias, the NLU did contain subtle acoustic cues that subject were sensitive to. The identification of this sensitivity is important as it highlights the degree through which NLUs operate⁷.

8.3.2 Interaction Effects

Significant interaction effects were identified between the different video conditions and the subject gender, with this being true for the Stroking, Flicking and Eye Covering action scenarios. In all cases it was found that the females provided more extreme ratings than the males overall, with this being a more prominent effect

⁷This insight provides a justification for the previous experiments in which no context was provided amid concerns regarding the potential confounding of results.

with the Action Videos than the NLU Videos. These effects may be a suggestion that the female subjects are able to empathise more with the robot; something that has been reported in a variety of HRI studies, however with inconsistent results (Rosenthal-von der Pütten et al., 2012).

It is, however, important to consider the magnitude of the interaction effect and the conditions under which it has been found. The general size of effects due to gender and the familiarity with the Nao are all small (≤ 1.0). These differences cover a small range of the overall rating scale, and more importantly, cover a small proportion of the *working* range of responses (i.e. between the highest and lowest mean ratings). Also, these effects have only been found in the scenarios in which the affective ratings are not extreme (i.e. the flicking, stroking and eye covering action scenarios).

8.3.3 Methodological Remarks

While the results have shown support for H_2 , there are a number of drawbacks that need to be highlighted with respect to the methodology of the experiment in this chapter.

The fact that the experiment was conducted online using videos results in somewhat reduced ecological validity. The rationale for the use of videos was that subjects can be presented with exactly the same stimulus, and given the content of the videos, this is something that could not be achieved when presenting the action scenarios in real life. However, it is possible that the results found may have been different if the experiment were to have been conducted within the *real world* as it is likely that a medium of video provides a certain degree of *disconnection* between the subjects and the events shown within the videos. In particular with respect to a subjects' ability to empathise with the robot given that the subjects were *observing* the actions and reactions of the robot, rather than being the *cause* of these. Future work may seek to conduct a similar experiment within a real world setting, with subjects being both observers of the robot's behaviour as well as the cause of the behaviours.

The actions that were selected in this study may be considered as *iconic* and thus hold little ambiguity than other potential physical interactions (for example, someone moving the robot’s arm in a random manner). As a result, the notion that the situation overrides the NLU in the case of the actions presented is unsurprising, and it is a notion that may not be very generalisable to physical interactions with the robot in general. This proposition does not, however, consider how people may interpret random physical interactions with the robot. It may be that subjects (consciously or unconsciously) project meaning into physical interactions with robots that they either observe or engage in. If this is the case, then the underlying insight, that physical interaction appears to bias how NLUs are interpreted, may remain valid.

There is an issue surrounding the repetitiveness of the NLUs presented in the videos. While NLUs #1, 2, and 3 were presented as NLU only videos as a means to increase the number of NLUs subjects were presented with in an effort to disguise the repetition of NLUs #7 and 8 in the Action/NLU videos, it may be that the subjects were still able to recognise NLUs #7 and 8 and recall the ratings that they provided in the NLU only conditions. If this is the case, then this may be a factor influencing the observed effect due to the valence alignment.

Finally, all the NLUs covered a small region of the Likert scale with this region being located around what might be considered as “neutral”. Given this, it is not overly surprising that the actions override the NLU interpretation, however in this light it is surprising that effect of the alignment in valence between the NLU and action occurs. If the NLUs were to represent and evoke more extreme affective ratings, evidence supporting H_1 may emerge, particularly in the case of a misalignment in the valence of the action and NLU. This could result in evidence supporting H_2 being less prominent. However, this assumes that it is indeed possible for NLUs to hold interpretations with a similar magnitude as the Kissing and Slapping action scenarios without the inclusion of context. This is a more general assumption about the utility and communicative capabilities of NLUs that has a variety of complex facets (such as how exposure to NLUs through long-term

social interaction with the robot may impact how subjects in turn perceive and interpret NLUs) and is a subject that will be addressed in the final chapter of this thesis, as the findings of this chapter alone address this only in part.

8.3.4 Practical Use of NLUs

The results of this experiment also lend themselves to providing insights as to how NLUs may be used practically in robotic systems:

What NLU to make: As a general “rule of thumb”, if the context is clear, it is likely that any sound (within reason) is likely to produce an adequate alignment to the desired interpretation. However, the subtle sensitivity to NLUs that has been demonstrated here should not be overlooked. If both the valence of the action and the NLU are known, the affective interpretation of the scenario can be amplified to be *more* extreme (and perhaps robust). Though, caution must be drawn to the use of iconic sounds such as single tones with either a rising or falling frequency modulation as these tend to have clear and robust iconic meanings, and the interaction between iconic actions and iconic sounds has not been studied here.

When to make an NLU: This experiment has used NLUs as a means of animating how a robot reacts to physical actions it is subject to, each with an implied affective valence, rather than a means of replacing natural language interaction (NLI). In a similar light to providing feedback, NLUs that have been used here may potentially be used as reactions/feedback to external events that are encoded in language - NLUs could be used for *back-channelling* also. When a robot experiences salient sensory input, or performs a salient *physical* action, using NLUs will likely enhance how the robot is perceived by a human.

Replacing NLI: Depending on the type of interaction that is desired, NLUs, like *gibberish*, may be used to foster natural language responses (and other behaviours, such as gesturing) from a user, without the need to rely on NLP. As an example of this, Chao and Thomaz (2013) have used gibberish speech to circumvent the need for NLP (and the inherent complexities it adds to a system) while

still evoking natural dialogue from subjects when evaluating their model of turn-taking behaviour. In this light, both NLUs and gibberish might have potential in moving beyond Wizard of Oz (WoZ) approaches in HRI studies to overcome shortcomings in NLP (Riek, 2012), moving closer to fully autonomous social HRI.

8.4 Summary

This chapter has focused upon the influence that a situational content may have upon the affective interpretation of NLUs. The inclusion of context is an important aspect to investigate as situational context is also always inherent within real world social HRI.

Two hypotheses were tested. Firstly, that a subject's affective interpretation of an NLU overrides their affective interpretation of a situational context. Secondly, that the opposite is true; that a subject's affective interpretation of a context overrides that of an NLU. The experiment presented in this chapter tests these hypotheses by presenting subjects with five different video conditions:

- The robot emitting a positive NLU and a negative NLU.
- The robot being subject to an action (e.g. a slap on the head)
- Two videos in which the action is accompanied by either the positive or negative NLU.

The results have provided strong support for the second hypothesis in that, in all cases, that the interpretation of the action and NLU combinations were significantly different to that of the NLUs alone, and at the same time having a similar (and in the majority not significantly different) interpretation to that of the action alone. Furthermore, the results also indicate that while the context provides a more weightily cue for interpretation, subjects are still sensitive to the acoustic cues present within the NLU. It was found that when the interpretation of the NLU and the action were aligned (having the same valence), the overall interpretation was enhanced and more extreme, whilst when the valences were

misaligned there was no statistical difference between the action only and the action/NLU combination.

The apparent dominance of the context over the NLUs has important implications upon the practical use of NLUs, particularly with regard to the specification of the acoustic properties. Provided that the context is clear and unambiguous, it is likely that if a robot just makes an NLU (paying little attention to the details of the utterance) the overall interpretation of the scenario remains unchanged. As such, the robot may potentially make a completely random utterance, with little chance of an adverse outcome. Moreover, if the robot is able to make appraisals of salient sensory input, as well as of its planned, actions, the users' affective interpretations of these will likely be amplified if the valence of the NLU made by the robot is aligned with the users interpretation of the situation.

Chapter 9

Combining NLUs with Natural Language

Summary of the key points:

- As NLUs have a great deal of potential utility during linguistic interaction, and linguistic interaction is a rich source of situational context and mood in an interaction, an online experiment is conducted to assess the compatibility of NLUs and language when used in the same robot.
- Adults were shown videos a robot playing a guessing game with a human, where the robot either used only NLUs, natural language, or a combination of the two.
- People show preference for a robot that uses NLUs in combination with natural language than when a robot only uses NLUs, however, a robot that uses only natural language has the highest overall preference.
- If NLUs are to be used by a robot, this should be done in combination with natural language, rather than as a replacement for natural language.

The study presented in the last chapter found that when NLUs were presented within a clear context, the subjects' affective interpretation of the context overrode or significantly biased their affective interpretation of the NLUs. This in itself already sheds light on the factors that are influential in directing and biasing how people respond NLUs. As a result, guidelines regarding the practical use of NLUs were proposed.

However, while physical interaction with social robots and the surrounding environment provides a rich source for context within an interaction, it is not the only phenomenon that occurs regularly during social HRI - people use natural language to communicate with each other, and readily do the same when interacting with social robots (Breazeal, 2002). This also extends to machines and computers (Nass and Brave, 2005; Reeves and Nass, 1996). Thus, natural language is an important aspect to consider as this too is a rich source of situational context, mood and subject matters within an interaction. All of which are facets that will likely impact how NLUs are perceived and interpreted.

The field of HRI is now striving to understand how social robots engage in interactions with people over longer periods, with active efforts directed at developing new systems and technologies that facilitate this increasing in number (see current research projects such as ALIZ-E (Belpaeme et al., 2012) and LIREC (Leite et al., 2013a)). Initial outcomes of these efforts are some clear guidelines for designing these long-term capable systems, with competency in the use of Natural Language as a key capability required (Belpaeme et al., 2012). This raises issues that relate to NLUs: if utterances are indeed to be a useful means of communication and expression in complex and competent social robots, it is likely that their use will need to be integrated alongside natural language. The NLUs will *support*, rather than a *replace* natural language, as the latter is most likely to hamper the development of a long-term interaction bond. In this respect utterances can have, for example, a supportive role by providing backchannel feedback during spoken dialogue (Yngve, 1970) or be used to make expressive and affective displays. However, these potentials are founded on the assumption that natural language and

NLUs are indeed compatible and can be used together without having an adverse impact upon HRI. Furthermore, it must be recognised that natural language and NLUs both operate via the acoustic modality and thus are essentially in direct competition for the same resource when used alongside each other by the same agent. This competition may be the cause of adverse effects and thus should also be investigated.

In this light, the content of this chapter is focused upon gaining initial insights as to whether the modality of natural language and NLUs are fundamentally compatible with each other if used by the same agent. This will allow a first assessment of whether the theoretical potential of NLUs and language also has a practical foundation. More specifically, this chapter presents the results of a video based online experiment in which subjects were presented with four different videos, each showing the robot playing a game and using either natural language, NLUs, or a combination in reaction to events that occur as the game unfolds. Subjects were then asked to rate the vocal utterances made by the robot in the video with respect to their *appropriateness*, *expressiveness* and *naturalness* and how much they *liked* the robot, as well as providing a *preference* rating for the robot in each of the videos.

9.1 Experimental Setup

The experiment set out to test the following hypotheses:

- H_1 : A robot that uses only natural language will be rated as more appropriate/expressive/natural/preferable than a robot that uses NLUs alongside natural language.
- H_2 : A robot using only NLUs will be rated as less preferable/natural/appropriate than a robot that uses NLUs alongside natural language.
- H_3 : The ratings of appropriateness/expressiveness/naturalness/preference will be influenced by how NLUs and natural language are combined.

H_1 provides a clear hypothesis that a robot that solely uses natural language will have the highest overall ratings, while H_2 hypothesizes that a robot that uses only NLUs will have the lowest overall ratings across the measures. The rationale for these hypotheses is that natural language is an ultimate goal in HRI as it provides (in part) a truly social and natural interface for humans with which we have a vast amount of experience to draw upon in order to decode utterances. In comparison to this, NLUs may be considered as an unfamiliar modality of expression in that it is likely that the majority of people have little real-world experience to draw upon and relate to in order to *decode* utterances. H_3 states that how the two modalities are combined will impact the subjects' perception of the robot. This hints at the notion that some combinations of NLUs and natural language may be “*correct*” or “*better*” in some way than others, while not seeking to provide a specification of this, as this is likely to be highly context dependent.

To test these hypotheses, four videos were created, each showing the robot playing a game with a human, with the type of utterance (natural language, NLUs or combination of the two) being varied across the four videos. The game was based upon the “Cups and Balls” game, also known as the “Three Cups and Ball Routine”. In this game, an object (in this case, a small blue furry ball, see figure 9.3) is placed under one of three cups, and the cups shuffled. The objective of the game is for the observer (in this case, the robot) to guess under which of the cups the ball is hidden after they have been shuffled. This scenario was chosen as it facilitates the robot making a variety of different vocalisations throughout the scenario. These vocalisations range from linguistic comments regarding what is happening (such as it recognising the game), conversational fillers (to show that the robot is thinking about the guess it will make), and reactive and expressive vocalisations (used when the robot's guess is revealed to be incorrect). For the purpose of this experiment the unfolding of this game was scripted such that the robot's physical behaviour was always the same, that the robot always made the same (incorrect) guess and that the end location of the ball was always the same. Thus, only the utterances were the controlled dependent variable.

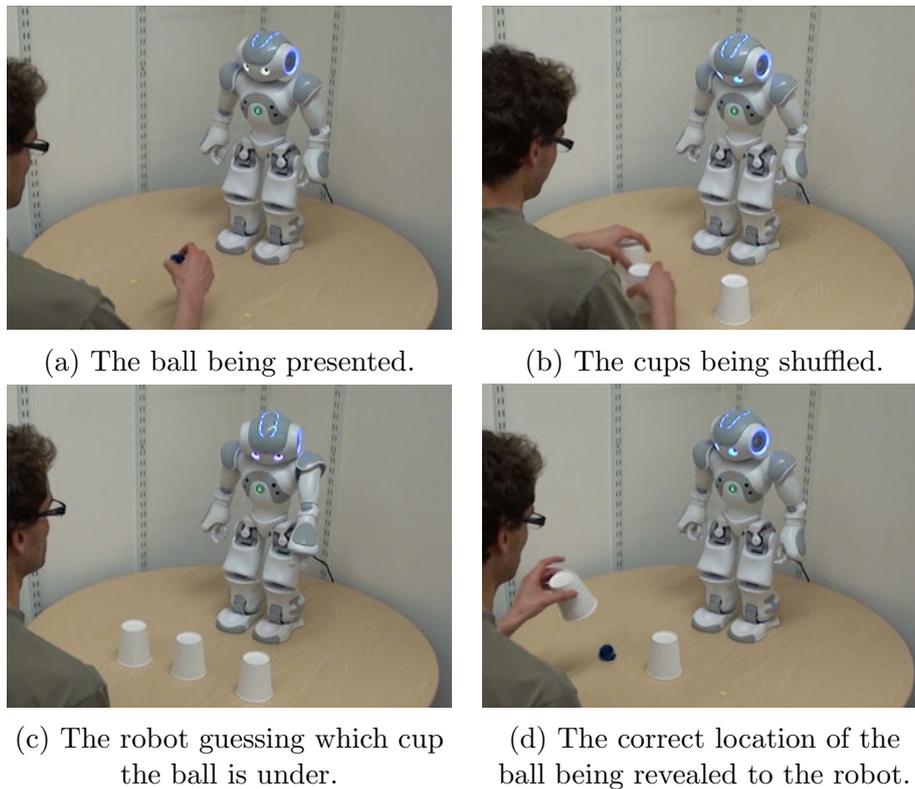


Figure 9.1: Frames from the videos depicting each of the four components to the cups and balls game.

To help formalise the conditions, the scenario may be broken down into the following four main *components* or *events* (see figure 9.1):

1. The human presents the ball to the robot, and then places the cups on the table with the ball being placed under the middle cup.
2. The cups are then shuffled by the human (approx. 10 seconds).
3. The robot behaves in such a manner as to show that it is thinking about which cup the ball is under. It makes a guess and indicates the guess to the human by pointing at the cup (on the right).
4. The guess made by the robot is revealed (to be incorrect) and then the actual location of the ball is revealed (under the left cup).

During each of these components, the robot made a vocalisation, or multiple vocalisations in order to provide feedback to the human either showing the robot's awareness and understanding of the situational context at the given time, or as a means of showing a reaction to what is happening within the context. These

vocalisations also have the benefit of helping animate the whole scenario and bring it more to “*life*”.

The four video conditions were as follows:

- V_1 : The robot only making natural language utterances.
- V_2 : The robot making only NLUs.
- V_3 : The robot making a combination of natural language and NLUs.
- V_4 : The robot making an *inversed* combination of natural language and NLUs.

Videos V_1 (natural language only condition) and V_2 (NLU only condition) serve as opposite control conditions by which conditions V_3 and V_4 (the language/NLU combined conditions) may be measured and compared against. Table 9.1 shows a breakdown of the language/NLU usage across the four conditions. The table shows that V_3 had 7 language samples and 5 NLU samples overall, while V_4 had 8 language samples and 4 NLU samples, with, in the majority, the NLUs and language samples being swapped over between the two conditions. The ratio of NLU to language utterances in V_3 was dictated by selecting utterances from the language only condition that could be replaced with NLUs while still preserving the coherence in the overall vocal behaviour of the robot throughout the game. As this study was not focused on varying the *ratio* (i.e. the frequency of NLUs vs that of natural language) of NLU to language utterances, this was not explicitly controlled. Certain language samples were kept constant between the two conditions however. These were language sample 1 which informed the viewer that the robot was indeed aware of how the game was played, and language samples 5 and 6 which provided an indication that the robot was thinking about which cup to select. All other NLU/language samples were inverted across video conditions V_3 and V_4 .

The remainder of this section outlines in more detail the process undertaken to produce the videos shown to subjects, the audio samples used and an overview of the experimental procedure that was undertaken.

Table 9.1: Breakdown of Language and NLUs used across each of the four video conditions. Note the inverses use of NLUs and language in conditions 3 and 4.

Component	Utterance #	Video Condition			
		1	2	3	4
1	1	Lang-1	NLU-1	Lang-1	Lang-1
2	2	Lang-2	NLU-2	Lang-2	NLU-2
	3	Lang-3	NLU-3	NLU-3	Lang-3
3	4	Lang-4	NLU-4	Lang-4	NLU-4
	5	Lang-5	NLU-5	Lang-5	Lang-5
	6	Lang-6	NLU-6	Lang-6	Lang-6
	7	Lang-7	NLU-7	NLU-7	Lang-7
	8	Lang-8	NLU-8	Lang-8	NLU-8
	9	Lang-9	NLU-9	NLU-9	Lang-9
4	10	Lang-10	NLU-10	NLU-10	Lang-10
	11	Lang-11	NLU-11	NLU-11	Lang-11
	12	Lang-12	NLU-12	Lang-11	NLU-12
Language Total		12	0	7	8
NLU Total		0	12	5	4

Table 9.2: The text input to the TTS engine to produce the language samples used during the videos.

Language Sample #	Speech
1	“Ah, I know this game.”
2	“Where’s it going?”
3	“Whoa, slow down!”
4	“Now then, where did it go?”
5	“Let me see”
6	“I think it’s...”
7	“It’s... It’s...”
8	“That one!”
9	“Am I right?”
10	“Drats!”
11	“Oh that’s a shame”
12	“I could have sworn that I was right!”

Table 9.3: Specification of the Generation Parameters used to generate each NLU.

NLU #	NLU Parameters								
	Base Freq	Freq Range	Speech Rate	Pause Rat.	Rhythm	Unit Count	Tremolo	Skew Rat.	Node Rat.
1	792	641	3	0.25	0.701	4	-0.0329	0.694	0.817
2	580	1000	2	1.011	0.05	4	0.0416	0.601	0.007
3	580	1000	2	0.992	0.05	4	-0.0943	0.393	0.866
4	1242	524	2.168	0.077	0.443	3	0.3681	0.562	0.205
5	580	500	1.95	0.15	0.3	3	0.066	0.809	0.110
6	772	1057	3.279	0.269	0.080	3	-0.2805	0.524	0.720
7	650	1000	3	0.15	0.3	1	0.0699	0.084	0.120
8	650	1000	3	0.15	0.3	2	0	0.937	0.168
9	650	1000	3.5	0.15	0.3	3	0	0.957	0.017
10	600	511	5.29	0.079	0.825	2	0.3847	0.655	0.334
11	518	1299	1.569	0.4826	0.340	1	-0.3031	0.405	0.621
12	1114	1171	3.5	0.15	0.3	3	0.0887	0.348	0.539

9.1.1 Stimulus Production

The language samples were pre-recorded by using the Nao's built in Text-To-Speech (TTS) engine and saving the output to a file (see table 9.2 for the specification of what was said - i.e. the input text strings). These audio files were then played back during the execution of the robot's scripted behaviour. Similarly, the NLUs were pre-recorded using the generation algorithm (chapter 3) and also called from the scripted behaviour. While the language monologue was designed to support the events occurring within the unfolding of the game, the NLUs were not intended (or specifically designed) to portray any particular affective state or have any communicative intent. This was done in the light of the findings outlined in the previous chapter (chapter 8) where the situational context was found to provide adequate cues and help direct how an utterance may be interpreted. Table 9.3 outlines the parameter specification for each NLU and figure 9.2 shows the spectrograms for the utterances indicating the pitch contour shapes. Once recorded, all the audio files underwent post-processing by being converted from a stereo to mono track file, as well as being normalised to -1 dB such that they all had roughly the same acoustic volume when played back via the robot.

Each video was recorded individually¹, with the audio being captured both via the video camera, and separately using professional audio equipment (see figure 9.3). These audio recordings were then converted to a mono-track, normalised to -1 dB and subject to noise removal. Using video editing software, the audio captured via the video camera was discarded and replaced with the post-processed recordings captured via the audio equipment to produce a higher quality end product.

At the end of each video either a number or word was displayed for 1500ms on screen. This was used as a means within the online survey to check whether subjects had indeed watched the video and paid attention to the content, through *validation* questions for each video. The total length of each of the four videos

¹As the videos were each recorded individually and separately, they were not completely identical however aside from the vocalisations made by the robot, there were only minor differences in the video due primarily to differences in the human's behaviour, which includes some differences in sounds made by the human interacting with the cups and balls.

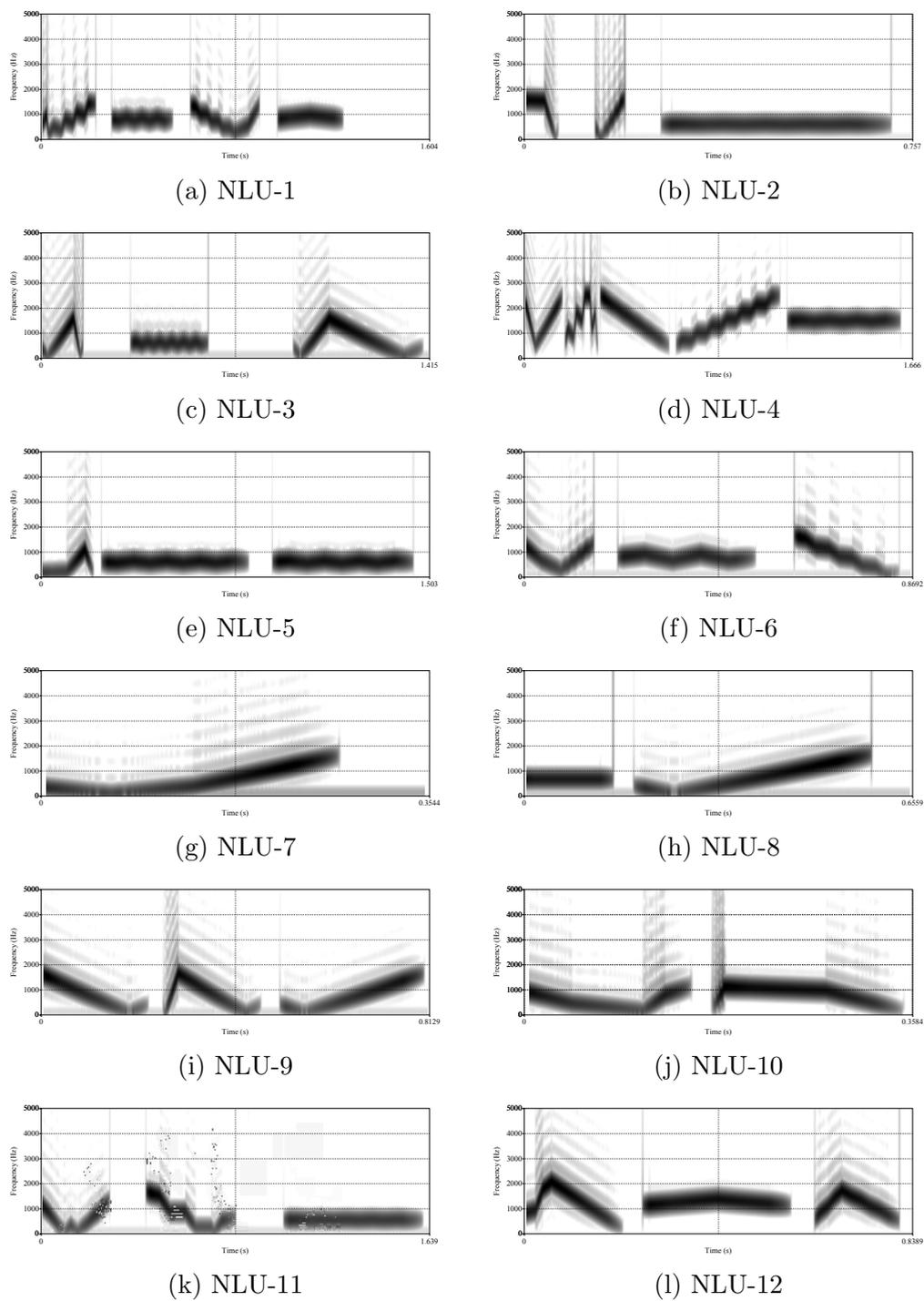
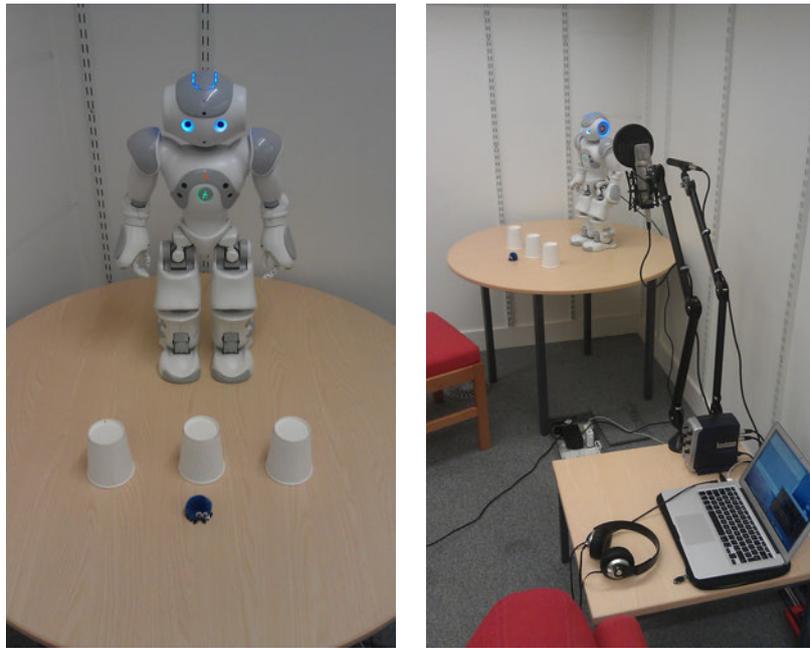


Figure 9.2: Spectrograms of the 12 NLUs used as stimulus.



(a) Image of the Nao, Cups and Ball used in the game.

(b) Image of the professional audio equipment setup.

Figure 9.3: Images of the apparatus used in the stimulus recording.

was 64 seconds.

9.1.2 Experimental Procedure

The online survey was facilitated, and subjects recruited via the *CrowdFlower* crowd sourcing service and were rewarded \$0.3 USD for their participation. Subjects were first asked to provide their age and gender, and then presented with the four videos in random order. After viewing every video, subjects were asked to rate how *appropriate*, *natural* and *expressive* they felt the speech/sounds made by the robot were. They were also asked how much they *liked* the robot in the video as well as answering the validation question confirming whether they indeed had watched the video. This question queried what the text shown (at the end of the video) was. Once all four videos had been presented and the video specific questions answered, subjects were asked to provide a rating of *preference* for each video. Both the *preference* and *liked* measures were included; while they measure similar aspects of the subject’s attitude toward the video, differences may arise due to the ordering of the videos and the fact that the preference ratings were required in the latter stages of the survey, while the likeability ratings were spread

out and queried after each video was viewed.

All ratings were collected using a 9-point Likert scale with 1 representing the most negative aspect of the rating (e.g. for the preference rating, 1 = Least Preferred) and 9 representing the most positive aspect (e.g. for the preference rating, 9 = Most Preferred).

This was followed by 3 more general validation questions, asking details unrelated to focus of the experiment. These questions asked what colour t-shirt the human in the video was wearing, the colour of the object that was covered by the cup and the colour of the robot's eyes. All validation questions were forced choice from a list of four possible answers. Finally subjects were asked whether they had seen the robot before.

9.2 Results

In total, 480 people completed the experiment online, however the data for 210 respondents was not used in the analysis presented in this section as the time taken for them to complete the survey fell below 5 minutes. This threshold was set as each of the four videos was just over one minute, thus it would take approximately 4 minutes 20 seconds to watch all the videos, an additional 40 seconds were added to this to account for subjects also answering all the questions.

Of the remaining 270 subjects whose data was included, 89 were male (mean age = 36.74, std = 10.8) and 181 were female (mean age = 37.74, std = 11.0). 166 subjects reported to have seen the robot before (57 males, mean age = 36.6, std = 10.5, and 109 females, mean age = 39.61, std = 10.15). The average time taken to complete the survey was 8.16 minutes, std = 3.43 minutes, and all validation questions were answered correctly.

Cronbach's α was used as a measure of the agreement between subjects, for each of the 5 different rating scales, across all the video conditions (see table 9.4). High α values were obtained for all the scales with the exception of the preference ratings, which had a very low value indicating little agreement between subjects in their ratings across the four conditions.

Table 9.4: Cronbach’s α ratings for each of the rating scales, across all of the language/NLU conditions.

Rating Scale	α
Appropriateness	0.762
Expressiveness	0.720
Naturalness	0.724
Preference	0.172
Rating	0.841

3-way repeated measures ANOVAs were performed for each of the DVs using the ratings for appropriateness, expressiveness, naturalness, preference and the rating of likability as the multiple dependant variables (DVs), and the different video conditions as the repeated measures. Subject gender and robot familiarity were used as independent, between subjects, variables (IVs). The main effects for the different video conditions were followed up with multi-comparison tests that included a Bonferroni correction². Multi-comparison tests were not performed for the gender or robot familiarity factors as these only had two levels.

When significant three-way interaction effects were found, these were followed up by four 2-way repeated measures ANOVAs, splitting the data with respect to subject gender, using the video condition and robot familiarity as the factors, and then by splitting the data by robot familiarity and using the video conditions and subject gender as the factors. Doing this provides a deeper insight as to the nature of the interactions that have been found. Significant main effects were again subject to the same multi-comparison format outlined above. In the case of significant two-way interactions, post-hoc independent samples test were performed to test for differences between either the two genders, or subjects familiar/unfamiliar with the robot, for each of the four video conditions individually.

The rest of this section presents the results of the ANOVAs for each of the 5 units of measure.

²Bonferroni corrections are applied in order to reduce the chance of finding Type 1 errors (rejecting the null hypothesis when it is in fact true) when comparing the means of multiple groups simultaneously.

9.2.1 Appropriateness Ratings

The Univariate test found that there was a main effect due to the video condition ($F(3, 798) = 49.278$, $MSE = 68.123$, $p < 0.0005$), and a three-way interaction effect between the video condition, subject gender and robot familiarity ($F(3, 798) = 3.790$, $MSE = 5.240$, $p = 0.01$).

Given the presence of the three-way interaction, the main effect was not followed up as it is likely not a true representative of the nature of the relationships between the variables. However, the three-way interaction effect was followed up by four two-way repeated measures ANOVAs, splitting the subjects by the two levels of the robot familiarity variable, using the four video conditions (within subjects) and subject gender (between subjects) as the two factors, and then splitting the subjects by gender and using the video conditions and robot familiarity as the two factors.

9.2.1.1 Splitting Subjects by Robot Familiarity

For the subjects whom were *unfamiliar* with the robot a main effect was found for the video condition ($F(3, 798) = 18.997$, $MSE = 26.254$, $p < 0.0005$) as well as a significant interaction effect between the video condition and subject gender ($F(3, 798) = 2.929$, $MSE = 4.049$, $p = 0.0328$). With respect to the main effect, post-hoc multi-comparison tests revealed that both V_1 and V_2 were significantly different than V_3 and V_4 ($p < 0.01$), with V_1 (mean = 7.358, 95% CI = [7.073 7.642]) receiving the highest ratings and V_2 (mean = 6.052, 95% CI = [5.578 6.526]) receiving the lowest ratings. Video conditions V_3 (mean = 6.757, 95% C = [6.405 7.109]) and V_4 (mean = 6.938, 95% CI = [6.600 7.275]) were not found to be significantly different. Post-hoc independent sample t-tests were used to uncover the nature of the interaction effect between the video conditions and the subject gender. These test found no significant differences between the genders for any of the video conditions.

For the subjects who were *familiar* with the robot, a main effect was found for the video condition ($F(3, 798) = 34.234$, $MSE = 47.308$, $p < 0.0005$) with no sig-

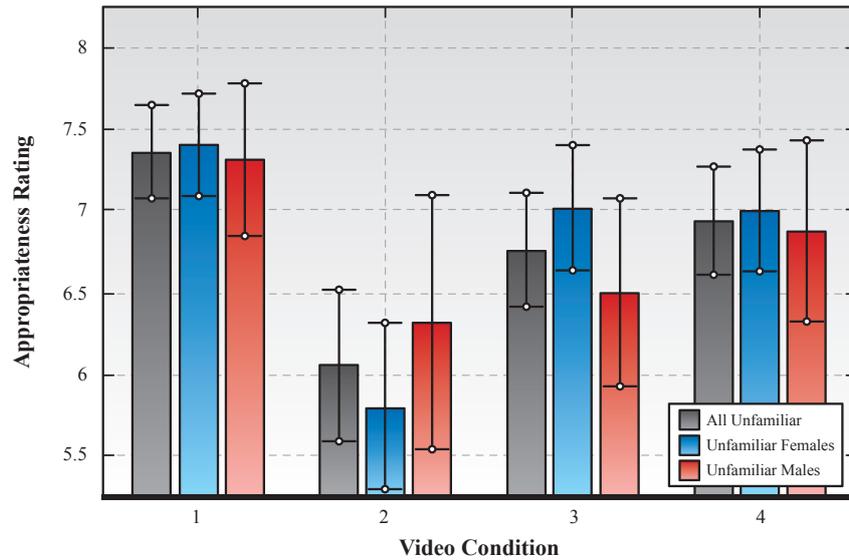
nificant interaction between the video condition and subject gender ($F(3, 798) = 2.384$, $MSE = 3.294$, $p = 0.068$). The post-hoc multi-comparison tests revealed that as with the subject unfamiliar with the robot, V_1 and V_2 had ratings that were significantly different from all the other video conditions ($p < 0.001$). Again, V_1 (mean = 6.160, 95% CI = [5.844 6.476]) received the highest rating while V_2 (mean = 7.515, 95% CI = [7.311 7.719]) received the lowest overall rating. Video conditions V_3 (mean = 6.911, 95% CI = [6.703 7.119]) and V_4 (mean = 7.047, 95% CI = [6.841 7.253]) were not significantly different from each other ($p = 0.431$).

These results are displayed graphical in figure 9.4 and summarised in table G.1.

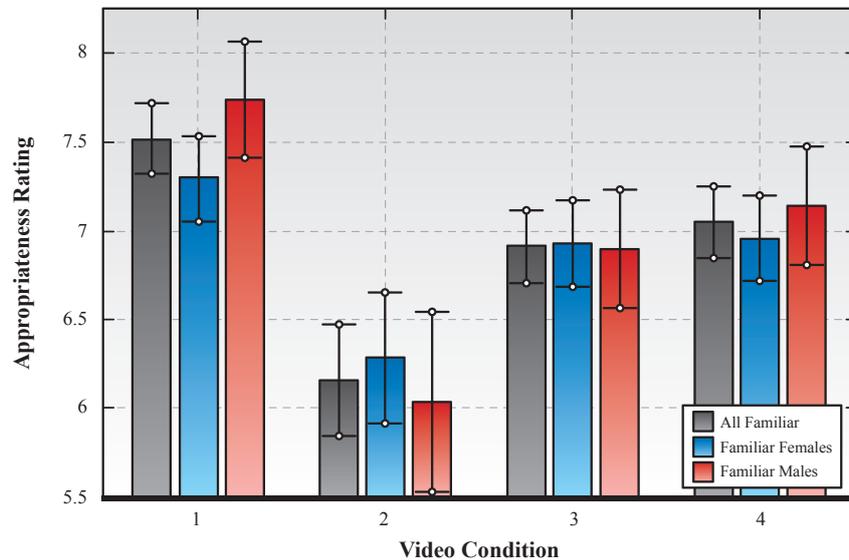
9.2.1.2 Splitting Subjects by Gender

When isolating the *male* subjects, a significant main effect was found for the video conditions ($F(3, 798) = 18.993$, $MSE = 26.249$, $p < 0.0005$), however no interaction effect was found between robot familiarity and the video conditions ($F(3, 798) = 1.583$, $MSE = 2.189$, $p = 0.191$). The post-hoc multi comparison tests revealed that V_1 (mean = 7.525, 95% CI = [7.261 7.789]) received ratings that were significantly higher than all the other video conditions ($p \leq 0.01$), while V_2 (mean = 6.174, 95% CI = [5.746 6.602]) received the lowest overall ratings and was significantly different from all the other video conditions ($p < 0.05$). V_3 and V_4 were not found to be significantly different ($p = 0.157$), with V_4 (mean = 7.008, 95% CI = [6.715 7.300]) receiving and overall higher rating than V_3 (mean = 6.697, 95% CI = [6.355 7.040]).

For the female subjects, a main effect was found for the video condition ($F(3, 798) = 39.195$, $MSE = 54.168$, $p < 0.0005$), as well as an interaction effect between the video condition and robot familiarity ($F(3, 798) = 2.601$, $MSE = 3.595$, $p < 0.0005$). The post-hoc multi-comparison tests found that V_1 (mean = 7.348, 95% CI = [7.147 7.549]) received the highest ratings and was significantly different than all the other video conditions ($p < 0.005$). V_2 (mean = 6.038, 95% CI = [5.718 6.358]) again received the lowest overall ratings and



(a) Subjects unfamiliar with the Nao robot.



(b) Subjects familiar with the Nao robot.

Figure 9.4: Plots showing the mean and 95% confidence intervals for the appropriateness ratings and interaction between male/female subjects and video conditions, with the data split across the robot familiarity factor.

too was significantly different from all the other video conditions ($p < 0.0005$). V_3 (mean = 6.970, 95% CI = [6.726 7.179]) and V_4 (mean = 6.977, 95% CI = [6.760 7.194]) received similar ratings and were not found to be significantly different ($p = 1.00$). With respect to the interaction effect identified, post-hoc independent sample t-tests were performed to check for differences between the female subjects who were familiar/unfamiliar with the robot, for each of the four

video conditions. Each of these tests found no significant differences either video conditions 1 ($t(179) = 0.536, p = 0.592$), 2 ($t(179) = -1.521, p = 0.130$), 3 ($t(179) = 0.413, p = 0.680$) or 4 ($t(179) = 0.209, p = 0.835$).

These results are displayed graphical in figure 9.5 and summarised in table G.2.

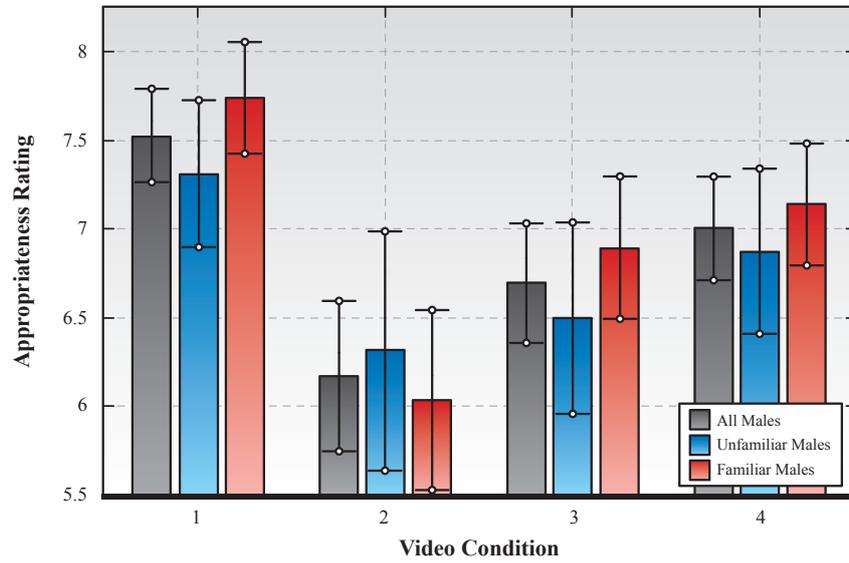
9.2.2 Expressiveness Ratings

The Univariate test found that there was a main effect due to the video condition ($F(3, 798) = 52.119, MSE = 86.818, p < 0.0005$), and a three-way interaction effect between the video condition, subject gender and robot familiarity ($F(3, 798) = 6.056, MSE = 10.088, p < 0.0005$).

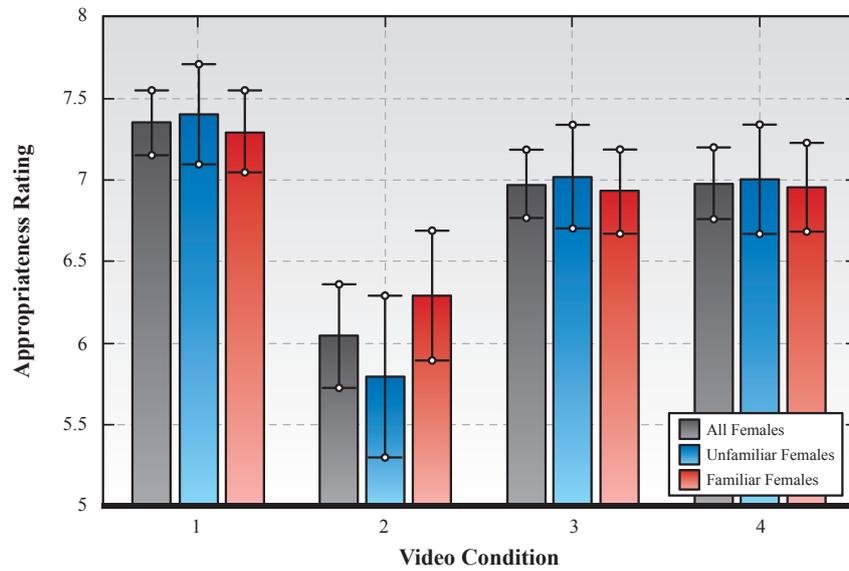
The three-way interaction effect was followed up by four two-way repeated measures ANOVAs, splitting the subjects by the two levels of the robot familiarity, using the four video conditions (within-subjects) and subject gender (between subjects) as the two factors, and then splitting the subjects by gender and using the video conditions and robot familiarity as the two factors.

9.2.2.1 Splitting Subjects by Robot Familiarity

When isolating the subjects *unfamiliar* with the robot, a main effect was found for the video condition ($F(3, 798) = 22.640, MSE = 37.719, p < 0.0005$), as well as an interaction effect between the subject gender and video condition ($F(3, 798) = 5.230, MSE = 8.714, p = 0.001$). Post-hoc multi-comparison tests showed that for the main effect there were no significant differences in the ratings for V_1 (mean = 7.047, 95% CI = [6.738 7.355]), V_3 (mean = 6.688, 95% CI = [6.338 7.037]) and V_4 (mean = 6.960, 95% CI = [6.644 7.277]), ($p > 1.58$), and that V_2 (mean = 5.628, 95% CI = [5.108 6.149]) was rated lower than all the other videos to a statistically significant degree ($p < 0.001$). Regarding the interaction effect identified, the post-hoc independent samples t-tests found that only for V_2 there was a significant difference between the two genders in their ratings of *expressiveness* ($t(102) = -2.132, p = 0.035$).



(a) Male Subjects



(b) Female Subjects

Figure 9.5: Plots showing the mean and 95% confidence intervals for the appropriateness ratings and interaction between subjects familiar and unfamiliar with the robot and the video conditions, with the data split across the subject gender factor.

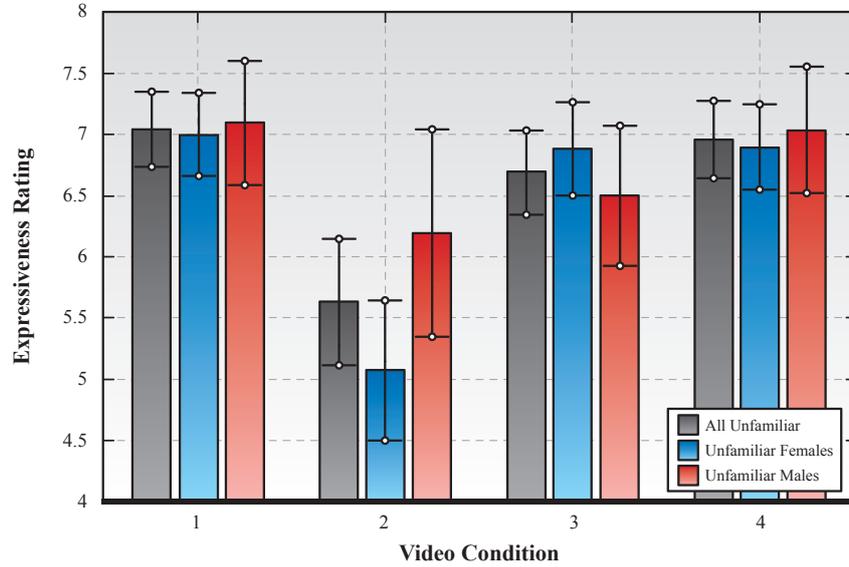
For the subjects *familiar* with the robot, again a main effect was found for the video condition ($F(3, 798) = 32.331$, $MSE = 53.864$, $p < 0.0005$), however no interaction was found between the genders and video condition ($F(3, 798) = 2.300$, $MSE = 3.823$, $p = 0.0759$). The post-hoc tests revealed that for the main effect, V_1 (mean = 7.300, 95% CI = [7.076 7.525]) again received the highest rating and was significantly different to all the other videos ($p < 0.01$), and that V_2 (mean = 5.913, 95% CI = [5.588 6.237]) received the lowest rating and was also significantly different to all the other videos ($p < 0.0005$). V_4 (mean = 6.990, 95% CI = [6.765 7.215]) was rated as more marginally expressive than V_3 (mean = 6.829, 95% CI = [6.673 7.110]), however this difference was not found to be statistically significant ($p = 1.000$).

These results are shown in figure 9.6 and summarized in table G.3.

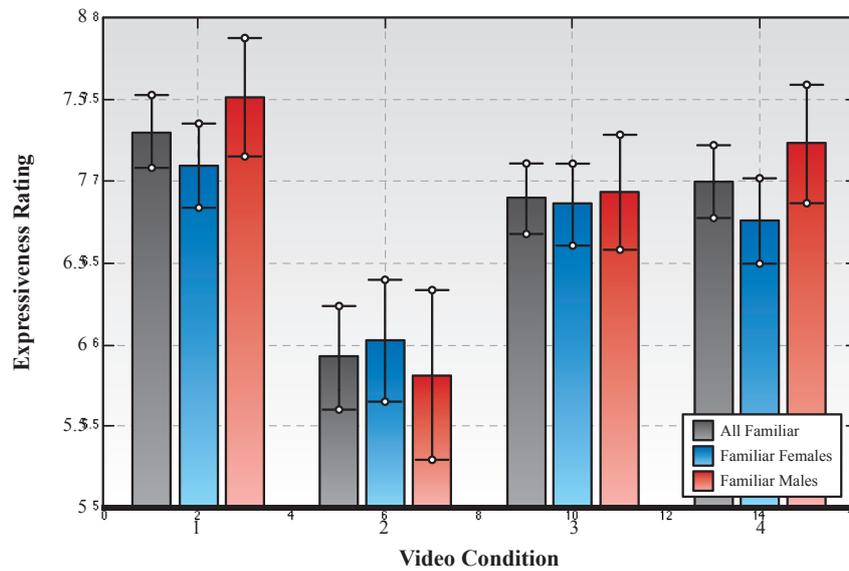
9.2.2.2 Splitting Subjects by Gender

When isolating the *female* subjects, the ANOVA found a main effect for the video condition ($F(3, 798) = 49.575$, $MSE = 82.592$, $p < 0.0005$) and an interaction effect between familiarity with the robot and the video condition ($F(3, 798) = 6.361$, $MSE = 10.599$, $p < 0.0005$). The post-hoc tests show that V_1 (mean = 7.046, 95% CI = [6.828 7.263]) received the highest expressiveness rating, V_2 the lowest rating (mean = 5.544, 95% CI = [5.204 5.883]), and V_3 (mean = 6.864, 95% CI = [6.649 7.079]) and V_4 (mean = 6.821, 95% CI = [6.579 7.044]) received near identical ratings. No significant differences were found between V_1 , V_3 and V_4 ($p > 0.1$), while all three videos were rated significantly higher than V_2 ($p < 0.0005$). For the interaction effect, the post-hoc independent samples t-tests found that there was only a difference between the female subjects familiar/unfamiliar with the robot for V_2 ($t(179) = -2.758$, $p = 0.006$).

For the *male* subjects, a main effect was again found for the video condition ($F(3, 798) = 16.578$, $MSE = 27.260$, $p < 0.0005$), however no interaction effect was found between robot familiarity and the video condition ($F(3, 798) = 1.763$, $MSE = 2.946$, $p = 0.151$). The post-hoc tests revealed that again, V_1 (mean



(a) Subjects unfamiliar with the Robot



(b) Subjects familiar with the Robot

Figure 9.6: Plots showing the mean and 95% confidence intervals for the expressiveness ratings and interaction between male/female subjects and video conditions, with the data split across the robot familiarity factor.

= 7.301, 95% CI = [7.005 7.598]) received the highest rating of expressiveness, V_2 the lowest (mean = 5.997, 95% CI = [5.547 6.447]). V_4 (mean = 7.130, 95% CI = [6.842 7.417]) was rated as more expressive than V_3 (mean = 6.715, 95% CI = [6.374 7.056]), with this also being statistically significant ($p = 0.011$). The difference between V_1 and V_4 was not found to be statistically significant ($p = 0.993$), while the difference in rating between videos 1 and 3 was found to be statistically significant ($p = 0.001$). The rating for video 2 was significantly lower than that of all the other videos ($p < 0.05$). These results are shown in figure 9.7 and summarised in table G.4.

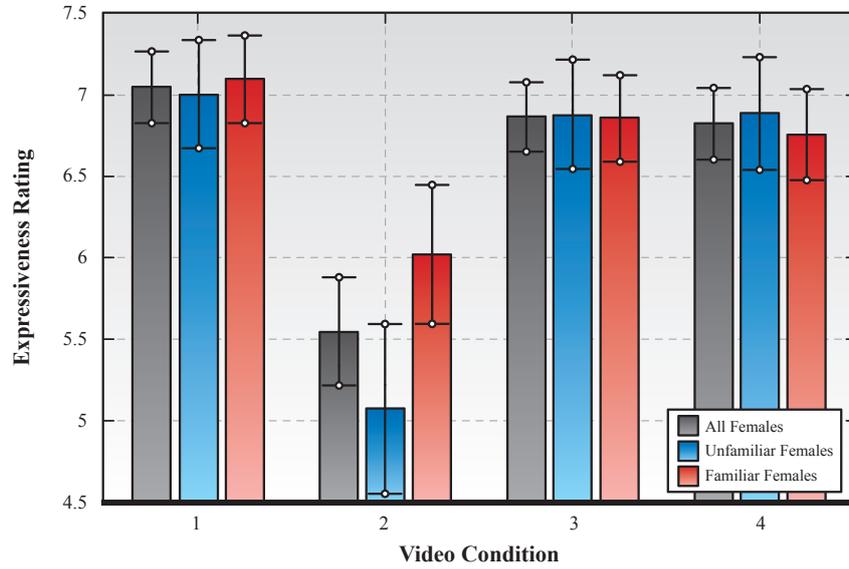
9.2.3 Preference Ratings

The Univariate tests found that there were main effects due to the video condition ($F(3, 798) = 67.4$, $MSE = 145.264$, $p < 0.0005$) and the familiarity with the robot ($F(1, 266) = 6.959$, $MSE = 53.002$, $p = 0.009$)

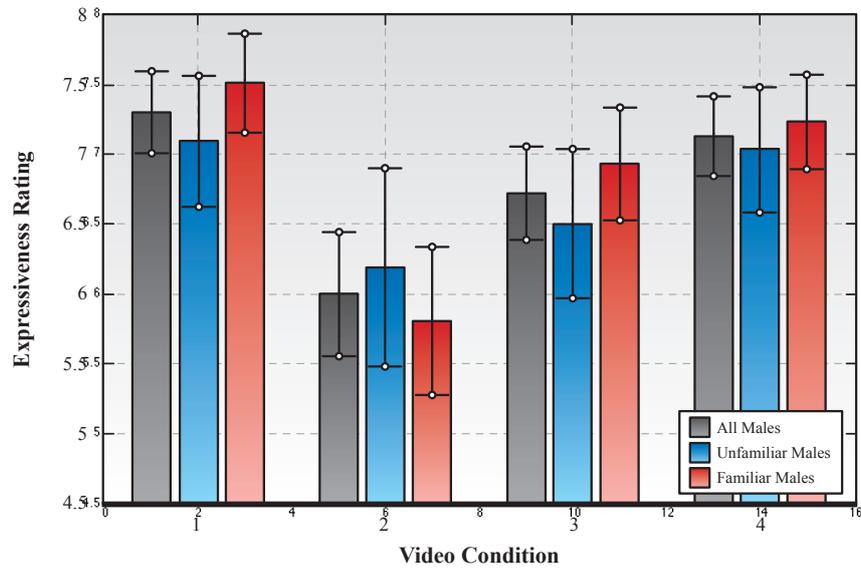
With respect to the video condition main effect, multi-comparison tests showed that V_1 had the highest rating (mean = 6.845, 95% CI = [6.645 7.063]) and was significantly higher than all the other conditions ($p < 0.001$). V_2 was found to have the lowest rating (mean = 4.969, 95% CI = [4.659 5.319]) and was significantly different than all the other conditions ($p < 0.001$). V_4 had the second highest rating (mean = 6.470, 95% CI = [6.256 6.684]) and V_3 the third (mean = 6.260, 95% CI = [6.044 6.475]), with no significant difference found between these two conditions.

With respect to the main effect due to robot familiarity, multi-comparison tests revealed that subjects who had seen the robot before provided higher ratings (mean = 6.387, 95% CI = [6.165 6.609]) than subjects whom had not seen the robot before (mean = 5.899, 95% CI = [5.611 6.188]).

The results for both main effects are shown graphically in figure 9.8 and summarised in table G.5.



(a) Female Subjects



(b) Male Subjects

Figure 9.7: Plots showing the mean and 95% confidence intervals for the expressiveness ratings and interaction between the subjects familiar and unfamiliar with the robot and video conditions, with the data split across the subject gender factor.

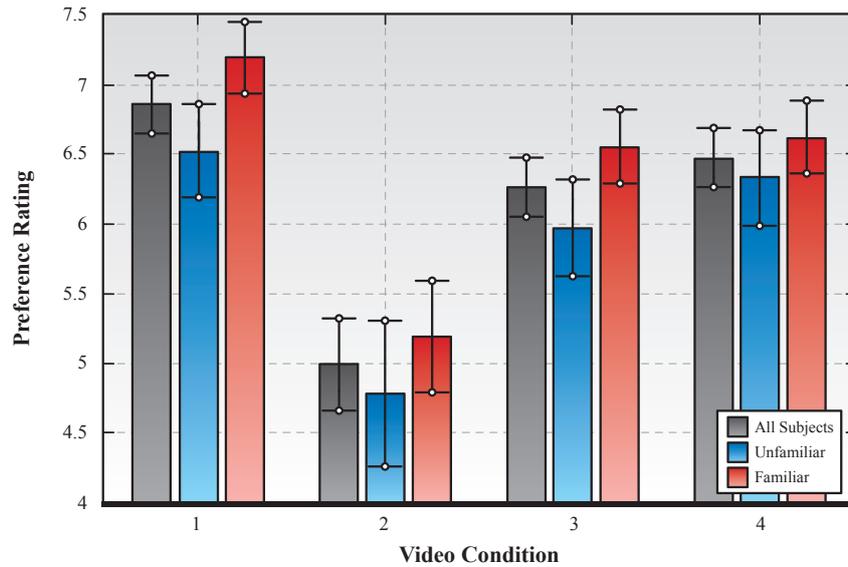


Figure 9.8: Plot of the mean ratings and 95% Confidence Intervals for the Preference ratings for all the subjects combined, and the subjects whom were familiar/unfamiliar with the robot.

9.2.4 Naturalness Ratings

The Univariate tests found a main effect due to the video condition ($F(3, 798) = 56.834$, $MSE = 206.102$, $p < 0.0005$) and subject gender ($F(1, 266) = 10.298$, $MSE = 43.316$, $p = 0.001$), and a two-way interaction effect between the video condition and subject gender ($F(3, 798) = 3.445$, $MSE = 12.494$, $p = 0.016$).

With respect to the main video condition effect, multi-comparison tests revealed that V_1 had the highest rating (mean = 7.165, 95% CI = [6.936 7.393]) and was significantly different to all the other video conditions ($p < 0.05$). Similarly, V_2 was found to have the lowest rating (mean = 4.981, 95% CI = [4.630 5.331]) and was significantly different to all the other video conditions ($p < 0.0005$). V_4 was found to have the second highest rating (mean = 6.781, 95% CI = [6.577 6.985]), while V_3 had the second lowest rating (mean = 6.586, 95% CI = [6.372 6.800]). No significant difference was found between V_3 and V_4 .

With respect to the main effect due to subject gender it was found that male subjects provided a significantly higher mean rating (mean = 6.599, 95% CI = [6.376 6.822]) than the female subjects (mean = 6.158, 95% CI = [6.004 6.311]). With respect to the interaction effect between the video conditions and subject

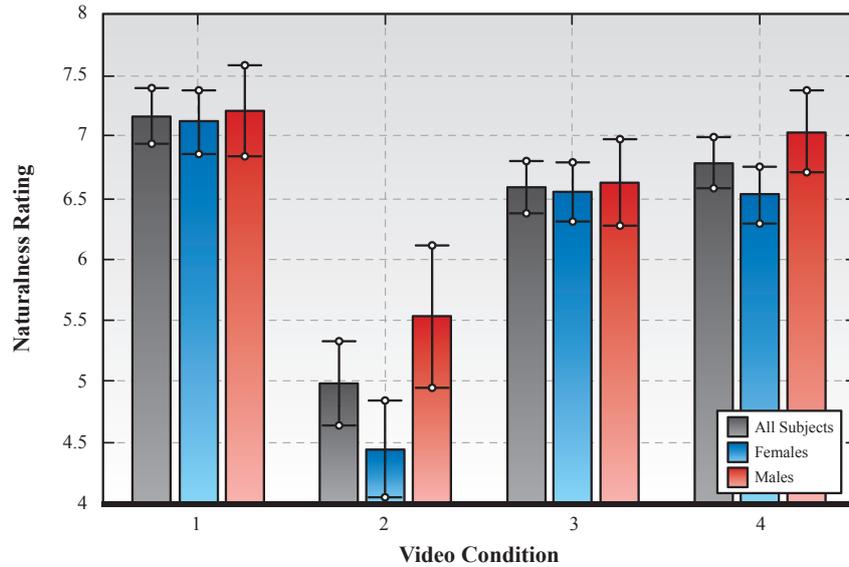


Figure 9.9: Plot of the mean ratings and 95% Confidence Intervals for the Naturalness ratings across the four video conditions, for all the subjects collectively and subject genders.

gender, the post-hoc independent sample t-tests found no significant differences between the two genders across any of the four video conditions.

The results for main and interaction effects are shown graphically in figure 9.9 and summarised in table G.6.

9.2.5 Like-ability Ratings

The Univariate tests found a main effect due to the video condition ($F(3, 798) = 20.396$, $MSE = 28.711$, $p < 0.0005$) and robot familiarity ($F(1, 266) = 7.988$, $MSE = 68.288$, $p = 0.005$), Two-way interaction effects were found between the video condition and subject gender ($F(3, 798) = 3.617$, $MSE = 5.092$, $p = 0.013$) and between subject gender and robot familiarity ($F(1, 266) = 3.948$, $MSE = 33.750$, $p = 0.048$).

The multi-comparison tests found that, for the main effect due to the video condition, that V_1 had the highest rating (mean = 7.305, 95% CI = [7.091 7.519]), while V_2 had the lowest rating (mean = 6.462, 95% CI = [6.176 6.748]). V_1 was found to be statistically different from V_2 and V_3 ($p < 0.05$) but not from V_4 . V_2 was different from all the other conditions to a statistically significant degree ($p < 0.0005$). V_4 had the second highest rating (mean = 7.095, 95% CI

= [6.878 7.311]) and was significantly different from V_2 ($p < 0.0005$). V_3 had the second lowest rating (mean = 7.001, 95% CI = [6.783 7.220]) and was significantly different from V_1 and V_2 ($p < 0.05$).

For the main effect due to robot familiarity the multi-comparison tests revealed that subjects whom had seen the robot before had a mean rating (mean = 7.243, 95% CI = [7.007 7.478]) that was significantly higher than the subject whom had not seen the robot before (mean = 6.689, 95% CI = [6.383 6.995]).

Post-hoc independent samples t-tests were performed to uncover the nature of the interaction effect between the video conditions and subject gender. These tests were performed to compare the ratings of each gender, for each video condition independently (thus four t-tests were performed). For V_1 , no significant differences ($t(268) = -1.534$, $p = 126$) were found between the males (mean = 7.426, 95% CI = [7.073 7.779]) and females (mean = 7.184, 95% CI = [6.942 7.426]). The ratings for V_2 were found to be significantly different ($t(268) = -3.93$, $p = 0.002$), with the ratings for the males (mean = 6.846, 95% CI = [6.375 7.318]) being higher than those for the females (mean = 6.077, 95% CI = [5.753 6.402]). No significant difference was found for V_3 ($t(268) = -0.930$, $p = 0.353$), though the male subjects (mean = 7.031, 95% CI = [6.670 7.391]) had a higher overall rating than the females (mean = 6.972, 95% CI = [6.725 7.220]). The ratings for V_4 were found to be significantly different ($t(268) = -2.412$, $p = 0.017$), with the males (mean = 7.302, 95% CI = [6.945 7.659]) again having a higher overall rating than the females (mean = 6.887, 95% CI = [6.641 7.133]).

With respect to the interaction effect between subject gender and robot familiarity, that post-hoc independent samples t-tests found that the female subjects whom had seen the robot before had a mean rating (mean = 6.862, 95% CI = [6.587 7.138]) that was higher than the female subjects who had not seen the robot before (mean = 6.698, 95% CI = [6.359 7.037]), but this was not significantly different ($t(179) = -0.750$, $p = 0.454$). Similarly, male subjects who had seen the robot before had an significantly higher overall mean rating (mean = 7.623, 95% CI = [7.242 8.004]) than males who had not seen the robot before (mean = 6.680,

95% CI = [6.171 7.189]), $t(87) = -2.849$, $p = 0.005$. The tests also revealed that when isolating the subjects familiar with the robot, the male subjects had a significantly higher overall mean rating than the females ($t(164) = -3.186$, $p = 0.002$), while this was not the case for the subjects who were unfamiliar with the robot ($t(102) = 0.059$, $p = 0.953$).

The interaction effects are shown graphically in figures 9.10a and 9.10b and summarised in tables G.7 and G.8.

9.2.6 Summary of Results

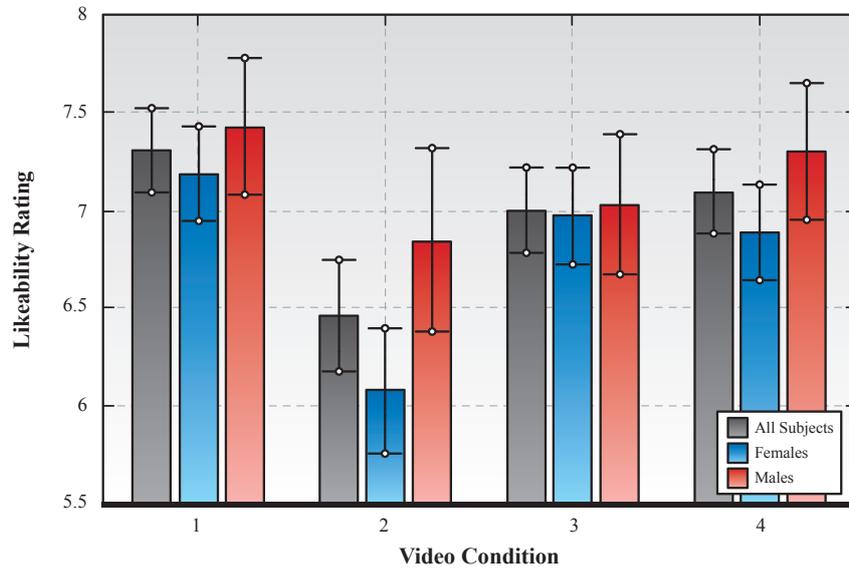
As this section of results has been densely populated with statistical results, this section serves to provide a summary of the main, important findings.

Each of these ANOVAs found a main effect due to the video condition. For all of these, it was shown that V_1 received the highest rating, and V_2 the lowest, with these two conditions being significantly different in all cases. For the *appropriateness*, *naturalness*, *preference* and *likeability* ratings, V_3 and V_4 were not found to be significantly different from each other, but were found to be significantly different (and rated lower) than V_1 and (rated higher) than V_2 . With respect to the *expressiveness* ratings, V_1 , V_2 and V_4 were not found to be significantly different, but were all significantly different from V_2 .

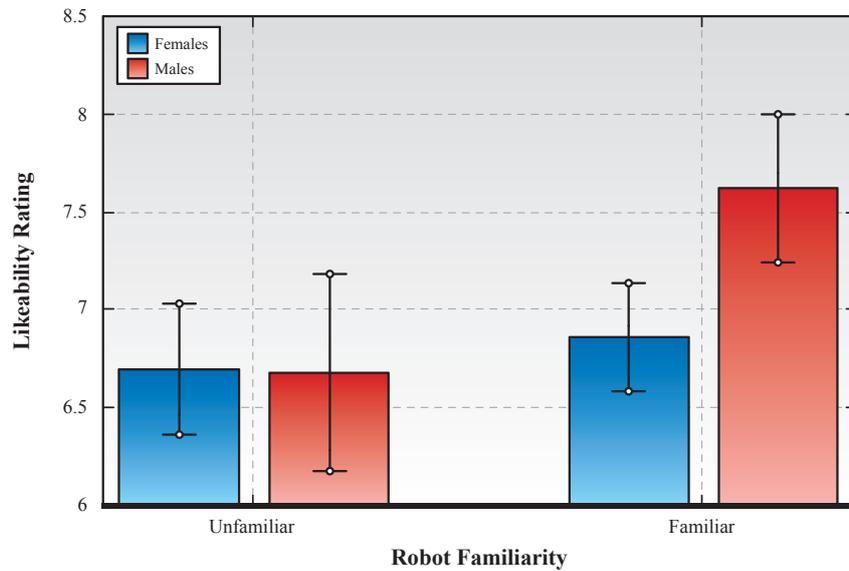
Other main and interaction effects are outlined for each unit of measure here:

Appropriateness Ratings: It was found that for the subjects unfamiliar with the robot there was an interaction between the subject gender and video condition. Similarly, it was found that when isolating the female subjects, there was an interaction effect between the robot familiarity factor and the video condition. However for both of these interaction effects, the post-hoc t-tests found no significant differences between the testing factors for any of the video conditions.

Expressiveness Ratings: As with the appropriateness ratings, it was found that for the subjects unfamiliar with the robot there was an interaction between the subject gender and video condition. Similarly, it was found that when isolating the female subjects, there was an interaction effect between the robot familiarity



(a) Plot of the interaction effect between the video condition and the subject gender.



(b) Plot of the interaction effect between subjects familiar/unfamiliar with the robot and the subject gender.

Figure 9.10: Plot of the mean ratings and 95% Confidence Intervals of the Likeability across the four video condition for all the subjects collectively and the genders, and the interaction effect between the subject gender and robot familiarity.

factor and the video condition. In both cases, it was found that there was a significant difference between the between-subjects factor (gender or robot familiarity) for V_2 , but no other video conditions.

Naturalness Ratings: A main effect was found due to subject gender, where male subjects gave, overall, a higher mean rating than the females. An interaction effect between subject gender and the video condition was also found, however, the post-hoc t-tests found no significant differences between the genders for any of the video conditions.

Preference Ratings: The preference ratings also had a main effect where the subjects who were familiar with the robot tended to provide a higher overall preference rating than subject who were unfamiliar with the robot.

Like-ability Ratings: A main effect due to the robot familiarity was also found, where subjects familiar with the robot provided a higher overall rating than subject unfamiliar with the robot. Also, two, two-way interaction effects were also found. Firstly between the video condition and subject gender, where the males provided significantly higher mean ratings for V_2 and V_4 than the females. The second interaction effect identified was between subject gender and robot familiarity. The post-hoc tests revealed that when unfamiliar with the robot, the two genders had no significant differences in their ratings. However, when subject were familiar with the robot, the males provided a significantly higher overall mean rating than the females.

9.3 Discussion

The results of the tests have clearly shown that the four different video conditions have elicited different ratings for each of the units of measure overall, for each of the videos shown, with these differences being statistically significant in some cases. From the results it is clear that subjects found the language only condition (V_1) to be most appropriate, expressive and natural as well as having the highest overall preference. Conversely, the opposite was true for the NLU only condition (V_2). Together, these results paint a clear picture that a robot using only natural

language holds more promise during an interaction than a robot that uses only NLUs, at least in settings where the human counterparts speak the *same* language as the robot (this experiment has not tested with subjects who *do not* understand English). These findings confirm both H_1 and H_2 which state respectively that a robot that uses only natural language will have the highest overall ratings, and a robot that uses only NLUs will have the lowest overall ratings.

What is also interesting is that the NLU and natural language combinations (V_3 and V_4) represent a *middle-ground* with respect to the overall ratings. Both of the combination conditions were rated higher than the NLU only condition, and in the majority of cases (the expressiveness ratings being the exception) received ratings that were significantly lower than those for the language only condition. Furthermore, few differences were found between the V_3 and V_4 . This shows that the combination of NLUs and natural language do not appear to have a drastic, detrimental effect on people ratings (i.e. the ratings were not the same as the ratings for the NLU only condition). This finding does not confirm H_3 which states that how the NLUs are combined with language will influence the subjects' ratings. This is a promising result suggesting that there is indeed potential for NLUs and natural language to be used in combination.

Furthermore, within the set up of this experiment this makes sense as in both these conditions the robot used natural language to make important statements regarding its knowledge of the game (Lang-1 utterance) and feeding back to the human it is was thinking about which cup the ball was under (Lang-5 and Lang-6 utterances). All the other utterances essentially provided *colouring* to the scenario and were not strictly required and thus may be used interchangeably. With respect to the ratings of expressiveness, while conditions V_3 and V_4 were rated lower than V_1 , this was not found to be statistically significant. This indicates that subjects found natural language to be a far more expressive means of vocalisation than NLUs, and that even when combined with NLUs, the subjects perception of the expressiveness of language remained totally dominant.

It is also interesting to note that in general, the standard error in the ratings for

the NLU only condition were notably higher than those for all the other conditions, showing the subjects provided a far broader range of ratings for this condition than the other three (see the tables in appendix G). This is further evidence of subjects struggling to find a general coherent understanding and interpretation of NLUs in general, as suggested in the experimental results presented in previous chapters (chapters 5 and 7). Also, while the units of measure do not query affect in any way, one might suspect that if they did, one might well get a smaller range of ratings for the V_1 , V_3 and V_4 conditions also.

It is also shown that subject gender (for the naturalness ratings) and the familiarity with the robot (for the appropriateness and preference ratings) impact how the robot is perceived. With respect to the naturalness ratings the male subjects provided higher overall ratings of naturalness than the females. While this effect was found to be significant, it is noteworthy that the difference in the ratings between the genders is rather small, relative to the overall range of the ratings for the video, as well as the overall range of the Likert scale also. As such, it may be prudent to apply little weight to this finding. With respect to the appropriateness and preference ratings found that the subjects familiar with the robot provided, in general, higher overall ratings than the subjects who were unfamiliar with the robot, with the overall profile of the ratings relative to each video being the same across the subject groups. This indicates that the levels of appropriateness and preference appear to be a function of familiarity, and that as experience with the robot is gained, people become more accustomed to the robot and may have a more *positive* view of the robot.

The results also identified a number of interaction effects, however. It was found in the majority that there were no significant differences between the levels of the factors that interacted with the video condition. As a result, it is difficult to interpret these interactions meaningfully. For the like-ability ratings, an interaction effect was found between the two between-subjects factors (subject gender and robot familiarity). It was found that there was no difference in the ratings between the genders when the subjects were unfamiliar with the robot. However,

when subjects were familiar with the robot, the males gave an overall higher mean rating than the females. Furthermore, it was found that when isolating the males, the males who were familiar with the robot gave higher rating than the males unfamiliar with the robot, while there was no such difference for the female subjects. This suggests that it is perhaps the case that as males gain experience with the robot, they are more likely (or easily swayed) to *like* the robot more, while females might have more *inertia* to overcome in this respect.

The final set of results to address is the the low Cronbach's α value for the Preference ratings. This shows that subjects showed little agreement as to their preference for each of the conditions. This low value could well be due to the main effect where subjects familiar with the robot provided significantly higher overall ratings along this scale, as this was a factor that split the subjects almost into two equal halves, introducing a overall difference in the variances in the ratings for the two genders. This may be enough to disrupt the overall Cronbach's α value.

9.3.1 Methodological Remarks

This methodology is subject to shortcomings. Firstly, as with the experiment in the previous chapter (chapter 8), the use of videos rather than real robots is a line of criticism, with the same arguments for the use of videos applying - with real robots it would be highly cumbersome to recreate each of the conditions, for each participant.

Authors have argued both ways with respect to this issue. For example, Woods et al. (2006a,b) and Walters et al. (2011) have all advocated the use of videos as a valid means of measuring HRI, citing that benefits such as reaching a larger subject size and demographic, less effort to administer, and increased control for standardised methodologies (e.g. identical robot behaviours, balanced subject groups and conditions, etc) provide adequate justification for this methodology. These arguments are of course valid and attractive in the eyes of an experimenter, however, studies that have investigated the influence of a physical robot as opposed to a virtual robot present compelling counter arguments. For instance, Bainbridge

et al. (2010) have found that humans obey the instructions of a physical robot more than a virtual robot, even when the instructions are likely to be considered as *wrong* by the subject (e.g. throwing their Professors' text book in a bin). Along a similar line, Leyzberg et al. (2012) have found that the human subjects able to perform a cognitive task with a higher degree of performance when they have been given tuition on the task by a physical robot than when taught by a virtual robot. Here the message is simple: a robot that is physically present makes people behave differently.

The arguments for the use of videos are orientated around the ease of performing high quality, well controlled experiments, which is attractive, however, the arguments for the use of real robots are based around the fact that people behave *differently* when interacting with a real robot and are very persuasive for all HRI. However while these arguments exist, with respect to this experiment, it was deemed the having accurate repeatability through the use of videos and placing the subject in the perspective of an observer (rather than interacting with the robot) took precedence. Though, it is noted that had the subjects been observing a real robot, their ratings may have been different. Furthermore, it seems plausible to argue that had the subject been *interacting* with the robot rather than observing it, then again, one might expect to have obtained different results. These arguments are also applicable to the work regarding morphology, and the biasing that situational context has upon the affective interpretations of NLUs presented in chapters 4 and 8 as these too were online experiments conducted without the use of a real robot.

The experiment presented here is ultimately limited in the insights that it can provide on this issue due to the set up and overall methodology employed. However, it has generally provided a coarse answer to whether NLUs are indeed compatible with natural language, with the answer being, broadly speaking, "yes". This is, however, only scratched the surface of what is very likely a deep avenue of research that extends far beyond the scope of this thesis, which is, in part, why this particular experiment has been set up and conducted in the manner that it

has - it provides an insight into the further possible directions that future work may follow, without deviating too far from the rest of the body of work presented.

9.3.2 Broader Remarks

This section serves to provide a slightly broader interpretation and discussion of the results obtained, with the aim of providing initial pointers to areas that either require further research or may be interesting and fruitful entry points for avenues of further research.

These results would suggest that while language certainly appears to be the preferred modality for vocalisations made by the robot, mixing natural language and NLUs is also acceptable for adults. It is also clear that, when compared with both natural language and the combination of language and NLUs, the sole use of NLUs is the least desirable means for expression. However, this is currently limited to situations in which the humans and the robot speak the same natural language (English in this case), and thus in situations where the robot and human do not speak the same language this may be different.

With respect to the question of whether NLUs may be used alongside language, the results provide evidence that this is indeed possible and has potential, though the ratings for the NLU combinations sit approximately in the middle of the ratings for the language only condition and the NLU only condition, suggesting that there may be a direct, linear relationship between the ratio of the NLUs to natural language and the ratings for each of the five units of measure. It is also unclear as to what the impact of this NLU/natural language combination may be upon real HRI. This study has not controlled the ratio of NLUs to natural language when they are combined, rather it has provided four rather coarse conditions representing roughly the two extremes and the centre point (with two different combinations). This ratio is also a factor likely to influence how the robot is perceived and the potentially (and more importantly) the *quality* of the interaction. It would be beneficial to have an insight as to this what relationship is, as a non-linear relationship in the form of a step function would provide a

useful guideline as to how frequently a robot may use NLUs in combination with natural language before severely impacting the quality of interaction in an adverse manner. This is considered as a valuable avenue for future research.

However, this chapter has only provided a brief insight into the combined use of NLUs and language, and as such, there remain a number of important open questions. Firstly, and perhaps most importantly, the issue of exactly *when* an NLU is made by the robot, relative to any context set by natural language has not been investigated or defined in any manner. *Incorrect* use of NLUs may have potent impacts upon how the robot is perceived by the user(s) and is likely to influence how an interaction ultimately unfolds, as well as the overall *quality* of the interaction. Furthermore, there is also an issue surrounding the *frequency* of use - i.e. the ratio of NLUs to spoken language. NLUs and natural language have a fundamental rivalry rooted in that fact that they both operate through the acoustic modality and thus are competing for “airtime”. As this ratio between the use of NLUs and language changes, the interaction quality is also likely to change.

Similarly to the findings of chapter 8, in which the influence of a context based upon a physical interaction influences the affective interpretation of an NLU, the influence of context defined by what the robot says also requires investigation, particularly given the rich nature of context that may be derived from natural language. This too is considered to be a deep and valuable avenue of future work that will likely lead to important findings that will help inform how NLUs may be best used in social robots that engage in natural language conversations with people.

Finally, another potential avenue of utility may come as a tool for providing back-channel feedback during *active listening* in Sensitive Artificial Listeners (Schröder et al., 2012). In this scenario, it is not so much the combination with language that is of importance, but whether NLUs are able to give (in part) an impression to the speaker that the robot is indeed listening and interested in the subject matter of the conversation as well as the robots affective appraisal of

the subject matter (which in turn influences the unfolding of the conversation). This may be an area of interest particularly with aesthetically and behaviourally simpler robots such as Keepon (Kozima et al., 2009), which have found traction as a means of providing therapy for young users, an area that appears to hold considerable potential as a real world application area of robotic technology.

9.4 Summary

This chapter has presented the results of an experiment aimed at uncovering whether NLUs may be used alongside natural language by a single robotic agent. Using an online crowd sourcing facility, subjects were presented with four different videos, each depicting the same scenario (a ball being placed under one of three cups, the cups shuffled and the robot incorrectly guessing which cup the ball is under), but varying the type of vocalisations that were made by the robots as the scenario unfolded. Specifically, the four videos showed the robot using only natural language, using only NLUs, and in two cases using a (different) combination of NLUs and natural language. After watching each video, subjects were asked to rate the speech/sounds made by the robot with respect to their *appropriateness*, *expressiveness* and how *natural* they felt the vocalisations were. Also, subjects were asked to rate how much they *liked* the robot, and finally, only all the videos had been viewed, provide a *preference* rating for each video.

The results have conclusively shown that subjects gave the robot that used only natural language the highest ratings across all the units of measure, while the robot that used only NLUs was rated lowest. The videos showing the combinations of NLUs and natural language were found to have only minor differences, which were in the majority not statistically significant, with both the videos reciting similar ratings that were approximately in the middle of the ratings for the extreme conditions. While these results are rather intuitive, they do present the message suggesting that it is indeed possible to combine natural language and NLUs in a single robotic agent.

Chapter 10

Conclusions, contributions and future work

This chapter serves to provide an overview of the topics that have been covered within this thesis as well as some concluding remarks. The main themes are recapitulated and reflected upon using the insights that have been gained through the experiments conducted. The overall contributions to the field of social Human-Robot Interaction are discussed, as are the shortcomings and limitations. Finally, a collection of suggestions of future avenues of research are outlined.

10.1 Summary

This thesis has investigated how abstract, robotic Non-Linguistic Utterances may be used during social HRI, particularly with respect to how they may be utilised for making affective displays by a social robot, and what factors are influential in achieving this. The use of NLUs in fictional robotic characters has enjoyed a rich amount of success in the world of Animation and Film, however there is little understanding of how this rich and vibrant expressive modality can be transferred to social robots in the real world. It has been established that there is only a very small community of researchers who have been active in the areas of both NLUs and gibberish speech (as applied to social HRI), and as such, an overall body of research and literature that covers and reviews the topics in their current states,

as a collective, has been lacking. This is arguably the biggest part of the problem, and this thesis has sought to provide an initial step forward in regard to this, and is intended to serve as a stepping stone for other other research efforts to draw and build upon.

After reviewing the directly related and relevant prior works on both gibberish speech and NLUs, it is clear that much of the previous work has been concerned with two things: basic methods that can be used to generate expressive utterances, and how these utterances may be charged to convey different affective states to people, mainly adults. There has been very little work beyond this looking at what factors might impact affective interpretation, and more generally, what affordances NLUs can have during social HRI. Komatsu (2005); Komatsu and Yamada (2007); Komatsu et al. (2010, 2011) appear to be the only authors who have addressed this, using very simple NLUs rather than gibberish, and have focused upon how a robot is able to convey either a positive or negative attitude, and whether a robot's use of NLUs can impact upon how people behave. These are the kinds of questions that are vital if the use of NLUs is to gain increased traction in broader efforts investigating real world HRI both with respect to research, but also with respect to commercial robots. Understanding how to create expressive utterances to convey different affective states is important, but confirming whether these utterances have the same effect during a real world interaction, and what may influence this, is equally essential.

As robots come in a variety of embodiments with different shapes and sizes, a facet that people are sensitive to, this thesis began with an investigation into how a robot's morphology influences the perception that people have of the NLUs that the robot makes (chapter 4). This served two main purposes. Firstly, to probe whether people have a different interpretation (with respect to conveyed affect and intention, as well as what is deemed as acceptable) of the same NLUs made by robot's with different morphologies (an Aldebaran Nao vs. a Sony Aibo), making comparison to utterances that might be more associated with the particular embodiments (e.g. utterances made by humans vs. utterances made by animals).

Secondly, and more importantly for this thesis specifically, whether the Nao robot is an appropriate platform through which NLUs may be embodied and studied. It transpires that when NLUs are presented in a context-free manner, people exhibit very little coherence in how they affectively interpret utterances made by the robot in general, a finding that is echoed in other experiments in this thesis also (see chapters 5, 6 and 7). People also show a preference to having an alignment between the type of utterances made by a robot and the morphology of the robot. More specifically, people found human-like utterances made by the Nao more appropriate than animal-like utterances, and *visa versa*, where animal-like utterances were deemed more appropriate than human-like utterances when made by the Aibo dog robot. Furthermore, people also showed a preference for NLUs being made by the Nao rather than the Aibo. This latter finding essentially provided a basic confirmation that the Nao robot is a suitable platform through which NLUs may be studied.

Given the general small volume of research that has directly investigated NLUs (rather than gibberish speech), there is also a distinct lack of methods and tools available to generate and synthesise utterances beyond single sine waves. Moreover, utterances have all been hand crafted for particular experiments, as well as for animation and commercial robots. As a result of this, it has been necessary to develop a custom tool for designing, characterising and synthesising NLUs¹. The tool takes inspiration from the algorithms described by Breazeal (2002) and Oudeyer (2003) to generate and synthesise expressive gibberish speech, as well as having built in parameters that characterise utterances in a manner that is analogous to the fields psychology and musicology, and could be manipulated independently in a systematic manner. While the tool essentially provided a means of creating an utterance, it was lacking a means or a set of *rules* or a *specification* regarding how different acoustic features of an utterance evoke different affective interpretations in people - a mapping between the utterance parameters and affective interpretations was missing. To address this, two experiments were

¹This tool will also be made freely available to the general public and scientific communities at a later date to serve as a stepping stone for future research efforts.

conducted in which a broad range of different utterances with different parameter values were rated by both children and adults (see chapters 5 and 6), with the main goal of using this data as training data.

This training data was then used to train a collection of feed-forward Multi-Layer Perception Artificial Neural Networks (ANNs) that learnt a mapping between a PAD affect space and each utterance parameter individually (chapter 7) such that a desired affective interpretation could be input to the networks and the values for each utterance parameter, in order to elicit this interpretation, would be output. Interestingly, the mappings that were learnt through this process had stark similarities with the reported acoustic correlates of both the human voice and music with respect to emotional expression (with respect to their dynamics). A small collection of utterances was produced using the output of the ANNs with various affective inputs that correspond to basic affective prototypes in the AffectButton affect space, and were presented to, and affectively rated by young school children as part of a human subject evaluation of the mappings. It was found that while the utterances had acoustic characteristics that followed those that are found in both the human voice and music, people again did not exhibit coherence in their affective interpretations of the utterances.

While the experiments in chapters 4, 5 and 7 found that people do not exhibit coherence in how they affectively interpret NLUs, there was one prevailing trend in how people affectively rated utterances. This was that ratings appeared to be subjects to a perceptual magnet effect and were drawn to particular affective prototypes (or basic emotions). The hypothesis was that people exhibit Categorical Perception when affectively rating NLUs. Using methods that have matured in psychology for measuring categorical perception, the experiments presented in chapter 6 were designed to test this hypothesis head on. It was revealed that children do not appear not exhibit signs of Categorical Perception while adults do, particularly when the scope of potential affective interpretations is narrowed.

Real world HRI inherently has situational context that influences how events that occur within this context are perceived and interpreted by people. It is thus

important to investigate and understand how situational context influences people's affective interpretations of NLUs and this provides more relevant insights regarding how NLUs may be used by a robot during an interaction to better effect. This has been lacking in the literature. Furthermore, such insights also help explain and account for some of the findings from the previous experiments, namely the observation that people generally show little coherence when interpreting NLUs that are presented in a context-free manner. Chapter 8 presented an experiment designed to test how a physical interaction with a robot biases how people interpret the NLUs that the robot makes. Specifically, adults were asked to rate a collection of videos where the Nao robot was subject to a physical action from a human (each with a varying degree of positive or negative valence), where the robot only made an NLU with a perceived valence and no physical interaction occurred, and finally to rate videos where both the physical interaction and NLU made by the robot were combined. This allowed assessment of whether the affective interpretation of the action overrode that of the NLU, or *visa versa*. The results showed that the physical interaction with the robot biases, arguably to the point of dictation, how NLUs are interpreted. A more subtle and interesting effect was also identified: when the interpretation of an action and NLU are *aligned* the overall affective interpretation of the interaction is more *extreme*, showing that while the context plays the predominant role in directing the overall interpretation, people are subtly sensitive to the acoustic features of the utterances that the robot makes.

Finally, this notion of situational context having a prominent bias is extended to the realm of natural language in chapter 9. Natural language is a rich source of situational context and mood within a social interaction and thus will also likely have a similar influence. Furthermore, assuming that NLUs can be used alongside natural language rather than as a replacement, there is a great deal of potential for NLUs to have a supportive utility in this respect by facilitating social cues such as affect bursts, back-channel feedback and affective listening. However, before one can engage with an investigation of these potentials, the assumption upon

which they are founded must be validated. In an experiment, adults were shown videos of a robot playing a guessing game with a human, where the robot either used only natural language, only NLUs, or a combination of the two. It was found that while people prefer a robot that uses only natural language, they prefer a robot that combines natural language and NLUs more than a robot that only uses NLUs. The guideline drawn from this is that if a robot is going to use NLUs, this is best done in combination with natural language rather than as a replacement.

10.1.1 Summary of the main contributions

Overall, this thesis has presented a dense volume of information and data regarding six experiments designed to gain deeper insights into the use of NLUs during real world HRI and factors that impact how NLUs are perceived by people. A number of important findings have been uncovered as a result, and while these are eluded to in the previous section, they are explicitly outlined here:

- **The design and development of a new method of parameterising and creating NLUs** that can be used to generate and systematically explore the acoustic characteristics of NLUs beyond single tones with either a rising or falling pitch envelope.
- **Not all robotic platforms are compatible with NLUs: morphology matters.** People deem it less acceptable for some robots with a given morphology to use NLUs than others with a different morphology, and as such there is an alignment that needs to be made between a robot's physical design and the *vocal* behaviour that it exhibits.
- **People are not coherent in their inferred affective meanings of NLUs** when they are presented in a context-free manner. When different people are presented with the same utterance, they readily perceive it as having an affective meaning, but across people this meaning is not the same: they are not coherent in interpreting NLUs.

- **The acoustic correlates in the human voice and music during affective expression do not appear to have the same effects when transfed to NLUs.** When NLUs have acoustic features that are associated with a particular emotion meaning in the human voice or music, people do not interpret the NLUs as having that same emotion.
- **People exhibit Categorical Perception when affectively interpreting NLUs.** When listening to, and rating NLUs with regard to their affective meaning, peoples' interpretations are drawn to particular (basic) emotional states.
- **Situational context biases greatly how NLUs made by a robot are affectively interpreted.** While two different utterances may have a different meaning in the eyes of a person when they are both made by a robot during an interaction with a clear context and affective mood, the utterances adopt the affective interpretation of the context and lose their original affective meaning.
- **During a vocal interaction NLUs are better used along-side language than on their own.** People show a preference for a robot that uses both NLUs and natural language in a combined manner over a robot that solely uses NLUs.

10.2 Discussion

The tools that have been used throughout this body of work, as well as the experiments that have been conducted are believed to provide a robust arrangement through which NLUs can be studied in a systematic and thorough manner. Nevertheless, there are some remarks and points to be made that require discussion. These issues are addressed in this section.

10.2.1 The custom method for creating NLUs

An important part of this body of work has concerned the design and development of a custom method for creating, parameterising and synthesising NLUs (chapter 3). This was a necessity as no such tool exists for the NLUs, while this is the case for gibberish speech. While on one hand this contributes considerably to the novelty of the research and thesis, the sole use of this method is also an important limiting factor with respect to how the findings of this thesis can be generalised, applied to and compared with other work on NLUs as well as broader areas of HRI also. In essence, the insights underpinning this thesis are heavily tied to the specific method for generating NLUs. If a different method were to have been used, it is unclear whether the same main thesis would have emerged.

Comment may also be made regarding the simplistic approach that has been taken regarding utterances that consist of multiple, single sine wave carrier signals and have a limited number of characterised frequency modulation and are separated by small temporal pauses. While the underlying parameters that control how these simple carrier signals are modulated is designed to be analogous to the characteristics that the fields of psychology and musicology focus upon when studying affective expression via the human voice and musical pieces, the NLUs generated are very simplistic acoustic signals in comparison. It is possible to argue that by employing such a simplistic type of acoustic signal, there are limitations as to the extent of the parallels that may be drawn between the findings of the literature on the human voice and music and those of the NLUs. Furthermore, one can also argue that due to the simplistic nature of the NLUs, potentially interesting and rich results regarding the acoustic correlates of affect in NLUs and in turn how utterances are interpreted have been missed.

To address these, indeed it is possible that the simplistic nature of the NLUs generated indeed do limit the potential parallels that may be drawn the music and the human voice. It is possible that given the abstract nature of the NLUs in comparison to the human voice and music, that these signals are processed differently by the human brain. This is plausible as it has been shown that

there are differences, at a neurological level, in how human speech and music are processed (Tervaniemi et al., 2006). As NLUs sit somewhere in the middle, it is likely that they too are processed differently in the brain.

10.2.2 Using a single robotic platform

The Aldebaran Nao robot has been the sole platform used throughout the body of work that has been presented here. It has been deemed, through human based experimentation, that the Nao is indeed an appropriate platform in which to embody NLUs. However, there are both benefits and drawbacks to the sole use of this platform. The main motivation for using a single platform is that as this work has sought to provide a broad exploration as to how NLUs may be used during social HRI, it was deemed unwise to use multiple different robots with different embodiment and morphologies as this would introduce a discontinuity between the findings across the different experiments. Maintaining a single platform throughout the work circumvents this and keeps the focus of the research on NLUs and the factors that impact their use in real robots.

However, this also results in shortcomings with respect to the generality of the findings in this work. Does the validity findings and conclusions that have emerged from the experiments transfer to the use of NLUs in different robotic platforms? For example it is unclear whether the findings of this thesis apply in the same manner when NLUs are used with robots that do not have the same type embodiment, for example the ASIMO robot or Sony's QRIO humanoid, or other robots that already use NLUs such as Keepon.

Furthermore, it is unwise to assume that the findings here apply to service robots such as the iRobot Roomba, simply because these robot not only have a different embodiment, but they are also not necessarily considered as being *social*. The Nao is a humanoid robot and as a result people tend to anthropomorphise it to a great degree than robots that do not have such an anthropomorphic embodiment. However, it has been shown that people do attribute emotion to such robots through observed motion (Saerbeck and Bartneck, 2010), as well as by adding an-

imated accessories such as a tail (Singh and Young, 2012). This particular issue is addressed in the future work section of of this chapter.

10.2.3 The AffectButton measuring tool

Similarly to the use of a single, custom method of creating NLUs, the use of the AffectButton to capture affective ratings from people is both the source of novelty in this research, and of limitations also. The AffectButton presents a paradigm in which people are asked to explicitly (i.e. self report (Plutchik, 1994)) provide their affective rating or interpretation of a stimulus by selecting an expressive facial gesture from a continuum that matches their interpretation, where the facial gestures are also associated with a three-dimensional co-ordinate within an affect space. The use of this tool has a number of important benefits. Firstly, it provides a way of capturing affect through a continuous manner, rather than using discrete affective labels or categories, meaning that any subtle effects with respect to how people could rate the stimulus are captured, as can larger, more coarse effects. The value of this has been shown in via the experiments in chapter 6 where it was confirmed that peoples' ratings of NLUs are subject to Categorical Perception.

Secondly, given that ratings are captured via facial gestures, this has lent itself to the use with both young children and adults (chapters 5, 6 and 7) in that the underlying dimensional representation of affect has been hidden from the subjects and thus it was not necessary to explain the nature of affective dimensions, which is known to be a cumbersome task (Broekens et al., 2010). However, as the validity of the tool has only been confirmed with teenagers and adults (see Broekens and Brinkman (2013)), it was still necessary to take care to explain to subjects' how the tool worked, and confirm that the young children who partook in the experiments did indeed use the tool in a similar manner as the adults, allowing comparison between the two age groups to be facilitated. The results of these experiments have shown that this is the case.

Marrying these two benefits with the fact that the Nao robot does not have an

expressive face, and that the stimulus presented use the acoustic modality, means that this was a well suited tool that provided an intuitive means for people to provide ratings, and that the social modalities through which the measuring tool and stimulus being rated did not conflict. Though, one must be wary of the effects that have been shown regarding how people perceive expressive gestures made by the human face.

There are drawbacks to the use of the AffectButton also however. Firstly is the concern regarding the compatibility between the rather specific mapping between the facial gestures and the underlying PAD co-ordinates (and more specifically the actual PAD space and mapping itself) with other dimensional models of affect which are used in other tools such as FEELTRACE (Cowie et al., 2000), EMuJoy (Nagel et al., 2007) and the Self Assessment Manikin (Lang and Bradley, 1994). Generally speaking, there is still no firm consensus in the world of Emotion Theory regarding the exact nature of emotion (Plutchik, 1994; Scherer, 2013), and this also extends to the fields concerned with measuring affect also (Cowie and Cornelius, 2003) and as a result a wide variety of different tools have been developed for both the explicit and implicit measurement of affect, and their compatibility with each other is currently rather unclear, a pitfall that this research also falls into.

10.2.4 Child and Adult Evaluations

It is noted in chapter 2 that none of the previous work has actually attempted to perform evaluations with both adults and young children, rather evaluations were only performed with a small number of adults. The work in this thesis has tried to address this, recognising that Child-Robot Interaction is a very promising area of HRI, in which NLUs likely have many exploits. As such, this work has performed evaluations with both children and adults where possible.

However, it has not been possible to perform comparative evaluations between both for all the experiments. This is primarily due to time constraints of the work. Contrary to what intuition might suggest, it is more time effective to perform human based evaluations with children in local schools, rather than adults within

a lab setting. Conducting evaluations in a school means that subjects are readily available, and there is a very small turn around time between subjects. This is not always the case with adults. That said, in the case of the online studies, where people were rating videos of the robot, it was easier to gain access to adults.

As a results of the different type of experiments that have been performed, some of the conclusions and contributions outlined about may have general lacking generalisation to encompass both adults and children.

10.2.5 Automating the generation of NLUs, their affective meaning and the role of situational context

It appears that on their own, without the influence on situational context, NLUs do not project a coherent affective impression on people: people do perceive NLUs have having an expressive, affective meaning, but they do not all see the same meaning from the same utterance. Furthermore, people readily perceive utterances to have very clear and distinct affective meanings, and exhibit Categorical Perception. The lack of general coherence, also holds true when utterances have similar general acoustic characteristics as the human voice when expressing affect, achieved via the use of machine learning. What is interesting about this is that while subjects are not coherent in their interpretations, the neural networks trained on data collected from people did in fact produce an affective mapping that had notable similarities with the acoustic correlates of the human voice and music. This implies that people are sensitive, to some degree, to the acoustic features of an utterance, even when rating NLUs in a context-free manner. This implication is made more concrete by the finding (refer to chapter 8) that when the perceived valence of a situational context and an NLU made within this context are aligned, people perceive the whole situation as having a more extreme valence than when the two valences do not align (in this case the valence of the context overrides that of the NLU).

This general notion has some interesting implications for the how NLUs can be used during an interaction, and how to generate these utterances. Firstly,

given that the situational context biases and directs how NLUs are interpreted affectively, it shows that while NLUs on their own do evoke different but prototypical affective interpretations, coherence of these interpretations emerges from the situation within which the utterances are used - the *other* events that are occurring within the environment provide more salient cues that guide a person's overall interpretation of the situation as it is unfolding. Moreover, given that this holds true even when the original interpretations of the NLUs are different, this suggests that it does not matter what utterance is made, rather that the utterance generally provides a means to further animate the robot.

From the perspective of generating utterances, one hand this can mean that using a *random* utterance, with little regard for the acoustic properties, will suffice. However, a robot is able to modulate the intensity of the affective meaning of an NLU if it aligns the acoustic features of an utterance with the overall valence of the situation in which the NLU is being used. Though, the challenge there is to get the robot to recognise the actual valence of the situation.

10.3 Future Work

This section serves out outline some possible future directions of research that are considered as valuable to further understanding how NLUs may be applied to HRI research and what factors are influential in the successful integration of this expressive modality. Firstly, further exploration with regard to how robots with different embodiments is deemed fruitful as this an initial aspect of an interaction in which expectations of how a robot behaves can be managed. Secondly, the use of different NLU synthesisers to firstly replicate the findings of the research presented here is required, but also to further explore other potential influences that NLUs can have on an interaction and visa versa. Thirdly, the use of NLUs alongside language is clearly an area of potential that needs further exploration as this thesis has only touched upon the matter. Natural Language Interaction is a vital component for engaging HRI, however, much of the field of HRI is contingent upon the state-of-the-art in natural language processing, and currently, it is not at

a level where it can be readily utilised. Rather, much of the HRI field (and other fields) are contingent upon the overall performance of NLP systems, however given the overall lacking performance (in real world settings, and with respect to general open ended conversations), the field of HRI in general is being hindered by this and is having to employ temporary work arounds (such as the WoZ technique). Finally, the use of NLUs in a robot that engages in Long-Term HRI is necessary in order to understand how people perceive and respond to utterances once they are no longer novel is required.

10.3.1 Robot embodiment and morphology

Robots come in all shapes and sizes, and are entering a large variety of different areas that impact the daily life of people. This ranges from service robots that clean our home environments, robots that can provide care and assistance for elderly members of society, through to robots that can be used in the classrooms and educational settings with young children in order to boost performance during the learning process, and as therapeutic tools for people suffering with social disorders.

Part of the research in this thesis, and previous works has touched upon the fact that different morphologies and embodiments causes people to perceive and interpret NLUs in different ways. However, there is little understanding regarding exactly what aspects of a robot's physical design and aesthetic have influence here. Something that is missing is a set of guidelines that robot designers can utilise to better inform the physical design of robots and how this relates to the kinds of vocal (and other) behaviours that the robot exhibits, as clearly there is an alignment that needs to be made between the two. Such a set of guidelines would not only benefit the robot designer, but also researchers who are seeking to examine and further explore either aspect. For example, in this research, the focus has been vocal behaviour, and as such an experiment was performed in order to confirm that the Nao robot was indeed an appropriate platform to be used. Here, the behaviour of the robot has been the priority, and the robot platform itself

has needed to be aligned to this. However, conversely, research into design may be interested in exploring different aesthetics of a robot, and so when evaluating this, it would be useful to have a guideline for the general types of behaviour that would be deemed acceptable for the robot to exhibit such that during human based evaluations, people do not have an adverse reaction to the (mis-)alignment between the robot's morphology and behaviour. A set of guidelines as proposed could be used to aid the design of such evaluations in real HRI.

10.3.2 Exploration of different types of NLU synthesisers and replication of results

When compared with classic examples of NLUs, such as those made by the robot R2D2, the NLUs used in this thesis are still very simple, and as outlined in the discussion above, many of the insights are likely confined to the specific method in which NLUs have been created. In order to validate the overall conclusions and insights gained through this thesis, it is necessary to address some of the same issues that have been covered in this thesis, but with another NLU synthesiser which can make different sounding NLUs. Replication of results is required.

Particularly the finding that people are generally not coherent in their interpreting NLUs when presented in a context free-manner needs replication, as does the finding that NLUs adopt the overall valence and affective interpretation of the situation in which they are used. An easy initial step in this direction would be to replicate the exact experiments and see if the same results emerge, but in the case of situational context, this is something that has many subtle and intricate facets to it, and as such it would also be valuable to explore different contexts also to identify whether this overriding of interpretation is a universal trait of NLUs. If so, this opens many avenues through which NLUs may be used in robotic systems in open, unconstrained HRI.

10.3.3 Using NLU alongside Natural Language

The use of NLUs alongside natural language has been shown to be a possibility in the last experiment presented in this thesis. The suggestion that people prefer a robot that uses NLUs alongside language rather than a replacement is an interesting result. However, the finding that people prefer a robot that speaks only language seemingly stands in the way of using NLUs in real robots. This however, is arguably superficial, in that the technologies for facilitating natural language interaction in robots is far from perfect, and the experiment in chapter 9 was not representative of this.

If used alongside natural language, NLUs have a large number of potential uses. Firstly, they may be used to try and *disguise* the times when NLP fails completely, or simply is taking too long to process the input and provide an appropriate output². Robots that have delayed responses during a conversation has a detrimental effect upon an interaction, and this maybe mitigated simply by making the robot do *something*, as opposed to *nothing*. Given that NLUs only gain coherent meaning when used in a contextual setting, their use in such situations is appealing - the person interacting with the robot projects meaning into what the NLU means, when it could only be a random NLU just to fill a gap of silence.

Further more, back channelling is known to be highly important for both Human-Human Interaction and Human-Robot Interaction, and the use of NLUs in this regard is also seen as advantageous. Not only would NLUs be able to support an interaction by providing vital feedback and social cues during a conversation, but they could also serve the practical function of buying more time for an NLP system if it is struggling, and would do so with little computational expense.

What is not currently known is how NLUs are best used alongside language, and there are a few important open questions regarding this. Firstly, the notion itself must be validated in a more robust manner. The experiment in this thesis only serves to provide a hint that this is possible - it does not confirm it completely.

²Identifying when the NLP system has *misunderstood* (i.e. speech recognition has failed) something a person has said and produces an incorrect response is a harder problem to overcome.

Furthermore, how frequently NLUs are used is also something that requires attention, as this will likely impact the quality of interaction. For example, if the robot uses NLUs very infrequently, will this have a negative impact, and if so, would it have more of a negative impact than if NLUs were used more frequently (meaning that people would come to expect these types of utterance from a robot)? These are the types of fundamental questions that need addressing if the potential use of NLUs during verbal conversations are indeed to be realised.

10.3.4 Long-Term Human-Robot Interaction

Establishing and maintaining Long-Term HRI is a very current goal of the field of HRI, and is something that is now beginning to be tackled head on. As such, it is not clear exactly how people will respond to robots, generally, after long periods of time, and this applies to a very large number of different aspects of HRI. As such, it is likely that many findings that have already been reported are likely to require validation and replication with respect to their validity during Long-Term HRI.

This is also true for the use of NLUs in robots. However, if NLUs indeed can be used in a robot for longer term interactions, a great number of interesting questions emerge from this. Let us assume for the time being that they can. This thesis has opted for the view that NLUs do not constitute a language as some fundamental issues regarding established cultural rules concerning vocabulary and grammar are missing. If NLUs were used in robots in a consistent manner, for long periods of time, it can be argued that through this, it is possible that an established vocabulary and grammar would emerge. This may be a potent inroad toward the evolution of NLUs as single abstract utterances to an artificial language that could at least be understood by people, through they would perhaps struggle to speak it themselves.

Furthermore, use of NLU through long term interactions may also be a way of solving the lack of coherence that people exhibit. It can be argued that an important factor that leads to this incoherence is that lack of prior experience

that can be drawn upon in order to decode the affective meaning that different utterances can have. Through having increased experience with the utterances a robot makes in different situations, it is likely that people will begin to form associations between the acoustic features of an utterance and the perceived affective meanings. Most of this is of course highly speculative, but if such developments were to begin emerging it would paint a very bright future for the use of NLUs in real social HRI in the long term.

Appendix A

Methods

Pseudo code implementation of the NLU generation algorithm

Algorithms 1, 2, 3, 5, 4, 6, 7 present the NLU generation algorithm in pseudo code, with this section serving to provide a descriptive, supportive overview. It is noteworthy that this description assumes that all Utterance parameters are applied globally at an utterance level - that is to say that all parameters aside from the pitch contour are applied uniformly to all sound units.

The algorithm has in essence two stages. The first stage is the generation process whereby the *blueprint* for an utterance is generated. The second part is taking this blueprint, and transforming it to real-world units of Frequency and Time ready for synthesis through SuperCollider. Referring to Algorithm 1, once the all the Global Utterance parameters have been specified, the sound units are recursively generated and concatenated, being stored in a List which is a member of the Utterance instance. Each sound unit has an overall start and end time stamp, and a duration specification, as do the spoken and silent components. Initially, these values do not have any specific unit of measure, but rather are there in place to provide a relative relationship and proportional relationship between all sound units within the utterance with respect to their respective temporal durations.

It can be seen that the duration of the spoken component is determined by the `rhythm` parameter, which in essence controls the random variance in temporal duration (t) that a utterance may have from a fixed length of 1, where $0 \leq t \leq 1$. This is the only aspect of the whole generation algorithm that is left to random chance. The start time stamp is set to the total duration of all the previous sound units within the utterance (this is why there is a running total as each sound unit is generated). The end time stamp of the spoken component is set to the addition of the start time stamp and the duration value. With respect to the silent component, the start time stamp is set to the same value as the end time stamp of the spoken component, and the duration is proportional to the duration

of spoken component, using the pause ratio as the control. At the sound unit level, the start time stamp is set to the same value as the startStamp of the spoken component and the end time stamp is set to the endStamp value of the silent component.

The next phase is to recursively create and modulate the carrier signals and respective frequency and amplitude envelopes for the spoken component, the number of which is specified by the `waveCount` parameter. The envelopes are stored in a `Wave` object, which has two members, a frequency and amplitude envelope. Once created, each word object is appended to the `waveList` which is a member of the `SoundUnit` object. Algorithm 6 outlines the procedure to generate the amplitude envelopes. As mentioned previously, the amplitude envelopes have been kept constant throughout this work, and as such, the `GENERATEAMPLITUDEENVELOPE` function is in essence a *static* process and applies equally to all sound units that are generated¹. Algorithm 2 outlines the process for generating the Frequency Envelope of a carrier signal and begins by determining the type of pitch contour modulation that is to be generated as the process is somewhat different for each type of contour. The pitch contour shape may have one of five different values: “flat”, “rising”, “falling”, “rising-falling”, and “falling-rising”.

These five contour characterisations may all be broken down to, and built up from the generation of a single, horizontal, linear array of equally spaced and temporally sequenced nodes, with the frequency values being set as the tangent of angle specified by the `tremolo` parameter as to produce the tremolo effect. The result of this alone is the “Flat” contour. Rotating this array about first node (located at point (0,0) in the bi-normal space) through the `angle` parameter value, a “Rising” contour is produced. By inverting the frequency values of the nodes, the rising contour may be transformed to a “Falling” contour. Finally, “Rising-Falling” and “Falling-Rising” contours may be produced by creating and concatenating the “Rising” and “Falling” contours in the respective orders (this process is described below).

As a result of the Rising-Falling and falling-Rising contours being created

¹This is true for all the work that is described in this Thesis.

from the addition of two individual node array, in Algorithm 2 the process for generating the Flat, Rising and Falling contours is notably difference than the process for the Rising-Falling and Falling-Rising contours. In the case of the first three contours, only a single node array is generated, with the nodeCount (n), tremolo angle (t) and rotation being specified (a), as well as a invert flag (i). In the case of Rising-Falling and Falling-Rising contours, two node arrays must first be generated, and as a result, more generation parameters must be considered also as the nodeRatio and skewRatio parameters only apply to these contour shapes. To cater for these two other parameters, a nodeCount must be specified for each array in order to adhere to the nodeRatio². This is done by multiplying the overall node count by the nodeRatio, and rounding this value up to the nearest integer for the first array, and the nodeCount for the second array set to the total nodeCount minus the nodeCount for the first array³. Also, each array also requires a different invert flag, as one array will be rising, and the other falling. Finally, in addition to calling the GENERATEARRAY function twice to generate two arrays, the ADDARRAYS function is also called, returning a concatenation of the two arrays.

Algorithm 2 outlines the GENERATEARRAY function which beings by creating the horizontal array of nodes, adding a tremolo effect, and then rotating them. However, before checking and acting upon the invert flat, the array is normalized using the NORMALIZEARRAY function (see Algorithm 4). This function takes an input node array and scales all the node values such that their values fully cover the range $[-1\ 1]$ in the frequency dimension, and the range $[0\ 1]$ in the time dimension. The pitchContour shape of the array is also passed to the function, as in the case of a Flat contour, the frequency values are not scaled as this would distort any tremolo effect⁴.

The ADDLINES function is used to concatenate two node arrays, and is used

²A note about the nodRatio parameter: this particular parameter is only useful when used in conjunction with the tremolo parameter, and when the tremolo $\neq 0$

³There is also an subtle but important detail to highlight here, and that is that the second array has 1 added to the value. The practical nature of this will be outlined later in this text as it is related to the ADDARRAY function.

⁴The result of this would require more complex control of the frequency range parameter, at the (local) sound unit level, rather than the (global) utterance level.

only to create the rising-falling and falling-rising pitch contours. It receives three inputs, the two lines that are to be concatenated, and the skew ratio that is to be applied to the envelope. In order to apply the skew ratio, the total duration of the two lines must be calculated, and then two scalar constants (r_1 and r_2) are calculated, one for each input node array respectively.. These constants determine what proportion of the spoken duration each array must cover. Each array is then recursed and the respective constant applied to the temporal component (x) of each Node. There are some noteworthy features of the function. Firstly, as all the input lines to this function are already normalised, the frequency components of each Node (y) have a value of 1 added to them placing them into the range $[0\ 2]$. This is temporary, and done in order to reuse the NORMALIZELINE function and adhere to the assumption that all node values are greater or equal to zero, and once normalized, the frequency values fall back into the range $[-1\ 1]$. Secondly, when iterating the Nodes in $line_2$, $n = 2$ rather than 1. This is the aforementioned detail that relates to the addition of 1 to the nodeCount of the second node array for Rising-Falling and Falling-Rising pitch contours and has the function of skipping the first Node in the array, avoiding overlap in both the time and frequency dimensions with the last Node in $line_1$.

Once all the sound units have been created, the blueprint for the utterance is complete. Currently all the values of the properties of the utterance have no specific units of measure associated with them. As such, the next phase of the generation process is to transform the blueprint to an utterance with real world units. This is done via the SCALETOREALUNITS function (Algorithm 7). This function has four inputs: a specific sound unit whose values are to be scales, the scalar constant to be applied to all duration values, and the base and frequency range values which are used to calculate the frequency values for each node individually. With respect to the duration, the duration values and time stamps are corrected as these values are used to calculate the real world time values for each of the nodes. The frequency values for the nodes are calculated by multiplying the normalised node value but the absolute difference between the frequency base

and range values, and adding this proportional value to the base frequency value. This process must be completed if any changes are made to any of the utterance properties or parameters, and hence is only called just prior to utterance synthesis through SuperCollider.

Algorithms 1, 2, 3, 5, 4, 6, 7 present the NLU generation algorithm in pseudo code, with this section serving to provide a descriptive, supportive overview. It is noteworthy that this description assumes that all Utterance parameters are applied globally at an utterance level - that is to say that all parameters aside from the pitch contour are applied uniformly to all sound units.

The algorithm has in essence two stages. The first stage is the generation process whereby the *blueprint* for an utterance is generated. The second part is taking this blueprint, and transforming it to real-world units of Frequency and Time ready for synthesis through SuperCollider. Referring to Algorithm 1, once the all the Global Utterance parameters have been specified, the sound units are recursively generated and concatenated, being stored in a List which is a member of the Utterance instance. Each sound unit has an overall start and end time stamp, and a duration specification, as do the spoken and silent components. Initially, these values do not have any specific unit of measure, but rather are there in place to provide a relative relationship and proportional relationship between all sound units within the utterance with respect to their respective temporal durations.

It can be seen that the duration of the spoken component is determined by the `rhythm` parameter, which in essence controls the random variance in temporal duration (t) that a utterance may have from a fixed length of 1, where $0 \leq t \leq 1$. This is the only aspect of the whole generation algorithm that is left to random chance. The start time stamp is set to the total duration of all the previous sound units within the utterance (this is why there is a running total as each sound unit is generated). The end time stamp of the spoken component is set to the addition of the start time stamp and the duration value. With respect to the silent component, the start time stamp is set to the same value as the end time

stamp of the spoken component, and the duration is proportional to the duration of spoken component, using the pause ratio as the control. At the sound unit level, the start time stamp is set to the same value as the startStamp of the spoken component and the the end time stamp is set to the endStamp value of the silent component.

The next phase is to recursively create and modulate the carrier signals and respective frequency and amplitude envelopes for the spoken component, the number of which is specified by the `waveCount` parameter. The envelopes are stored in a `Wave` object, which has two members, a frequency and amplitude envelope. Once created, each word object is appended to the `waveList` which is a member of the `SoundUnit` object. Algorithm 6 outlines the procedure to generate the amplitude envelopes. As mentioned previously, the amplitude envelopes have been kept constant throughout this work, and as such, the `GENERATEAMPLITUDEENVELOPE` function is in essence a *static* process and applies equally to all sound units that are generated⁵. Algorithm 2 outlines the process for generating the Frequency Envelope of a carrier signal and begins by determining the type of pitch contour modulation that is to be generated as the process is somewhat different for each type of contour. The pitch contour shape may have one of five different values: “flat”, “rising”, “falling”, “rising-falling”, and “falling-rising”.

These five contour characterisations may all be broken down to, and built up from the generation of a single, horizontal, linear array of equally spaced and temporally sequenced nodes, with the frequency values being set as the tangent of angle specified by the `tremolo` parameter as to produce the tremolo effect. The result of this alone is the “Flat” contour. Rotating this array about first node (located at point (0,0) in the bi-normal space) through the `angle` parameter value, a “Rising” contour is produced. By inverting the frequency values of the nodes, the rising contour may be transformed to a “Falling” contour. Finally, “Rising-Falling” and “Falling-Rising” contours may be produced by creating and concatenating the “Rising” and “Falling” contours in the respective orders (this process is described below).

⁵This is true for all the work that is described in this Thesis.

As a result of the Rising-Falling and falling-Rising contours being created from the addition of two individual node array, in Algorithm 2 the process for generating the Flat, Rising and Falling contours is notably different than the process for the Rising-Falling and Falling-Rising contours. In the case of the first three contours, only a single node array is generated, with the nodeCount (n), tremolo angle (t) and rotation being specified (a), as well as an invert flag (i). In the case of Rising-Falling and Falling-Rising contours, two node arrays must first be generated, and as a result, more generation parameters must be considered also as the nodeRatio and skewRatio parameters only apply to these contour shapes. To cater for these two other parameters, a nodeCount must be specified for each array in order to adhere to the nodeRatio⁶. This is done by multiplying the overall node count by the nodeRatio, and rounding this value up to the nearest integer for the first array, and the nodeCount for the second array set to the total nodeCount minus the nodeCount for the first array⁷. Also, each array also requires a different `invert` flag, as one array will be rising, and the other falling. Finally, in addition to calling the `GENERATEARRAY` function twice to generate two arrays, the `ADDARRAYS` function is also called, returning a concatenation of the two arrays.

Algorithm 2 outlines the `GENERATEARRAY` function which begins by creating the horizontal array of nodes, adding a tremolo effect, and then rotating them. However, before checking and acting upon the invert flag, the array is normalized using the `NORMALIZEARRAY` function (see Algorithm 4). This function takes an input node array and scales all the node values such that their values fully cover the range $[-1\ 1]$ in the frequency dimension, and the range $[0\ 1]$ in the time dimension. The pitchContour shape of the array is also passed to the function, as in the case of a Flat contour, the frequency values are not scaled as this would distort any tremolo effect⁸.

⁶A note about the nodeRatio parameter: this particular parameter is only useful when used in conjunction with the `tremolo` parameter, and when the `tremolo` $\neq 0$

⁷There is also a subtle but important detail to highlight here, and that is that the second array has 1 added to the value. The practical nature of this will be outlined later in this text as it is related to the `ADDARRAY` function.

⁸The result of this would require more complex control of the frequency range parameter, at the (local) sound unit level, rather than the (global) utterance level.

The `ADDLINES` function is used to concatenate two node arrays, and is used only to create the rising-falling and falling-rising pitch contours. It receives three inputs, the two lines that are to be concatenated, and the skew ratio that is to be applied to the envelope. In order to apply the skew ratio, the total duration of the two lines must be calculated, and then two scalar constants (r_1 and r_2) are calculated, one for each input node array respectively.. These constants determine what proportion of the spoken duration each array must cover. Each array is then recursed and the respective constant applied to the temporal component (x) of each Node. There are some noteworthy features of the function. Firstly, as all the input lines to this function are already normalised, the frequency components of each Node (y) have a value of 1 added to them placing them into the range $[0\ 2]$. This is temporary, and done in order to reuse the `NORMALIZELINE` function and adhere to the assumption that all node values are greater or equal to zero, and once normalized, the frequency values fall back into the range $[-1\ 1]$. Secondly, when iterating the Nodes in $line_2$, $n = 2$ rather than 1. This is the aforementioned detail that relates to the addition of 1 to the `nodeCount` of the second node array for Rising-Falling and Falling-Rising pitch contours and has the function of skipping the first Node in the array, avoiding overlap in both the time and frequency dimensions with the last Node in $line_1$.

Once all the sound units have been created, the blueprint for the utterance is complete. Currently all the values of the properties of the utterance have no specific units of measure associated with them. As such, the next phase of the generation process is to transform the blueprint to an utterance with real world units. This is done via the `SCALETOREALUNITS` function (Algorithm 7). This function has four inputs: a specific sound unit whose values are to be scales, the scalar constant to be applied to all duration values, and the base and frequency range values which are used to calculate the frequency values for each node individually. With respect to the duration, the duration values and time stamps are corrected as these values are used to calculate the real world time values for each of the nodes. The frequency values for the nodes are calculated by multiplying

the normalised node value but the absolute difference between the frequency base and range values, and adding this proportional value to the base frequency value. This process must be completed if any changes are made to any of the utterance properties or parameters, and hence is only called just prior to utterance synthesis through SuperCollider.

Algorithm 1

```
U = new Utterance
U.duration = 0
U.soundCount = set to value
U.bluePrint = new List
U.realPrint = new List

//Create Utterance Blueprint
for n = 1 to U.soundUnitCount do
  s = new SoundUnit

  s.spoken = new SpokenComponent
  s.spoken.startStamp = Global.totalDuration
  s.spoken.duration = (1 - U.rhythm) + [(1 - U.rhythm) · rand()]
  s.spoken.endStamp = s.spoken.startStamp + s.spoken.duration
  s.waveList = new List

  s.silent = new SilentComponent
  s.silent.startStamp = s.spoken.endStamp
  s.silent.duration = s.spoken.duration · U.pauseRatio
  s.silent.endStamp = s.silent.startStamp + s.silent.duration

  s.startStamp = s.spoken.startStamp
  s.duration = s.spoken.duration + s.silent.duration
  s.endStamp = s.startStamp + s.duration

  U.duration += (s.spoken.duration + s.silent.duration)

  for i = 1 to U.waveCount do
    freqEnv = GENERATEFREQUENCYENVELOPE() ▷ Goto Algorithm 2
    ampEnv = GENERATEAMPLITUDEENVELOPE() ▷ Goto Algorithm 6
    s.spoken.waveList[-1].freqEnv = freqEnv
    s.spoken.waveList[-1].ampEnv = ampEnv
  end for

  U.bluePrint.append(s)
end for

//Create Utterance Real Print
for all in U.bluePrint do
  r = new SoundUnit
  k = (soundUnitCount / speechRate) / totalDuration
  r ← SCALETOREALUNITS(s, k, U.baseFreq, U.freqRange) ▷ Goto
Algorithm 7
  U.realPrint.append(r)
end for
```

Algorithm 2

```
function GENERATEFREQUENCYENVELOPE
  if wave.pitchContour == “Flat” then
     $n = \text{wave.nodeCount}$ 
     $t = \text{wave.tremolo}$ 
     $a = 0$ 
     $i = \text{false}$ 
     $\text{wave.nodeList} \leftarrow \text{GENERATEARRAY}(n, t, a, i)$     ▷ Goto Algorithm 3

  else if wave.pitchContour == “Rising” then
     $n = \text{wave.nodeCount}$ 
     $t = \text{wave.tremolo}$ 
     $a = (2\pi)/8$ 
     $i = \text{false}$ 
     $\text{wave.nodeList} \leftarrow \text{GENERATEARRAY}(n, t, a, i)$     ▷ Goto Algorithm 3

  else if wave.pitchContour == “Falling” then
     $n = \text{wave.nodeCount}$ 
     $t = \text{wave.tremolo}$ 
     $a = (2\pi)/8$ 
     $i = \text{true}$ 
     $\text{wave.nodeList} \leftarrow \text{GENERATEARRAY}(n, t, a, i)$     ▷ Goto Algorithm 3

  else if wave.pitchContour == “Rising-Falling” then
     $n_1 = \text{ceil}(\text{wave.nodeCount} \cdot \text{wave.nodeRatio})$ 
     $n_2 = (\text{wave.nodeCount} - n_1) + 1$     ▷ Add one, this will be removed later
     $r = \text{wave.skewRatio}$ 
     $t = \text{wave.tremolo}$ 
     $a = (2\pi)/8$ 
     $i_1 = \text{false}$ 
     $i_1 = \text{true}$ 
     $l_1 \leftarrow \text{GENERATEARRAY}(n_1, t, a, i_1)$                 ▷ Goto Algorithm 3
     $l_2 \leftarrow \text{GENERATEARRAY}(n_2, t, a, i_2)$                 ▷ Goto Algorithm 3
     $\text{wave.nodeList} \leftarrow \text{ADDARRAYS}(l_1, l_2)$                 ▷ Goto Algorithm 5

  else if wave.pitchContour == “Falling-Rising” then
     $n_1 = \text{ceil}(\text{wave.nodeCount} \cdot \text{wave.nodeRatio})$ 
     $n_2 = (\text{wave.nodeCount} - n_1) + 1$     ▷ Add one, this will be removed later
     $r = \text{wave.skewRatio}$ 
     $t = \text{wave.tremolo}$ 
     $a = (2\pi)/8$ 
     $i_1 = \text{true}$ 
     $i_2 = \text{false}$ 
     $l_1 \leftarrow \text{GENERATEARRAY}(n_1, t, a, i_1)$                 ▷ Goto Algorithm 3
     $l_2 \leftarrow \text{GENERATEARRAY}(n_2, t, a, i_2)$                 ▷ Goto Algorithm 3
     $\text{wave.nodeList} \leftarrow \text{ADDARRAYS}(l_1, l_2, r)$                 ▷ Goto Algorithm 5
  end if
end function
```

Algorithm 3

```
function GENERATEARRAY(nodeCount, theta, angle, invert)
  nodeList = new List
   $k = 1/(\text{nodeCount} - 1)$   $\triangleright$   $k$  is used to keep the x and y values in a relative
  scale.

  for  $i = 0 \rightarrow (\text{nodeCount} - 1)$  do
    node = new Node

    if  $i == 0$  or  $i == \text{nodeCount} - 1$  then
      node.x =  $i \cdot k$ 
      node.y = 0
    else if  $i \bmod 2 == 0$  then
      node.x =  $i \cdot k$ 
      node.y =  $-\tan(\text{theta}) \cdot k$ 
    else
      node.x =  $i \cdot k$ 
      node.y =  $\tan(\text{theta}) \cdot k$ 
    end if

    nodeList[-1]  $\leftarrow$  node  $\triangleright$  Append this node to the nodeList
  end for

  // Rotate the Line.
  for all node  $\in$  nodeList do
    node.x =  $(\cos(\text{angle}) \cdot \text{node.x}) + (\sin(\text{angle}) \cdot \text{node.y})$ 
    node.y =  $(\sin(\text{angle}) \cdot \text{node.x}) - (\cos(\text{angle}) \cdot \text{node.y})$ 
  end for

  nodeList  $\leftarrow$  NORMALIZELINE(nodeList)  $\triangleright$  Goto Algorithm 4

  // If needed, invert the frequency values of the line
  if invert == true then
    for all node in nodeList do
      node.y =  $(-1) \cdot \text{node.y}$ 
    end for
  end if

  return nodeList
end function
```

Algorithm 4

function NORMALIZEARRAY(*nodeList*, *baseFreq*, *freqRange*, *pitchContour*)
Assume that all node values in *nodeList* ≥ 0 !

$x_{max} = 0$

$y_{max} = 0$

for all *node* **in** *nodeList* **do**

if $x_{max} \leq \text{node.x}$ **then**

$x_{max} = \text{node.x}$

end if

if $y_{max} \leq \text{node.y}$ **then**

$y_{max} = \text{node.y}$

end if

end for

for all *node* **in** *nodeList* **do**

$\text{node.x} = (\text{node.x}/x_{max})$

if *pitchContour* \neq "Flat" **then**

$\text{node.y} = [(\text{node.y}/y_{max}) \cdot 2] - 1$

end if

end for

return *nodeList*

end function

Algorithm 5

```
function CONCATENATEARRAYS(line1, line2, ratio)
  nodeList = new List
  totalDuration = line1[-1].x + line2[-1].x
  r1 = totalDuration · ratio
  r2 = totalDuration · (1 - ratio)

  for n = 1 → line1.size do
    tmp = new Node
    tmp.x = line1.x · r1
    tmp.y = line1.y + 1
    nodeList.append(tmp)
  end for

  midPoint = line1[-1].x

  for n = 2 → line2.size do                                ▷ Note that n starts at 1 here!
    tmp = new Node
    tmp.x = midPoint + (line2.x · r2)
    tmp.y = line2.y + 1
    nodeList.append(tmp)
  end for

  nodeList ← NORMALIZELINE(nodeList)                            ▷ Goto Algorithm 4

  return nodeList
end function
```

Algorithm 6

function GENERATEAMPLITUDEENVELOPE

*node*₁ = *newNode*

*node*₁.*x* = 0

*node*₁.*y* = 0

*node*₂ = *newNode*

*node*₂.*x* = 0.05

*node*₂.*y* = 1.0

*node*₃ = *newNode*

*node*₃.*x* = 0.2

*node*₃.*y* = 0.7

*node*₄ = *newNode*

*node*₄.*x* = 0.85

*node*₄.*y* = 0.7

*node*₅ = *newNode*

*node*₅.*x* = 1.0

*node*₅.*y* = 0

return [*node*₁, *node*₂, *node*₃, *node*₄, *node*₅]

end function

Algorithm 7

```
function SCALETOREALUNITS(s, k, baseFreq, freqRange)
  s.spoken.startStamp* = k
  s.spoken.duration* = k
  s.spoken.endStamp* = k

  s.silent.startStamp* = k
  s.silent.duration* = k
  s.silent.endStamp* = k

  s.startStamp = s.spoken.startStamp
  s.duration = s.spoken.duration + s.silent.duration
  s.startStamp = s.silent.endStamp

  for all wave in sound.spoken.waveList do
    for all node in wave.freqEnv do
      node.x = s.spoken.startStamp + (node.x · s.spoken.duration)
      node.y = baseFreq + (node.y · |baseFreq − freqRange|)
    end for

    for all node in wave.ampEnv do
      node.x = s.spoken.startStamp + (node.x · s.spoken.duration)
      node.y = node.y · s.volumeIntensity
    end for
  end for

  return s
end function
```

Algorithm 8 Calculating Pleasure, Arousal and Dominance from Screen Coordinates.

$x_m = \text{mouse.x}$
 $y_m = \text{mouse.y}$
 $\text{factor} = 0.55$
 $\text{sensitivity} = 1.1$
 $p = x_m \cdot \text{sensitivity}$
 $a = 0$
 $d = y_m \cdot \text{sensitivity}$

if ($p \geq \text{factor}$) **or** ($d \geq \text{factor}$) **or** ($p \leq -\text{factor}$) **or** ($d \leq -\text{factor}$) **then**

if ($p = 0$) **then**

if then($d \leq 0$)

$x_1 = 0$

$x_2 = 0$

$y_1 = \text{factor}$

$y_2 = 1.0$

else

$x_1 = 0$

$x_2 = 0$

$y_1 = -\text{factor}$

$y_2 = -1.0$

end if

else

if ($p \geq d$) **and then**

$x_1 = \text{factor}$

$x_2 = 1.0$

$y_1 = x_1 \cdot rc$

$y_2 = x_2 \cdot rc$

else if ($p \geq d$) **and** ($-p \geq -d$) **then**

$y_1 = -\text{factor}$

$y_2 = -1.0$

$x_1 = y_1/rc$

$x_2 = y_2/rc$

else if and ($-p \geq d$) **then**

$x_1 = -\text{factor}$

$x_2 = -1.0$

$y_1 = x_1 \cdot rc$

$y_2 = x_2 \cdot rc$

else

$y_1 = \text{factor}$

$y_2 = 1.0$

$x_1 = y_1/rc$

$x_2 = y_2/rc$

end if

end if

$d_1 = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$

$a_1 = \sqrt{(x_1 - p)^2 + (y_1 - d)^2}$

$a = 2 \cdot (a_1/d_1) - 1$

else

$a = -1$

end if

Algorithm 9 Calculating Pleasure, Arousal and Dominance from Screen Coordinates.

```
if  $p \geq 1.0$  then
     $p = 1.0$ 
else if  $(p \leq -1.0)$  then
     $p = -1.0$ 
end if
```

```
if  $a \geq 1.0$  then
     $a = 1.0$ 
else if  $(a \leq -1.0)$  then
     $a = -1.0$ 
end if
```

```
if  $d \geq 1.0$  then
     $d = 1.0$ 
else if  $(d \leq -1.0)$  then
     $d = -1.0$ 
end if
```

```
return  $[p, a, d]$ 
```

Appendix B

Alignment of NLUs with Agent

Morphology

Table B.1: Ratings and χ^2 values for overall Affective Ratings across robot and sound categories. The columns for each utterance class are the overall percentage of ratings for a given affective label, the results of a χ^2 test checking that that rating is above chance, and the results of a Stewart-Maxwell test indicating whether there was an overall difference in the distribution of ratings due to the robot image that was also presented. Please see Figures 4.3a, 4.3b and 4.3c

Emotion	Robot	Utterance Class								
		Human			Animal			Technological		
		Rating (%)	χ^2	S&M (χ^2)	Rating (%)	χ^2	S&M (χ^2)	Rating (%)	χ^2	S&M (χ^2)
Anger	Nao	3.006	21.642*		15.738	5.876		17.486	20.082*	
	Aibo	3.279	20.208*		19.016	17.155*		17.304	18.951*	
Fear	Nao	3.006	21.642*		15.082	4.328		12.751	1.328	
	Aibo	1.366	31.281†		14.426	3.017		13.661	3.213	
Disgust	Nao	5.465	10.503		13.771	1.941		13.297	2.361	
	Aibo	4.645	13.773		12.459	0.499		13.115	1.984	
Happiness	Nao	11.202	0.003		9.508	0.705		10.383	0.262	
	Aibo	11.202	0.003		7.213	4.171		11.475	0.066	
Sadness	Nao	3.552	18.822*	8.346	11.803	0.132	8.347	8.015	4.738	8.649
	Aibo	1.923	27.872†		11.147	0.000		9.836	0.803	
Affection	Nao	12.295	0.462		12.459	0.499		6.011	12.853	
	Aibo	12.295	0.462		14.098	2.450		5.282	16.787†	
Boredom	Nao	14.481	3.740		6.230	6.541		9.472	1.328	
	Aibo	15.027	5.052		9.180	1.023		7.468	6.557	
Interest	Nao	33.880	170.765†		3.934	14.138		17.501	22.443†	
	Aibo	36.339	209.642†		2.295	21.335*		19.126	31.738†	
Relaxation	Nao	13.115	1.322		11.475	0.036		4.736	20.082*	
	Aibo	13.934	2.626		10.164	0.246		2.732	34.689†	

† : $p < 0.05$

* : $p < 0.025$

* : $p < 0.01$

† : $p < 0.005$

Table B.2: Gender differences for Affective Ratings across the Utterance Classes and Robot Morphologies. The table shows the percentages of each affective ratings for each robot across the two genders and the $\chi^2(8)$ values indicating whether this rating is above chance (compared to a flat uniform distribution).

Emotion	Subject	Human				Animal				Technological			
		Rating (%)		χ^2		Rating (%)		χ^2		Rating (%)		χ^2	
		Nao	Aibo	Nao	Aibo	Nao	Aibo	Nao	Aibo	Nao	Aibo	Nao	Aibo
Anger	Males	1.852	3.086	12.500	9.389	20.741	25.926	11.267	26.667 ‡	20.988	18.519	21.333 *	12.000
	Females	3.922	3.432	9.490	10.828	11.765	13.529	0.065	0.895	14.706	16.340	3.559	7.529
Fear	Males	1.235	1.235	14.222	14.222	16.296	14.815	3.267	1.667	11.111	10.699	0.000	0.037
	Females	4.412	1.471	8.240	17.064 †	14.118	14.118	1.383	1.383	14.052	16.013	2.382	6.618
Disgust	Males	8.025	6.173	1.389	3.556	11.852	11.111	0.067	0.000	11.934	13.580	0.148	1.333
	Females	3.431	3.431	10.828	10.828	15.294	13.529	2.677	0.095	14.379	12.745	2.941	0.735
Happiness	Males	10.494	6.790	0.056	2.722	5.185	5.182	4.267	4.267	10.288	10.288	0.148	0.148
	Females	11.765	14.706	0.078	2.373	12.941	8.824	0.512	0.801	10.458	12.418	0.118	0.471
Sadness	Males	4.321	3.704	6.722	8.000	8.148	8.889	1.067	0.600	8.642	11.111	1.333	0.000
	Females	2.941	0.490	12.255	20.712 *	14.706	12.941	1.977	0.512	7.516	8.824	3.559	1.441
Affection	Males	14.198	15.432	1.398	2.722	11.852	7.407	0.067	1.667	6.173	7.407	5.333	3.000
	Females	10.784	9.804	0.020	0.314	12.941	19.412	0.512	10.542	5.882	3.595	7.529	15.559 †
Boredom	Males	17.284	14.815	5.556	2.000	8.889	12.596	0.600	0.267	12.757	6.173	0.593	5.333
	Females	12.255	15.196	0.240	3.064	4.118	6.471	7.483	3.295	6.863	8.497	4.971	1.882
Interest	Males	32.099	36.420	64.222 ‡	93.389 ‡	2.963	2.222	8.067	9.600	12.757	18.519	0.593	12.000
	Females	35.294	36.275	107.373 ‡	116.255 ‡	4.706	2.353	6.277	11.736	21.895	19.608	32.029 ‡	19.884 *
Relaxation	Males	10.494	12.346	0.056	0.222	14.074	11.852	1.067	0.067	5.340	3.704	7.259	12.000
	Females	15.196	15.196	3.064	3.064	9.412	8.824	0.442	0.801	4.248	1.961	12.971	23.059 ‡

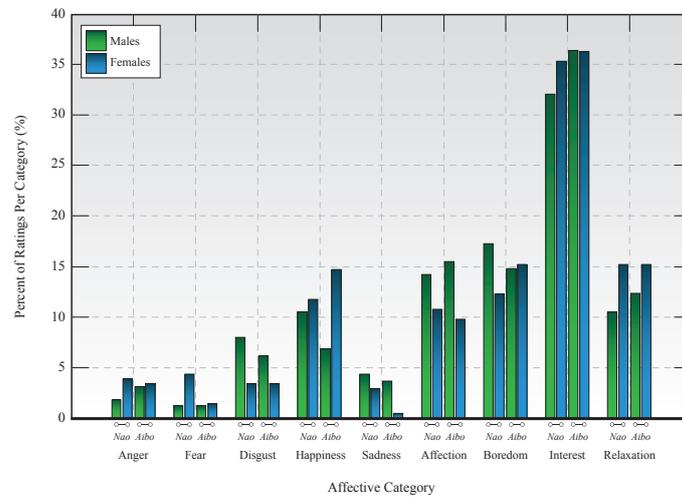
† : $p < 0.05$

* : $p < 0.025$

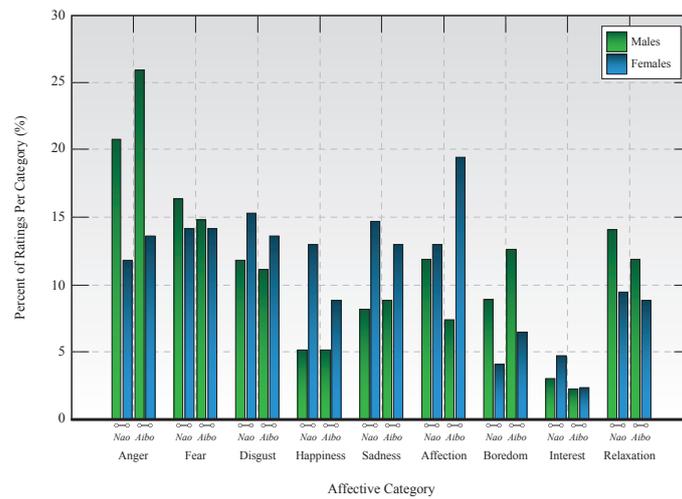
★ : $p < 0.01$

‡ : $p < 0.005$

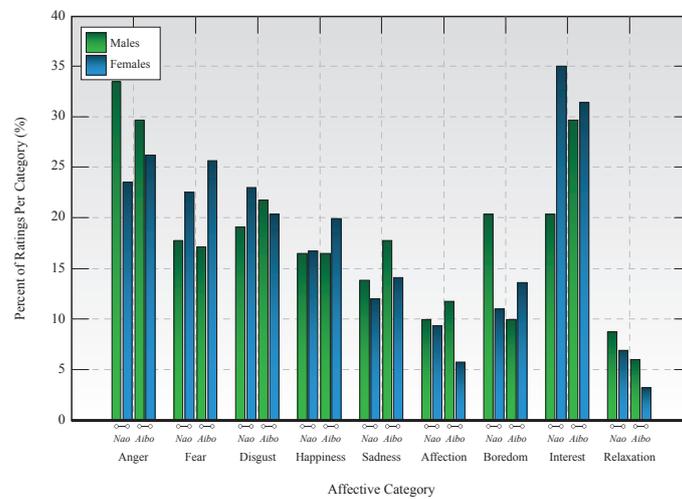
Figure B.1: Overall percentage of the Affective Ratings across both the two robot and subject gender. This is shown for each of the three utterance categories.



(a) Human Utterances

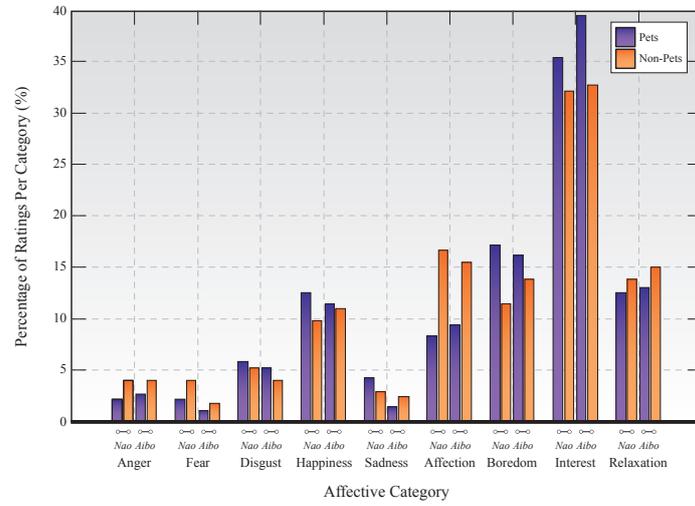


(b) Animal Utterances

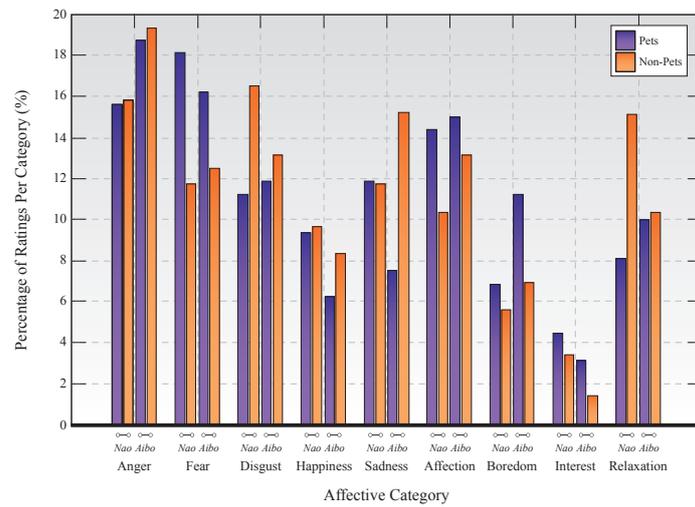


(c) Technological Utterances

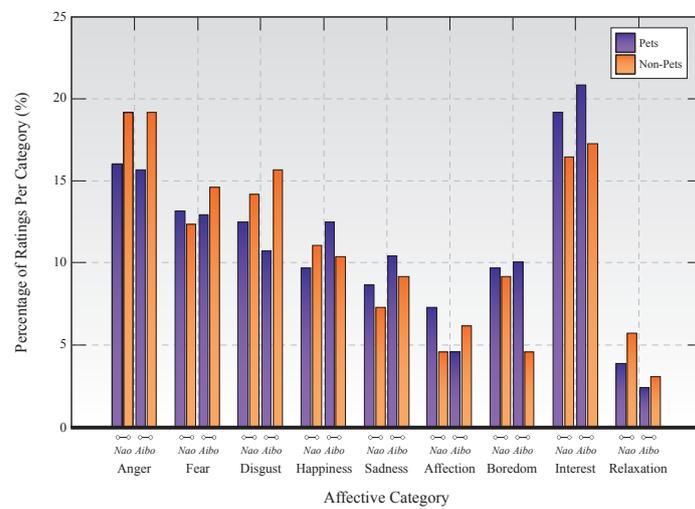
Figure B.2: Overall percentage of the Affective Ratings across both the two robot and pet/non-pet owners. This is shown for each of the three utterance categories.



(a) Human Utterances



(b) Animal Utterances



(c) Technological Utterances

Table B.3: Pet Ownership differences for Affective Ratings across the Utterance Classes and Robot Morphologies. The table shows the percentages of each affective ratings for each robot across the two genders and the $\chi^2(8)$ values indicating whether this rating is above chance (compared to a flat uniform distribution).

Emotion	Subject	Human				Animal				Technological			
		Rating (%)		χ^2		Rating (%)		χ^2		Rating (%)		χ^2	
		Nao	Aibo	Nao	Aibo	Nao	Aibo	Nao	Aibo	Nao	Aibo	Nao	Aibo
Anger	Pet	2.083	1.042	14.083	17.521 †	15.625	18.750	2.934	8.403	15.972	15.625	6.125	5.281
	Non-Pet	4.023	1.724	7.868	13.799	15.862	19.310	2.946	8.773	19.157	19.157	15.207	15.207
Fear	Pet	2.083	1.042	14.083	17.521 †	18.125	16.250	7.084	3.803	13.194	12.847	1.125	0.781
	Non-Pet	4.023	1.724	7.868	13.799	11.724	12.414	0.049	0.222	12.261	14.560	0.310	2.793
Disgust	Pet	5.729	5.208	5.005	6.021	11.250	11.875	0.003	0.084	12.500	10.764	0.500	0.031
	Non-Pet	5.172	4.023	5.523	7.868	16.552	13.103	3.863	0.518	14.176	15.709	2.207	4.966
Happiness	Pet	12.500	11.458	0.333	0.021	9.375	6.250	0.434	3.403	9.722	12.500	0.500	0.500
	Non-Pet	9.770	10.920	0.282	0.006	9.655	8.276	0.277	1.049	11.111	10.345	0.000	0.128
Sadness	Pet	4.167	1.563	8.333	15.755 †	11.875	7.500	0.084	1.878	8.681	10.417	1.531	0.125
	Non-Pet	2.874	2.299	10.626	12.161	11.724	15.172	0.049	2.153	7.280	9.195	3.448	0.862
Affection	Pet	8.333	9.375	1.333	0.521	14.375	15.000	1.534	2.178	7.292	4.514	3.781	11.281
	Non-Pet	16.667	15.517	4.833	3.040	10.345	13.103	0.077	0.518	4.598	6.130	9.996	5.828
Boredom	Pet	17.188	16.146	6.380	4.380	6.875	11.250	2.584	0.003	9.722	10.069	0.500	0.281
	Non-Pet	11.494	13.793	0.023	1.124	5.517	6.897	4.084	2.318	9.195	4.598	0.862	9.966
Interest	Pet	35.417	39.583	102.083 ‡	140.083 ‡	4.375	3.125	6.534	9.184	19.097	20.833	16.531 †	24.500 ‡
	Non-Pet	32.184	32.759	69.540 ‡	73.385 ‡	3.448	1.379	7.663	12.359	16.475	17.241	6.759	8.828
Relaxation	Pet	12.500	13.021	0.333	0.630	8.125	10.000	1.284	0.1778	3.819	2.431	13.781	19.531 *
	Non-Pet	13.793	14.943	1.126	2.299	15.172	10.345	2.153	0.077	5.747	3.065	6.759	15.207

† : $p < 0.05$

* : $p < 0.025$

★ : $p < 0.01$

‡ : $p < 0.005$

Table B.4: Ratings and χ^2 values for Intentional Ratings across both robot types and sound categories. The columns for each utterance class are the overall percentage of ratings for a given intentional label, the results of a χ^2 test checking that that rating is above chance, and the results of a Stewart-Maxwell test indicating whether there was an overall difference in the distribution of ratings due to the robot image that was also presented

Intention	Subject	Human			Animal			Techno		
		Rating (%)	χ^2	S&M (χ^2)	Rating (%)	χ^2	S&M (χ^2)	Rating (%)	χ^2	S&M (χ^2)
Neutral	Nao	23.771	2.602		18.361	0.410		21.129	0.350	
	Aibo	21.038	0.197		14.098	5.312		22.587	1.836	
Attention	Nao	38.525	62.79 †		26.885	7.230		33.698	51.503 †	
	Aibo	41.530	84.828 †		30.164	15.754 †		32.969	46.170 †	
Comfort	Nao	19.399	0.066	3.841	23.607	1.984	3.762	9.472	30.427 †	0.710
	Aibo	16.393	2.380		24.262	2.771		10.018	27.350 †	
Approval	Nao	12.842	9.378		7.869	22.443 †		12.022	17.472 †	
	Aibo	14.481	5.574		6.557	22.557 †		11.111	21.689 †	
Prohibition	Nao	5.465	38.665 †		23.279	1.639		23.679	3.716	
	Aibo	6.557	33.069 †		24.918	3.689		23.315	3.017	

† : $p < 0.05$

Table B.5: Gender differences for the Intention ratings. The table shows the percentages of each affective ratings for each robot across the two genders and the $\chi^2(8)$ values indicating whether this rating is above chance (compared to a flat uniform distribution).

Emotion	Subject	Human				Animal				Technological			
		Rating (%)		χ^2		Rating (%)		χ^2		Rating (%)		χ^2	
		Nao	Aibo	Nao	Aibo	Nao	Aibo	Nao	Aibo	Nao	Aibo	Nao	Aibo
Neutral	Males	18.519	18.519	0.178	0.178	19.259	14.074	0.037	2.370	18.107	24.280	0.435	2.226
	Females	27.941	23.039	6.432	0.942	17.647	14.118	0.471	2.941	23.529	21.242	1.906	0.2359
Attention	Males	38.889	44.444	28.900 ‡	48.400 ‡	25.926	27.407	2.370	3.704	33.745	29.630	22.954 ‡	11.267 *
	Females	38.235	39.216	33.918 ‡	37.663 ‡	27.647	32.353	4.971	12.971 *	33.660	35.621	28.550 ‡	37.334 ‡
Comfort	Males	20.988	16.667	0.079	0.900	18.519	20.471	0.148	0.037	11.111	11.111	9.600 †	9.600 †
	Females	18.137	16.177	0.354	1.491	27.647	27.059	4.971	4.235	8.170	9.150	21.412 ‡	18.011 ‡
Approval	Males	17.284	12.346	0.598	4.746	8.148	3.704	9.482	17.926 ‡	10.700	11.523	10.510 †	8.732 †
	Females	9.314	16.177	11.648 *	1.491	7.647	8.824	12.971 *	10.618 †	13.072	10.784	7.344	12.994 *
Prohibition	Males	4.321	8.025	19.912 ‡	11.616 *	28.148	34.074	4.482	13.370 *	26.337	23.457	4.880	1.452
	Females	6.373	5.392	18.942 ‡	21.766 ‡	19.412	17.647	0.029	0.471	21.569	23.203	0.377	1.569

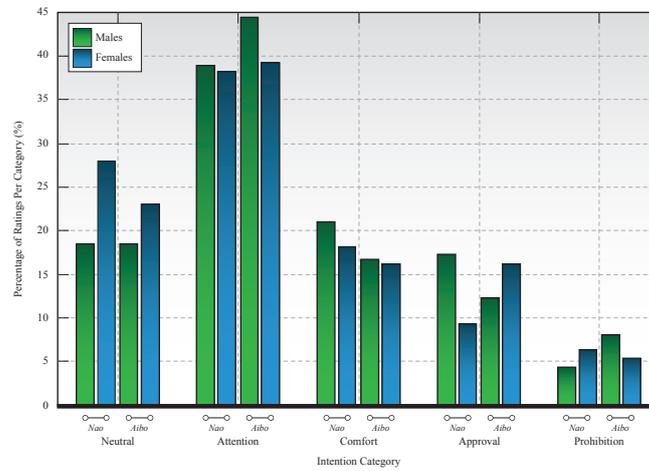
† : $p < 0.05$

* : $p < 0.025$

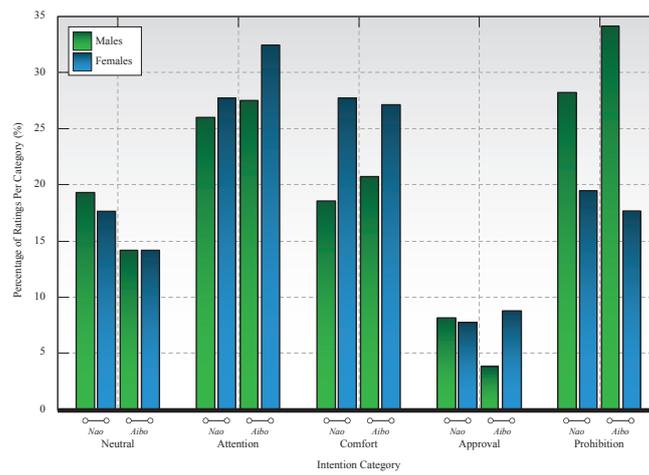
★ : $p < 0.01$

‡ : $p < 0.005$

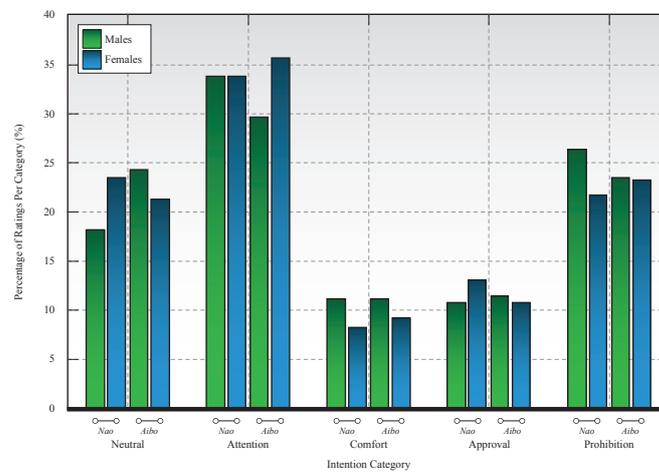
Figure B.3: Overall percentage of the Intention Ratings across both the two robot and subject gender. This is shown for each of the three utterance categories.



(a) Human Utterances



(b) Animal Utterances



(c) Technological Utterances

Table B.6: Pet Ownership differences for the Intention ratings. The table shows the percentages of each affective ratings for each robot across the two genders and the $\chi^2(8)$ values indicating whether this rating is above chance (compared to a flat uniform distribution).

Emotion	Subject	Human				Animal				Technological			
		Rating (%)		χ^2		Rating (%)		χ^2		Rating (%)		χ^2	
		Nao	Aibo	Nao	Aibo	Nao	Aibo	Nao	Aibo	Nao	Aibo	Nao	Aibo
Neutral	Pet	25.521	23.958	2.926	1.504	18.125	15.625	0.281	1.531	25.347	23.958	4.117	2.256
	Non-Pet	21.839	17.816	0.294	0.415	18.621	12.414	0.138	4.172	16.475	21.073	1.622	0.150
Attention	Pet	41.667	43.750	45.067 ‡	54.150 ‡	28.750	29.375	6.125	7.031	29.514	34.375	13.034 *	29.756 ‡
	Non-Pet	35.058	39.081	19.725 ‡	31.674 ‡	24.828	31.035	1.690	8.828	38.314	31.418	43.771 ‡	17.012 ‡
Comfort	Pet	18.229	16.146	0.301	1.426	26.875	24.375	3.751	1.531	11.458	10.417	10.506 *	13.225 *
	Non-Pet	20.690	16.667	0.041	0.967	20.000	24.138	0.000	1.241	7.2797	9.579	21.116 ‡	14.173 *
Approval	Pet	8.333	9.375	13.067 *	10.838 ‡	3.750	4.375	21.125 ‡	19.531 ‡	10.417	9.722	13.225 *	15.211 ‡
	Non-Pet	17.816	10.115	0.4149	0.001	12.414	8.966	4.172	8.828	13.793	13.644	5.028	7.062
Prohibition	Pet	6.250	6.771	18.150 ‡	16.801 ‡	22.500	26.250	0.500	3.125	23.264	21.528	1.534	0.336
	Non-Pet	4.598	6.322	20.639 ‡	16.277 ‡	24.138	23.448	1.241	0.862	24.138	25.287	2.235	3.648

‡ : $p < 0.05$

* : $p < 0.025$

★ : $p < 0.01$

‡ : $p < 0.005$

Table B.7: Appropriateness ratings for utterances and associated χ^2 values for the one-way χ^2 test comparing the ratings a flat uniform distribution, and the χ^2 values of the McNemar tests checking whether the difference in the rating distributions across the two robots are significantly different.

Utterance Class	Robot	“Yes” Ratings (%)	Chi Squared Text (χ^2)	McNemar Test (χ^2)
Human	Nao	74.590	44.262 ‡	27.480 ‡
	Aibo	57.924	4.596 †	
Animal	Nao	46.230	0.867	34.299 ‡
	Aibo	68.197	20.198 ‡	
Techno	Nao	59.745	10.427 ‡	4.198 †
	Aibo	54.827	2.558	

† : $p < 0.05$

‡ : $p < 0.005$

Table B.8: Appropriateness ratings for utterances and associated χ^2 values for the one-way χ^2 test comparing the ratings a flat uniform distribution, and the χ^2 values of the two-way tests checking whether the difference in the rating distributions between the genders are significantly different.

Utterance Class	Subject Group	“Yes” Rating (%)		1-way χ^2		2-way χ^2	
		Nao	Aibo	Nao	Aibo	Nao	Aibo
Human	Males	73.457	56.790	17.827 ‡	1.494	0.197	0.153
	Females	75.490	58.824	23.152 ‡	2.526		
Animal	Males	39.259	67.407	3.115	8.182 ‡	4.734 †	0.070
	Females	51.765	68.824	1.269	10.552 ‡		
Techno	Males	55.144	47.325	1.286	0.348	3.847 †	9.907 ‡
	Females	63.399	60.784	7.235 *	4.499 †		

† : $p < 0.05$

* : $p < 0.025$

★ : $p < 0.01$

‡ : $p < 0.005$

Table B.9: Appropriateness ratings for utterances and associated χ^2 values for the one-way χ^2 test comparing the ratings a flat uniform distribution, and the χ^2 values of the two-way tests checking whether the difference in the rating distributions between pet/non-pet owners are significantly different.

Utterance Class	Subject Group	“Yes” Rating (%)		1-way χ^2		2-way χ^2	
		Nao	Aibo	Nao	Aibo	Nao	Aibo
Human	Pet	75.000	55.729	24.000 ‡	1.260	0.036	0.798
	Non-Pet	74.138	60.345	22.321 ‡	2.371		
Animal	Pet	39.375	64.375	3.613	6.613 *	6.36 *	2.266
	Non-Pet	53.793	72.414	2.172	10.200 ‡		
Techno	Pet	59.722	54.514	5.444 *	1.174	0.000	0.024
	Non-Pet	59.770	55.172	5.236 *	1.274		

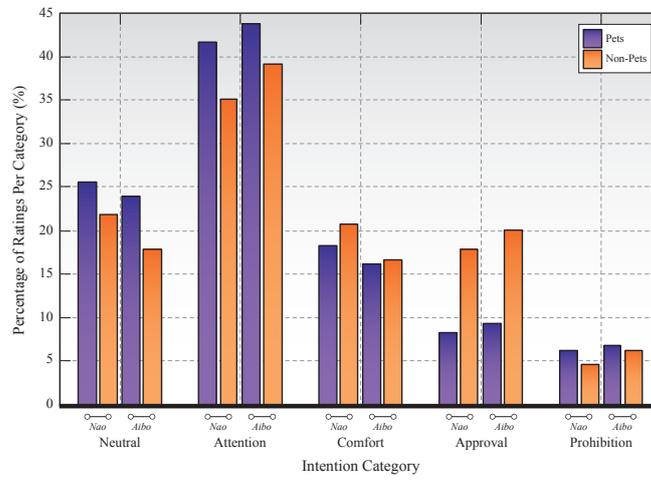
† : $p < 0.05$

* : $p < 0.025$

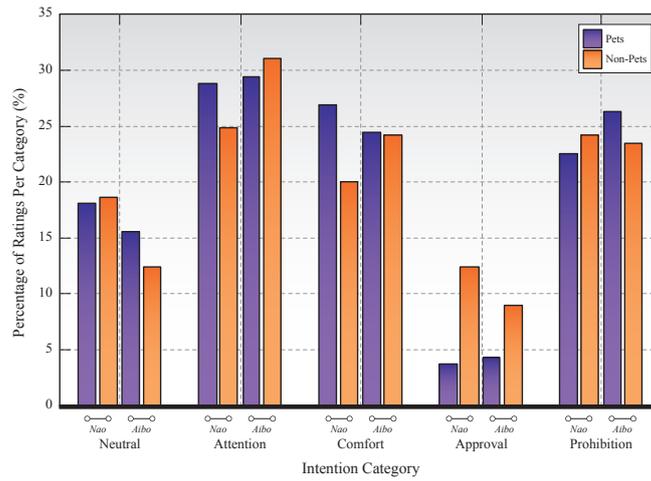
★ : $p < 0.01$

‡ : $p < 0.005$

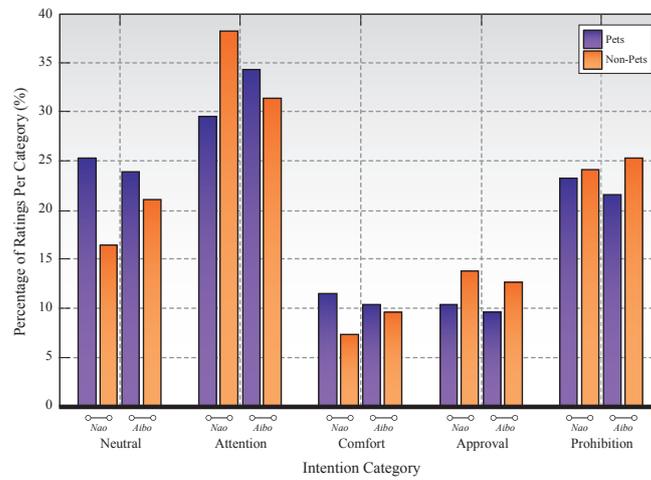
Figure B.4: Overall percentage of the Intention Ratings across both the two robot and pet/non-pet owners. This is shown for each of the three utterance categories.



(a) Human Utterances



(b) Animal Utterances



(c) Technological Utterances

Appendix C

Collecting Training Data for Machine Learning

C.1 Experiment #1 - Tables of Results

Table C.1: Results of the Kruskal-Wallice tests testing the difference in affective ratings due to the Tremolo values in Experiment #1.

Affect Dimension	d.f.	χ^2	p value
Pleasure	1	0.7	0.405
Arousal	1	0.67	0.412
Dominance	1	0.5	0.477

Table C.2: Descriptive statistics for the Box Plots shown in figure 5.4, showing the ratings for the Utterances with a Tremolo value of 0, grouped by the first pitch contour shape. The table shows the Median values, 1st and 3rd Quartiles, Inter-Quartile Range, and the Lower and Upper estimated 95% Median Confidence Intervals.

Dimension	Contour	Median	Q_1	Q_3	IQR	95% CI	
						Lower	Upper
Dominance	Flat	-0.423	-0.942	0.730	1.673	-0.874	0.027
	Rising	0.727	-0.276	0.910	1.186	0.387	1.0
	Falling	0.096	-0.473	0.847	1.319	-0.249	0.441

C.2 Experiment #2 - Tables of Results

Table C.3: Results of the Kruskal-Wallice tests checking for differences in the affective ratings across the two different sound unit counts in Experiment #2.

Affect Dimension	d.f.	χ^2	p value
Pleasure	1	2.03	0.154
Arousal	1	0.22	0.641
Dominance	1	1.04	0.308

Table C.4: Descriptive statistics for the Box Plots shown in figure 5.5, showing the ratings for the Utterances with a Rhythm value of 0, 0.5 and 1. The table shows the Median values, 1st and 3rd Quartiles, Inter-Quartile Range and the Lower and Upper estimated 95% Median Confidence Intervals.

Dimension	Rhythm	Median	Q_1	Q_3	IQR	95% CI	
						Lower	Upper
Arousal	0	0.433	-0.925	0.974	1.899	-0.007	0.873
	0.5	0.761	0.122	1.000	0.878	0.556	0.967
	1	0.788	0.548	1.000	0.452	0.643	0.933

C.3 Experiment #3 - Tables of Results

Table C.5: Results of the Kruskal-Wallice tests, collapsing the Frequency Range and testing only for the influence of the Base Frequency across the different sound unit counts, along each affective dimensions.

Affect Dimension	Unit Count	<i>d.f.</i>	χ^2	<i>p</i> value
Pleasure	3	2	2.448	0.294
Arousal		2	0.809	0.667
Dominance		2	1.468	0.480
Pleasure	5	2	0.815	0.665
Arousal		2	0.335	0.846
Dominance		2	2.667	0.264
Pleasure	Both	2	1.924	0.382
Arousal		2	0.973	0.615
Dominance		2	2.269	0.322

Table C.6: Results of the Kruskal-Wallice tests with interleaved Base Frequency and Frequency Range values, across the different sound unit counts, along each affective dimensions.

Affect Dimension	Unit Count	<i>d.f.</i>	χ^2	<i>p</i> value
Pleasure	3	8	8.878	0.353
Arousal		8	4.134	0.845
Dominance		8	10.855	0.210
Pleasure	5	8	13.677	0.091
Arousal		8	2.785	0.947
Dominance		8	9.452	0.306
Pleasure	Both	8	14.351	0.073
Arousal		8	3.395	0.907
Dominance		8	8.526	0.384

Table C.7: Descriptive statistics for the Box plots in figure 5.6, showing the ratings for all Utterances with 3 Sound Units, grouped by the Frequency Range values, along the Dominance dimension. Table shows the Median values, 1st and 3rd Quartiles, Inter-Quartile Range, and the lower and upper estimated Median Confidence Intervals.

Dimension	Base Freq	Median	Q_1	Q_3	IQR	95% CI	
						Lower	Upper
Dominance	500 Hz	-0.655	-0.924	-0.068	0.856	-0.889	-0.421
	1000 Hz	-0.020	-0.735	0.890	1.625	-0.434	0.394
	1500 Hz	0.010	-0.747	0.878	1.625	-0.472	0.492

Table C.8: Descriptive statistics for the Box plots in figure 5.7, showing the ratings for all Utterances with 5 Sound Units, grouped by the Frequency Range values, along the Pleasure dimension. Table shows the Median values, 1st and 3rd Quartiles, Inter-Quartile Range, and the lower and upper estimated Median Confidence Intervals.

Dimension	Freq Range	Median	Q_1	Q_3	IQR	95% CI	
						Lower	Upper
Pleasure	500 Hz	0.068	-0.907	0.702	1.609	-0.372	0.508
	1000 HZ	-0.178	-0.783	0.687	1.469	-0.543	0.187
	1500 Hz	0.623	-0.027	0.958	0.985	0.336	0.910

Table C.9: Descriptive Statistics for the Box Plots in figure 5.8, showing the ratings for utterances across two two difference Sound Unit Counts, for each affective dimension. Table shows the Median values, 1st and 3rd Quartiles, Inter-Quartile Range, and the lower and upper estimated Median Confidence Intervals.

Dimension	Unit Count	Median	Q_1	Q_3	IQR	95% CI	
						Lower	Upper
Pleasure	3	0.056	-0.869	0.860	0.173	-0.217	0.328
	5	0.074	-0.791	0.835	1.625	-0.179	0.327
Arousal	3	0.619	-0.361	0.934	1.295	0.415	0.824
	5	0.699	-0.587	1.0	1.587	0.452	0.946
Dominance	3	-0.335	-0.824	0.854	1.678	-0.600	0.071
	5	0.042	-0.671	0.942	1.613	-0.209	0.293

C.4 Experiment #4 - Tables of Results

Table C.10: Results of the Kruskal-Wallice tests, collapsing the Speech Rate and testing only for the influence of the Pause Ratio across the difference sound unit counts, along each affective dimension.

Affect Dimension	Unit Count	<i>d.f.</i>	χ^2	<i>p</i> value
Pleasure	3	2	0.94	0.625
Arousal		2	2.03	0.362
Dominance		2	3.42	0.181
Pleasure	5	2	0.03	0.984
Arousal		2	0.29	0.865
Dominance		2	1.48	0.476
Pleasure	Both	2	0.49	0.784
Arousal		2	1.52	0.468
Dominance		2	1.12	0.571

Table C.11: Results of the Kruskal-Wallice tests, collapsing the Pause Ratio and testing only for the influence of the Speech Rate across the difference sound unit counts, along each affective dimension.

Affect Dimension	Unit Count	<i>d.f.</i>	χ^2	<i>p</i> value
Pleasure	3	2	0.33	0.849
Arousal		2	2.15	0.341
Dominance		2	0.37	0.832
Pleasure	5	2	3.85	0.146
Arousal		2	2.53	0.283
Dominance		2	1.78	0.411
Pleasure	Both	2	3.24	0.198
Arousal		2	0.43	0.808
Dominance		2	0.44	0.801

Table C.12: Results of the Kruskal-Wallice tests, interleaving the Pause Ratio and Speech Rate, across the different sound unit counts, along each affective dimension.

Affect Dimension	Unit Count	<i>d.f.</i>	χ^2	<i>p</i> value
Pleasure	3	8	3.46	0.902
Arousal		8	4.73	0.786
Dominance		8	6.05	0.642
Pleasure	5	8	4.17	0.841
Arousal		8	4.72	0.787
Dominance		8	5.26	0.730
Pleasure	Both	8	5.31	0.724
Arousal		8	3.35	0.910
Dominance		8	3.55	0.896

Table C.13: Descriptive Statistics for the Box Plots in figure 5.9. Table shows the Median values, 1st and 3rd Quartiles, Inter-Quartile Range, and the lower and upper estimated Median Confidence Intervals.

Dimension	Unit Count	Median	Q_1	Q_3	IQR	95% CI	
						Lower	Upper
Pleasure	3	0.128	-0.855	0.846	1.701	-0.135	0.391
	5	0.238	-0.605	0.915	1.519	0.004	0.472
Arousal	3	0.690	-0.170	1.0	1.170	0.509	0.871
	5	0.779	-0.454	1.0	1.454	0.555	1.0
Dominance	3	0.180	-0.809	0.903	1.713	-0.085	0.445
	5	0.611	-0.421	0.984	1.405	0.395	0.827

C.5 Experiment #5 - Tables of Results

Table C.14: Descriptive Statistics for the Box Plots in figure 5.10. Table shows the Median values, 1st and 3rd Quartiles, Inter-Quartile Range, and the lower and upper estimated Median Confidence Intervals.

Dimension	PC	Median	Q_1	Q_3	IQR	95% CI	
						Lower	Upper
Pleasure	1	0.069	-0.771	0.886	1.657	-0.178	0.318
	2	0.064	-0.843	0.896	1.739	-0.194	0.322
Arousal	1	0.823	-0.126	1.0	1.126	0.655	0.992
	2	0.673	-0.437	1.0	1.437	0.459	0.886
Dominance	1	-0.246	-0.870	0.914	1.785	-0.513	0.022
	2	0.393	-0.529	0.902	1.431	0.181	0.606

Table C.15: Results of the Kruskal-Wallice tests checking for differences in ratings, along each affective dimension, between PC2 and PC3 in Experiment #5.

Affect Dimension	Unit Count	$d.f.$	χ^2	p value
Pleasure	3	1	3.63	0.057
Arousal		1	1.41	0.235
Dominance		1	0.28	0.594
Pleasure	5	1	1.15	0.284
Arousal		1	1.53	0.217
Dominance		1	0	0.951
Pleasure	Both	1	0.27	0.603
Arousal		1	0.25	0.616
Dominance		1	0.01	0.939

Appendix D

Categorical Perception of NLUs

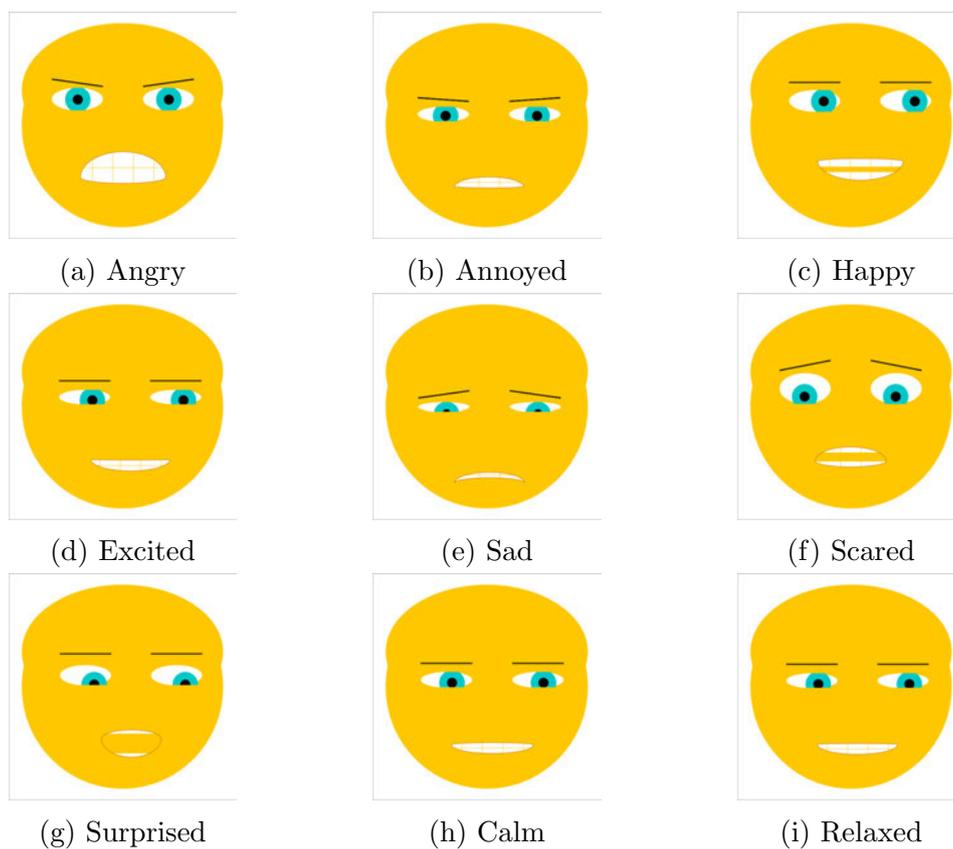


Figure D.1: AffectButton facial gestures of the mean ratings given by the adult subjects for each of the affective labels in the Labelling Task.

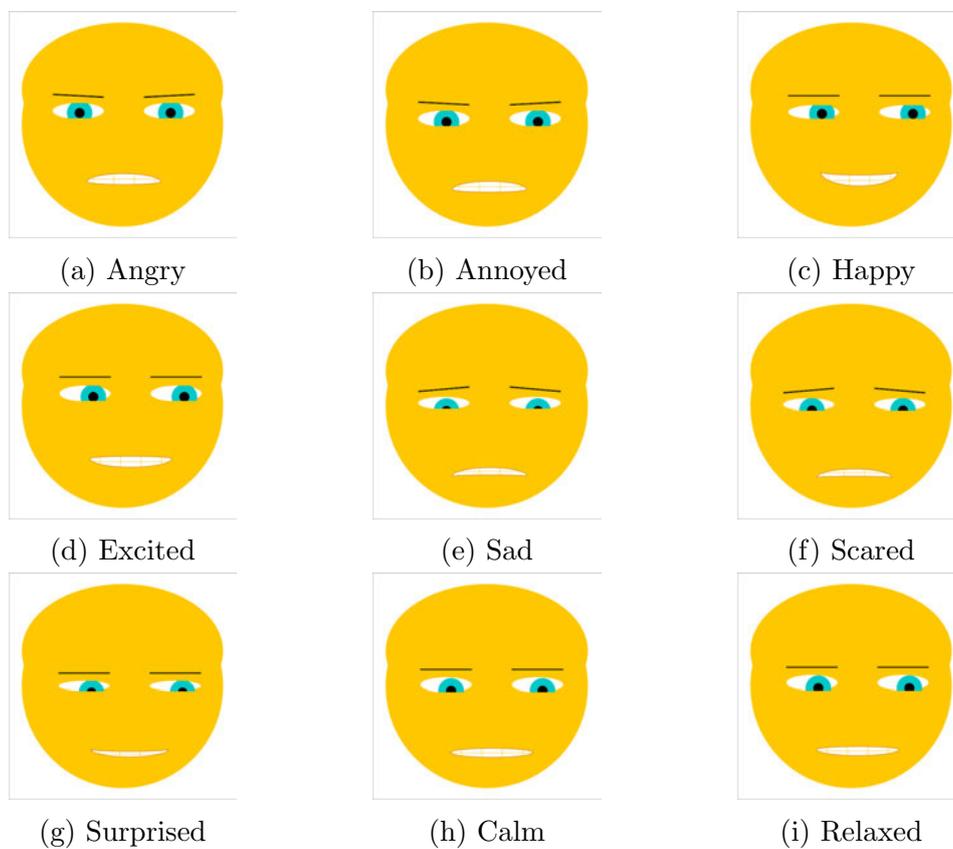


Figure D.2: AffectButton facial gestures of the mean ratings given by the children subjects for each of the affective labels in the Labelling Task.

Table D.1: Mean values and Standard Deviations of the ratings for each of the affective labels presented in the Labelling Task. The table shows the values for the adults and children, along each affective dimension of the AffectButton.

Subjects	Label	Affect Space Dimension					
		Pleasure		Arousal		Dominance	
		Mean	Std	Mean	Std	Mean	Std
Adults	Angry	-0.8096	0.2426	0.7981	0.5025	0.7541	0.3343
	Annoyed	-0.8096	0.2426	0.7981	0.5025	0.7541	0.3343
	Happy	0.7712	0.2260	0.8301	0.3211	0.7613	0.3560
	Excited	0.5511	0.4672	0.7391	0.4819	0.2800	0.7449
	Sad	-0.3907	0.3010	-0.5661	0.6868	-0.5229	0.1665
	Scared	-0.9343	0.2272	0.7385	0.6519	-0.5793	0.4908
	Surprised	0.6506	0.4601	0.9221	0.1359	-0.8251	0.2099
	Calm	0.1022	0.3628	-0.5639	0.5274	0.3051	0.3588
	Relaxed	0.2976	0.4131	-0.2938	0.7266	0.2109	0.4131
Children	Angry	-0.6207	0.5325	0.3010	0.9468	0.4406	0.8665
	Annoyed	-0.3933	0.6543	0.0176	0.8344	0.1994	0.7765
	Happy	0.5578	0.3360	0.2577	0.8661	0.5376	0.4675
	Excited	0.6266	0.2724	0.2964	0.8403	0.2234	0.6479
	Sad	-0.3444	0.4747	-0.3041	0.7549	-0.5997	0.3309
	Scared	-0.3076	0.4115	0.1838	0.8209	-0.2437	0.6608
	Surprised	0.4174	0.4534	0.4950	0.7017	-0.3803	0.6558
	Calm	0.0315	0.6203	-0.1941	0.7548	-0.2581	0.6076
	Relaxed	0.2306	0.5651	-0.2257	0.7869	-0.0841	0.6811

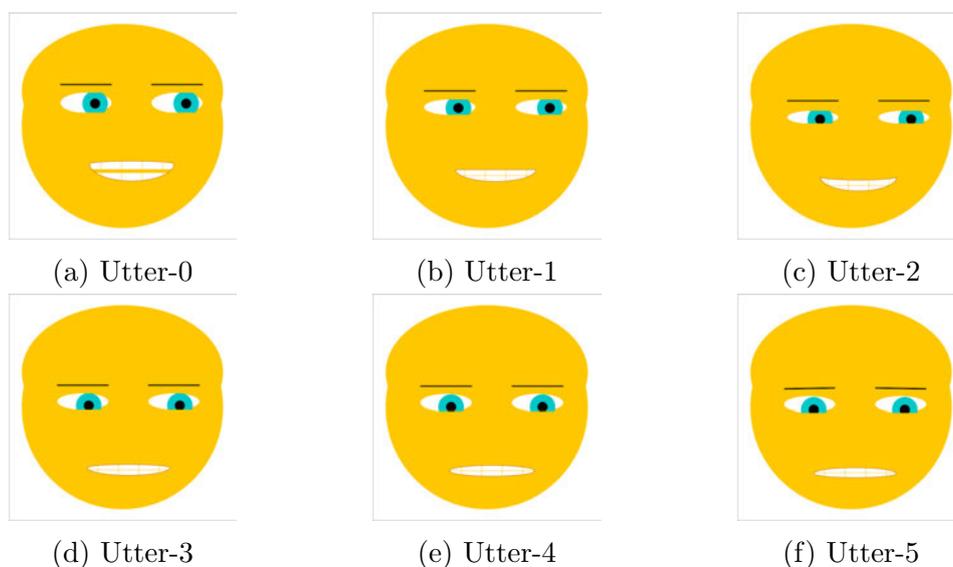


Figure D.3: Mean AffectButton facial gestures for both the Stimulus Set ratings provided by the adult subjects during the Identification Task.

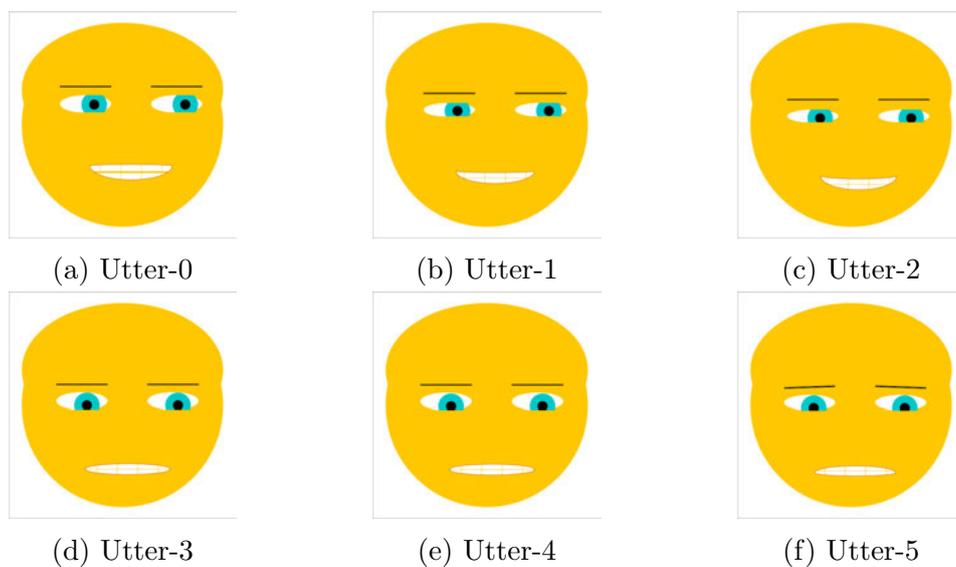


Figure D.4: Mean AffectButton facial gestures for the Stimulus Set 1 ratings provided by the adult subjects during the Identification Task.

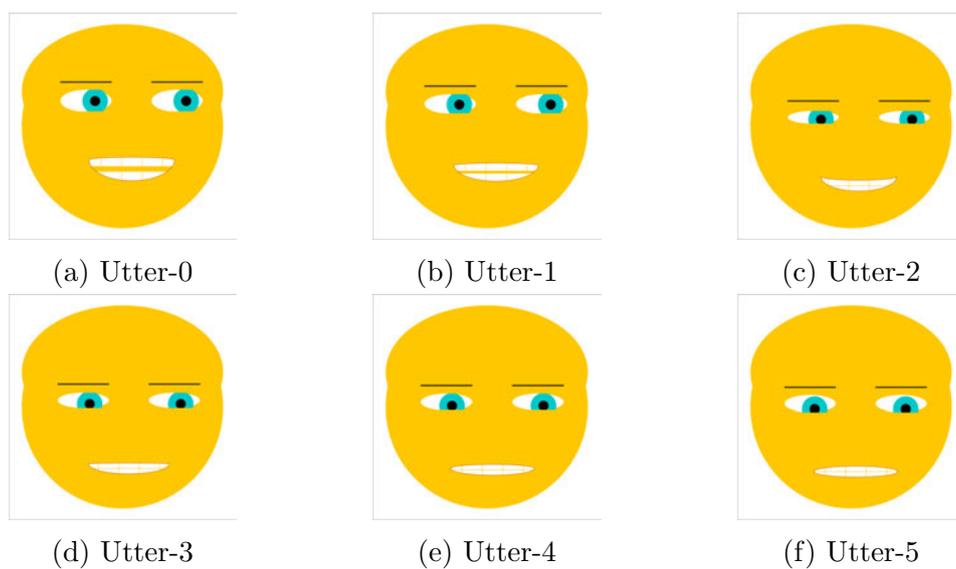


Figure D.5: Mean AffectButton facial gestures for the Stimulus Set 2 ratings provided by the adult subjects during the Identification Task.

Table D.2: Mean values and 95% confidence intervals for the overall adult ratings of stimuli in the Identification Task. The table shows the values for the overall different Utterance Parameter configurations, as well as those for Stimulus Set 1 and Stimulus Set 2.

Set	Utter	Mean	Std Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Both	0	0.748	0.055	0.635	0.862
	1	0.673	0.051	0.567	0.779
	2	0.473	0.043	0.385	0.561
	3	0.163	0.053	0.053	0.273
	4	0.061	0.054	-0.051	0.173
	5	-0.115	0.055	-0.229	-0.002
1	0	0.711	0.063	0.580	0.841
	1	0.614	0.073	0.464	0.764
	2	0.456	0.052	0.349	0.562
	3	0.043	0.073	-0.108	0.193
	4	-0.005	0.068	-0.146	0.135
	5	-0.141	0.072	-0.288	0.007
2	0	0.786	0.059	0.664	0.908
	1	0.732	0.055	0.619	0.845
	2	0.490	0.065	0.357	0.624
	3	0.284	0.065	0.151	0.417
	4	0.127	0.079	-0.035	0.290
	5	-0.090	0.065	-0.223	0.043

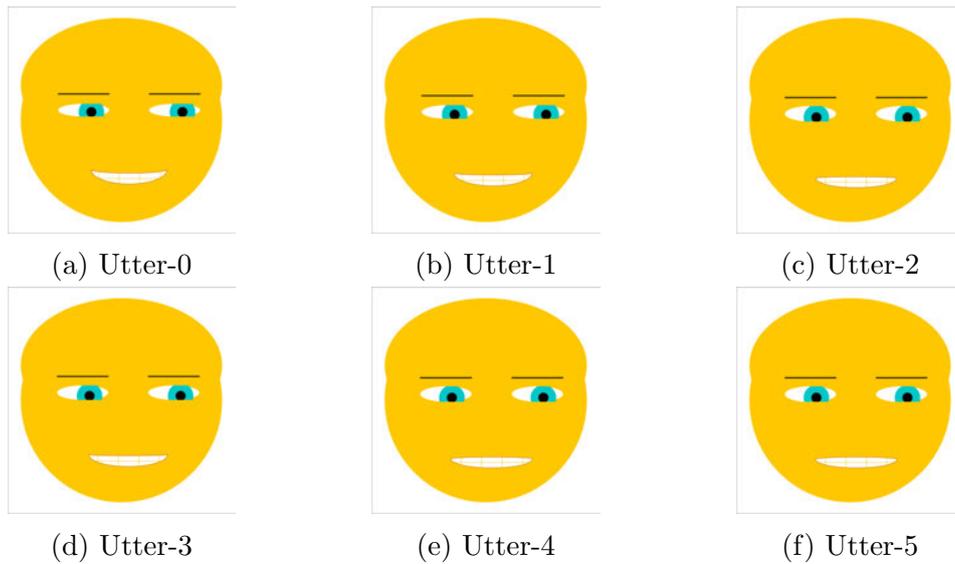


Figure D.6: Mean AffectButton facial gestures for both the Stimulus Set ratings provided by the child subjects during the Identification Task.

Table D.3: Mean values and 95% confidence intervals for the overall adults ratings of stimuli in the Identification Task. The table shows the values for the overall different Utterance Parameter configurations, as well as those for the two genders.

Subjects	Utter	Mean	Std Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Both	0	0.748	0.055	0.635	0.862
	1	0.673	0.051	0.567	0.779
	2	0.473	0.043	0.385	0.561
	3	0.163	0.053	0.053	0.273
	4	0.061	0.054	-0.051	0.173
	5	-0.115	0.055	-0.229	-0.002
Females	0	0.749	0.071	0.604	0.894
	1	0.715	0.066	0.579	0.850
	2	0.573	0.054	0.461	0.685
	3	0.235	0.068	0.095	0.376
	4	0.139	0.069	-0.004	0.282
	5	-0.005	0.070	-0.150	0.140
Males	1	0.747	0.085	0.572	0.923
	2	0.631	0.079	0.468	0.794
	3	0.373	0.066	0.238	0.508
	4	0.091	0.082	-0.078	0.260
	5	-0.017	0.084	-0.189	0.155
	6	-0.226	0.085	-0.401	-0.051

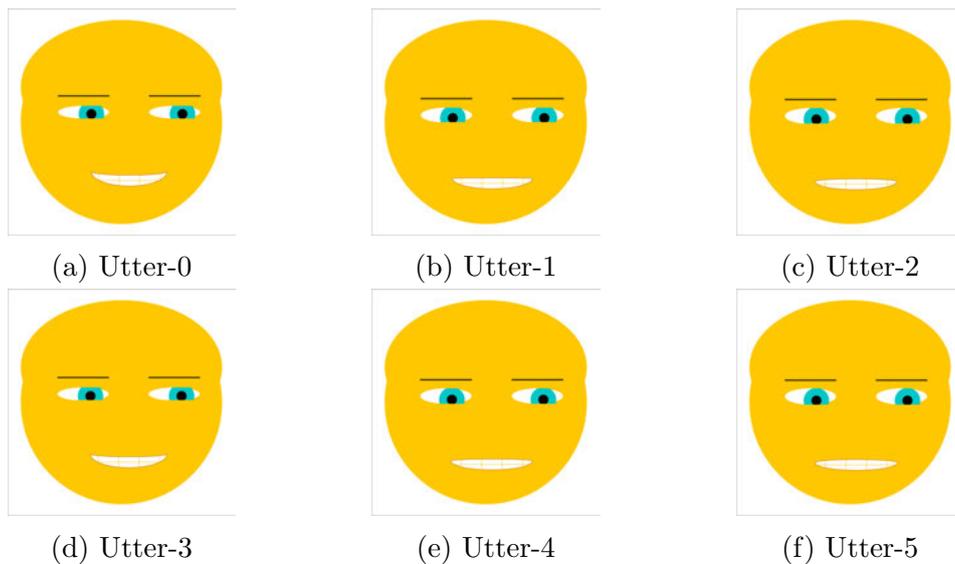


Figure D.7: Mean AffectButton facial gestures for the Stimulus Set 1 ratings provided by the child subjects during the Identification Task.

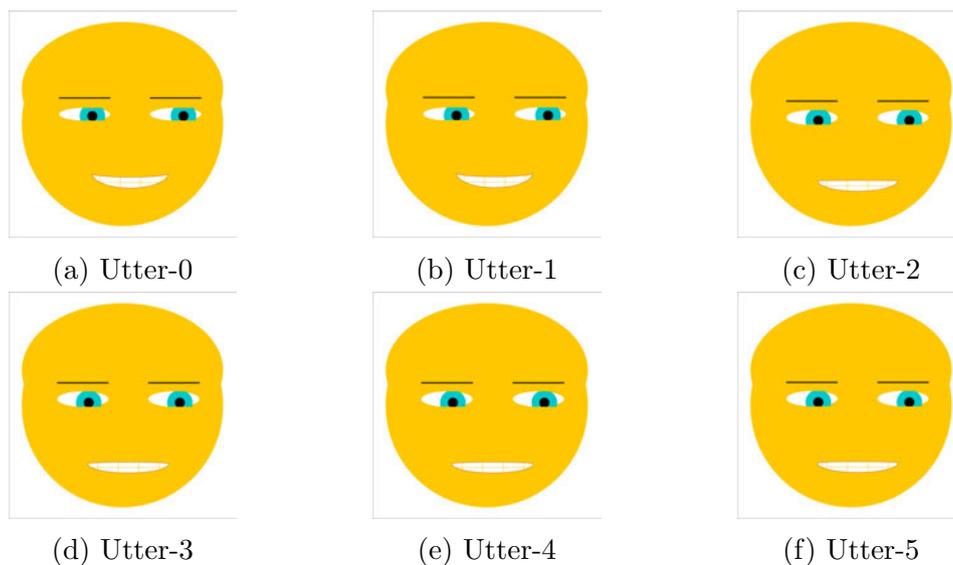


Figure D.8: Mean AffectButton facial gestures for the Stimulus Set 2 ratings provided by the child subjects during the Identification Task.

Table D.4: Mean values and 95% confidence intervals for the overall child ratings of stimuli in the Identification Task. The table shows the values for the overall different Utterance Parameter configurations, as well as those for Stimulus Set 1 and Stimulus Set 2.

Set	Utter	Mean	Std Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Both	0	0.511	0.065	0.376	0.645
	1	0.402	0.081	0.233	0.570
	2	0.226	0.076	0.067	0.385
	3	0.333	0.078	0.171	0.496
	4	0.224	0.080	0.057	0.392
	5	0.205	0.079	0.041	0.368
1	0	0.512	0.091	0.323	0.701
	1	0.283	0.098	0.080	0.487
	2	0.182	0.104	-0.034	0.398
	3	0.451	0.097	0.249	0.653
	4	0.225	0.095	0.028	0.422
	5	0.164	0.087	-0.016	0.344
2	0	0.509	0.088	0.326	0.693
	1	0.520	0.087	0.339	0.701
	2	0.270	0.096	0.070	0.469
	3	0.216	0.104	-0.001	0.433
	4	0.224	0.114	-0.013	0.461
	5	0.245	0.108	0.020	0.470

Table D.5: Mean values and 95% confidence intervals for the overall child ratings of stimuli in the Identification Task. The table shows the values for the overall different Utterance Parameter configurations, as well as those for the two genders.

Subjects	Utter	Mean	Std Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Both	0	0.511	0.065	0.376	0.645
	1	0.402	0.081	0.233	0.570
	2	0.226	0.076	0.067	0.385
	3	0.333	0.078	0.171	0.496
	4	0.224	0.080	0.057	0.392
	5	0.205	0.079	0.041	0.368
Females	0	0.442	0.085	0.265	0.619
	1	0.325	0.107	0.102	0.547
	2	0.222	0.101	0.013	0.432
	3	0.395	0.103	0.181	0.610
	4	0.369	0.106	0.149	0.590
	5	0.161	0.104	-0.054	0.377
Males	1	0.579	0.097	0.377	0.781
	2	0.478	0.122	0.225	0.732
	3	0.229	0.115	-0.010	0.468
	4	0.272	0.117	0.027	0.516
	5	0.080	0.121	-0.172	0.331
	6	0.248	0.118	0.002	0.494

Appendix E

Using Artificial Neural Networks to Automate NLU Production and Affective Charging

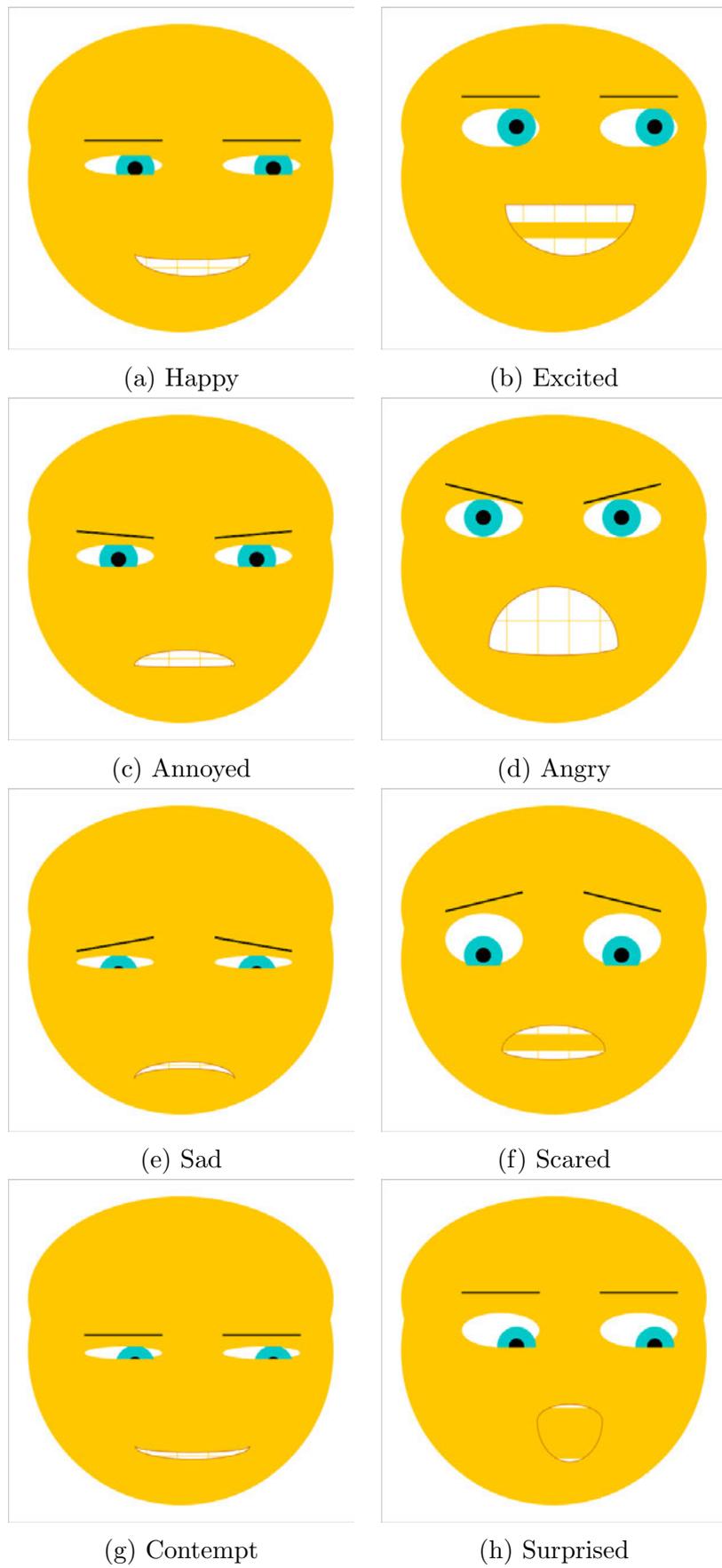


Figure E.1: Images of the AffectButton facial gestures used during the Matching Task in the ANN Evaluation experiment presented in chapter 7.

Table E.1: Mean and Standard Deviations for the Affective PAD ratings for each NLU parameter specifications, across the two pitch contour specifications.

NLU	Contour	Affective Dimension					
		Pleasure		Arousal		Dominance	
		Mean	Std	Mean	Std	Mean	Std
1	1	0.350	0.728	0.613	0.586	0.100	0.779
	2	0.335	0.660	0.417	0.717	0.225	0.687
2	1	-0.007	0.683	0.241	0.783	-0.281	0.733
	2	0.083	0.696	0.393	0.688	0.227	0.738
3	1	0.067	0.733	0.367	0.782	0.205	0.669
	2	-0.007	0.728	0.200	0.784	0.102	0.698
4	1	0.040	0.689	0.074	0.837	-0.250	0.628
	2	0.121	0.741	0.321	0.781	-0.085	0.725
5	1	0.129	0.675	0.167	0.763	0.219	0.656
	2	0.127	0.770	0.397	0.787	0.034	0.731
6	1	0.049	0.726	0.261	0.805	0.137	0.729
	2	0.169	0.773	0.509	0.652	0.372	0.695
7	1	0.240	0.678	0.415	0.765	0.259	0.689
	2	-0.041	0.72	0.133	0.831	-0.025	0.716
8	1	0.154	0.688	0.481	0.589	0.132	0.776
	2	-0.135	0.704	0.369	0.798	0.075	0.759
9	1	0.112	0.712	0.231	0.739	0.110	0.682
	2	0.129	0.692	0.279	0.764	0.064	0.742
10	1	0.263	0.716	0.377	0.849	0.227	0.743
	2	-0.049	0.748	0.282	0.791	0.139	0.702
11	1	0.024	0.748	0.293	0.764	0.175	0.706
	2	-0.038	0.704	0.182	0.812	-0.004	0.721
12	1	0.112	0.657	0.242	0.725	0.242	0.694
	2	0.179	0.697	0.239	0.834	0.046	0.701
13	1	0.265	0.729	0.473	0.685	0.076	0.768
	2	-0.204	0.722	0.324	0.759	0.252	0.673

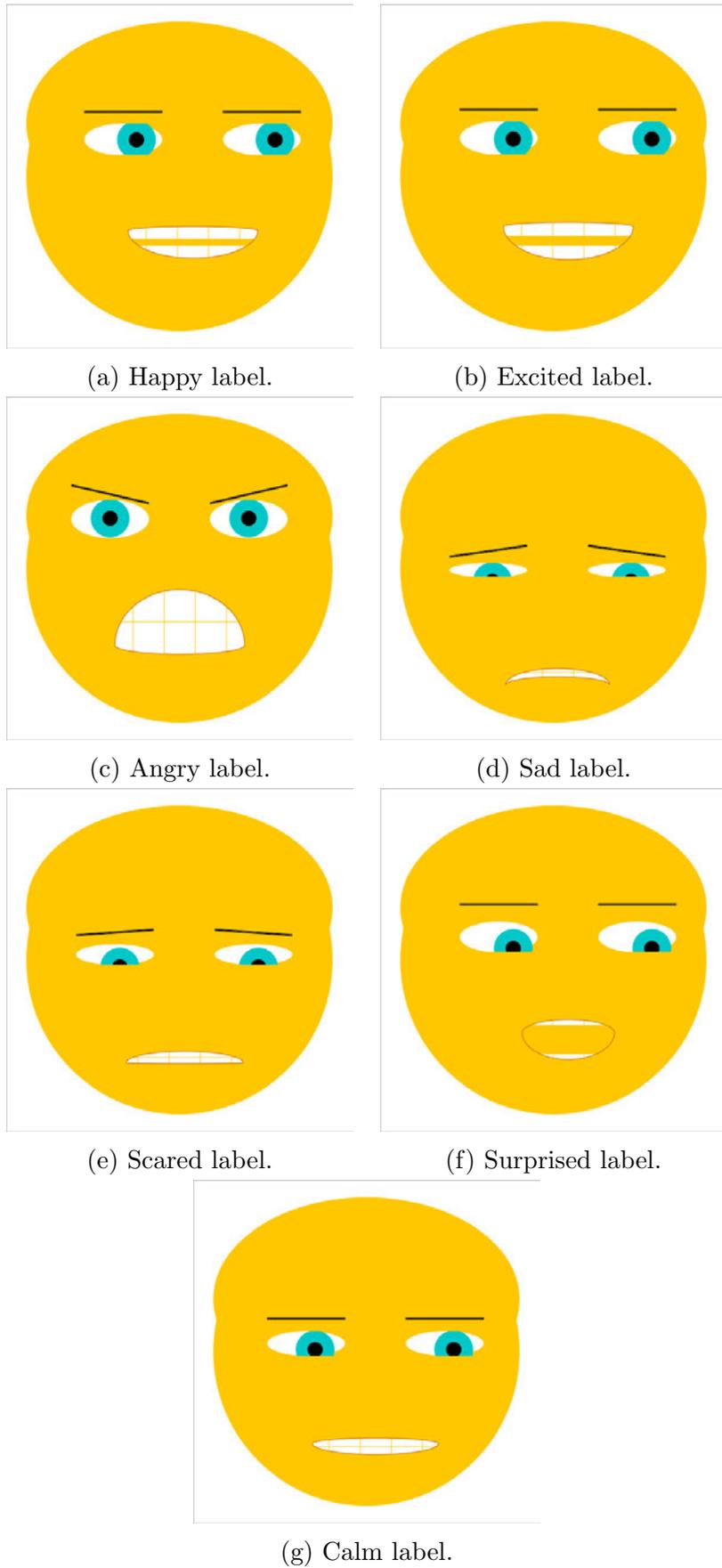


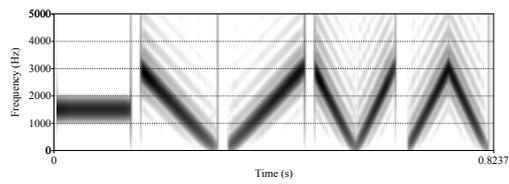
Figure E.2: Images of the AffectButton facial gestures for the mean ratings obtained during the Labelling Task in the ANN Evaluation experiment presented in chapter 7.

Appendix F

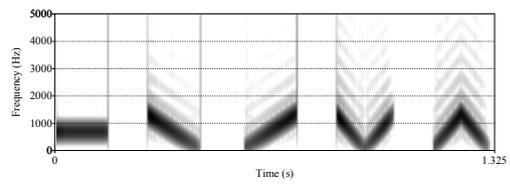
The Influence of Situational Context upon the Interpretation of NLUs

Table F.1: Parameter Configurations of the NLUs used in the Pilot and Main Survey presented in chapter 8. Spectrograms of these are shown in figure F.1 and outline the Pitch Contour

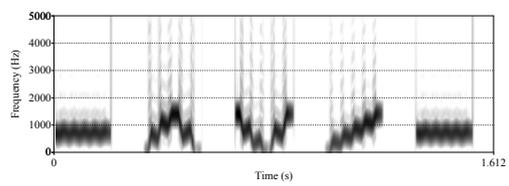
NLU	NLU Parameter						
	Base Freq. (Hz)	Freq. Range (Hz)	Speech Rate	Pause Ratio	Rhythm	S. Unit Count	Tremolo (rad)
1	1500	1500	6	0.05	1	5	0
2	666.67	666.67	3.5	0.633	1	5	0
3	700	700	3	0.5	1	5	0.3839
4	500	550	4	1	1	5	0.2181
5	500	600	5	0.75	1	5	0
6	1200	1000	3	0.5	1	5	0
7	600	500	3	0.75	1	5	0
8	1500	1500	4	0.25	1	5	0.2181



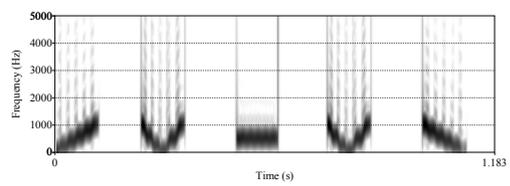
(a) NLU 1



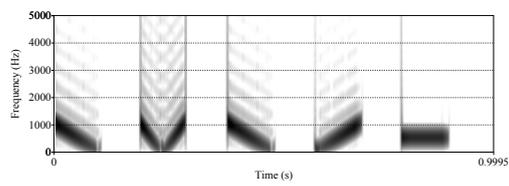
(b) NLU 2



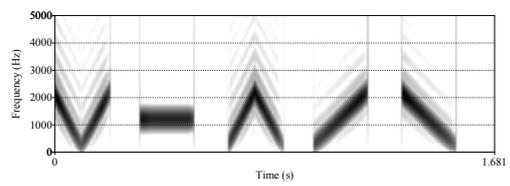
(c) NLU 3



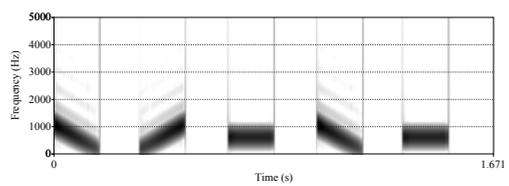
(d) NLU 4



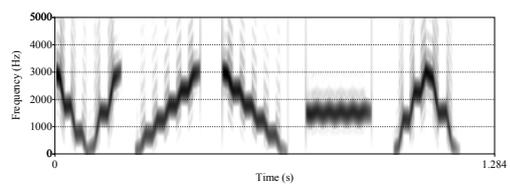
(e) NLU 5



(f) NLU 6



(g) NLU 7



(h) NLU 8

Figure F.1: Spectrograms of the NLUs outlined in table F.1

Table F.2: Question orders for the online CrowdFlower surveys. Questions with a (0) are action videos with no NLU (C_{Action}). Questions with a (1) are action videos combined with NLU #8 (C_{Action}^P). Questions with a (2) are action videos combined with NLU #7 (C_{Action}^N).

Question	Question Order		
	1	2	3
1	Subject Age	Subject Age	Subject Age
2	Subject Gender	Subject Gender	Subject Gender
3	NLU#1	Eye Cover (2)	Head Stroke (2)
4	Head Stroke (1)	Kiss (0)	Eye Cover (0)
5	Slap (2)	Face Flick (1)	Slap (2)
6	Eye Cover (0)	NLU#3	NLU#7
7	Face Flick (1)	Kiss (1)	NLU#2
8	Kiss (0)	NLU#2	Slap (0)
9	NLU#3	Face Flick (2)	Face Flick (2)
10	Eye Cover (2)	Slap (0)	Head Kiss (1)
11	NLU#7	NLU#7	NLU#1
12	Face Flick (2)	Face Flick (0)	NLU#8
13	NLU#8	Head Stroke (2)	Head Kiss (0)
14	Slap (0)	Slap (1)	Face Flick (1)
15	Kiss (1)	Kiss (2)	Eye Cover (2)
16	Slap (1)	Head Stroke (0)	Head Stroke (0)
17	Head Stroke (2)	NLU#1	Head Kiss (2)
18	Face Flick (0)	Eye Cover (1)	Eye Cover (1)
19	NLU#2	Slap (2)	Head Stroke (2)
20	Head Stroke (0)	NLU#8	NLU#3
21	Eye Cover (1)	Eye Cover (0)	Face Flick (0)
22	Kiss (2)	Head Stroke (1)	Slap (1)
23	Seen Robot Before	Seen Robot Before	Seen Robot Before

Table F.3: Mean and Standard Deviations of the Ratings for the Action Videos in the Pilot Study (refer to section 8.1.1.3 and figure 8.3).

Action	Rating	
	Mean	Std Dev
Slap on Head	2.066	1.279
Poke in Chest	3.533	1.187
Poke on Forehead	3.000	1.253
Flick to Head	3.333	1.345
Clicking Fingers	3.800	1.320
Covering Eyes	4.533	0.915
Waving in front of Eyes	5.200	1.740
Chin Tickle	6.333	1.046
Rib Tickle	4.866	1.355
Stroked on Head	6.133	0.990
Kiss on Head	6.533	1.684

Table F.4: Mean and Standard Deviations of the Ratings for the NLU Videos in the Pilot Study (refer to section 8.1.1.3 and figure 8.4).

NLU Video	Rating	
	Mean	Std Dev
1	5.733	1.387
2	4.933	1.751
3	4.000	1.464
4	4.066	1.162
5	6.133	1.060
6	5.000	1.732
7	3.733	1.387
8	6.466	1.807

Table F.5: Mean, Standard Error and 95% Confidence Intervals of the NLU Video Ratings obtained from the CrowdFlower Study (refer to section 8.2.1 and figure 8.6).

NLU Video	Mean	Std Error	95% Confidence Interval	
			Lower Bound	Upper Bound
1	6.100	0.090	5.923	6.277
2	5.781	0.082	5.621	5.942
3	5.866	0.083	5.703	6.029
7	6.102	0.089	5.926	6.278
8	5.219	0.084	5.053	5.385

Table F.6: Mean, Standard Error and 95% Confidence Intervals of the Action Video Ratings obtained from the CrowdFlower Study (refer to section 8.2.2 and figure 8.7).

Action Video	Mean	Std Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Covering Eyes	4.669	0.078	4.517	4.822
Kiss	6.687	0.111	6.469	6.905
Slap	3.045	0.110	2.828	3.262
Stroke	6.361	0.096	6.172	6.549
Flick	3.787	0.106	3.578	3.995

Table F.7: Mean, Standard Error and 95% Confidence Intervals for the overall ratings for the Video Conditions, across each action scenario (refer to section 8.2.3 and figure 8.8).

Action	Condition	Mean	Std Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Eye Cover	C_{NLU}^P	6.100	0.090	5.923	6.277
	C_{NLU}^N	5.219	0.084	5.053	5.385
	C_{Action}	4.669	0.078	4.517	4.822
	C_{Action}^P	4.611	0.096	4.422	4.799
	C_{Action}^N	4.497	0.089	4.322	4.672
Kiss	C_{NLU}^P	6.100	0.090	5.923	6.277
	C_{NLU}^N	5.219	0.084	5.053	5.385
	C_{Action}	6.687	0.111	6.469	6.905
	C_{Action}^P	7.550	0.087	7.378	7.721
	C_{Action}^N	7.017	0.099	6.822	7.212
Slap	C_{NLU}^P	6.100	0.090	5.923	6.277
	C_{NLU}^N	5.219	0.084	5.053	5.385
	C_{Action}	3.045	0.110	2.828	3.262
	C_{Action}^P	3.042	0.118	2.810	3.275
	C_{Action}^N	2.805	0.102	2.604	3.007
Stroke	C_{NLU}^P	6.100	0.090	5.923	6.277
	C_{NLU}^N	5.219	0.084	5.053	5.385
	C_{Action}	6.361	0.096	6.172	6.549
	C_{Action}^P	6.873	0.103	6.670	7.077
	C_{Action}^N	6.663	0.101	6.464	6.862
Flick	C_{NLU}^P	6.100	0.090	5.923	6.277
	C_{NLU}^N	5.219	0.084	5.053	5.385
	C_{Action}	3.787	0.106	3.578	3.995
	C_{Action}^P	3.784	0.116	3.556	4.012
	C_{Action}^N	3.456	0.104	3.251	3.661

Table F.8: Mean, Standard Error and 95% Confidence Intervals for the ratings of the interaction effect between Subject Gender and the Video Condition for the Flicking Action (refer to section 8.2.3.3 and figure 8.9a).

Gender	Condition	Mean	Std Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Females	C_{NLU}^P	6.119	0.095	5.932	6.306
	C_{NLU}^N	5.336	0.089	5.161	5.512
	C_{Action}	3.580	0.112	3.360	3.801
	C_{Action}^P	3.601	0.122	3.360	3.842
	C_{Action}^N	3.277	0.110	3.061	3.494
Males	C_{NLU}^P	6.081	0.153	5.780	6.382
	C_{NLU}^N	5.102	0.143	4.820	5.383
	C_{Action}	3.993	0.180	3.639	4.347
	C_{Action}^P	3.968	0.197	3.580	4.355
	C_{Action}^N	3.634	0.177	3.286	3.982

Table F.9: Mean, Standard Error and 95% Confidence Intervals for the ratings of the interaction effect between Subject Gender and the Video Condition for the Stroking Action (refer to section 8.2.3.4 and figure 8.9b).

Gender	Condition	Mean	Std Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Females	C_{NLU}^P	6.119	0.095	5.932	6.306
	C_{NLU}^N	5.336	0.089	5.161	5.512
	C_{Action}	6.553	0.101	6.354	6.752
	C_{Action}^P	7.159	0.109	6.944	7.374
	C_{Action}^N	7.129	0.107	6.920	7.339
Males	C_{NLU}^P	6.081	0.153	5.780	6.382
	C_{NLU}^N	5.102	0.143	4.820	5.383
	C_{Action}	6.169	0.163	5.848	6.489
	C_{Action}^P	6.588	0.176	6.242	6.933
	C_{Action}^N	6.196	0.171	5.859	6.534

Table F.10: Mean, Standard Error and 95% Confidence Intervals for the ratings of the interaction effect between Subject Gender and the Video Condition for the Eye Covering Action (refer to section 8.2.3.5 and figure 8.9c).

Gender	Condition	Mean	Std Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Females	C_{NLU}^P	6.119	0.095	5.932	6.306
	C_{NLU}^N	5.336	0.089	5.161	5.512
	C_{Action}	4.594	0.082	4.433	4.755
	C_{Action}^P	4.403	0.101	4.204	4.602
	C_{Action}^N	4.456	0.094	4.271	4.641
Males	C_{NLU}^P	6.081	0.153	5.780	6.382
	C_{NLU}^N	5.102	0.143	4.820	5.383
	C_{Action}	4.744	0.132	4.485	5.003
	C_{Action}^P	4.819	0.163	4.498	5.139
	C_{Action}^N	4.538	0.151	4.241	4.836

Table F.11: Mean, Standard Error and 95% Confidence Intervals for the ratings of the interaction effect between Robot Familiarity and the Video Condition for the Eye Covering Action (refer to section 8.2.3.5 and figure 8.9d).

Familiarity	Condition	Mean	Std Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Unfamiliar	C_{NLU}^P	6.032	0.135	5.765	6.298
	C_{NLU}^N	5.017	0.127	4.768	5.267
	C_{Action}	4.810	0.117	4.580	5.039
	C_{Action}^P	4.640	0.144	4.357	4.924
	C_{Action}^N	4.506	0.134	4.242	4.770
Familiar	C_{NLU}^P	6.168	0.119	5.935	6.402
	C_{NLU}^N	5.421	0.111	5.202	5.639
	C_{Action}	4.529	0.102	4.328	4.730
	C_{Action}^P	4.581	0.126	4.333	4.830
	C_{Action}^N	4.488	0.117	4.257	4.719

Appendix G

Combining NLUs with Natural Language

Table G.1: Results of the 2-way ANOVAs for Appropriateness ratings with the subjects split across the robot familiarity factor (refer to section 9.2.1.1). The Tables shows the Mean Values, Standard Errors and 95% Confidence Intervals. See figure 9.4.

Familiarity	Gender	Subjects (N)	Video Condition	Mean	Standard Error	95% Confidence Interval	
						Lower Bound	Upper Bound
Unfamiliar	All	104	1	7.358	0.144	7.073	7.642
			2	6.052	0.239	5.578	6.526
			3	6.757	0.178	6.405	7.109
			4	6.938	0.170	6.600	7.275
	Females	72	1	7.403	0.159	7.087	7.719
			2	5.792	0.265	5.266	6.318
			3	7.014	0.197	6.623	7.405
			4	7.000	0.189	6.626	7.374
	Males	32	1	7.313	0.239	6.839	7.786
			2	6.313	0.398	5.523	7.102
			3	6.500	0.295	5.914	7.086
			4	6.875	0.283	6.314	7.436
Familiar	All	166	1	7.515	0.103	7.311	7.719
			2	6.160	0.160	5.844	6.476
			3	6.911	0.105	6.703	7.119
			4	7.047	0.104	6.841	7.253
	Females	109	1	7.294	0.121	7.055	7.532
			2	6.284	0.187	5.914	6.655
			3	6.927	0.124	6.683	7.171
			4	6.954	0.122	6.713	7.196
	Males	57	1	7.737	0.167	7.406	8.067
			2	6.035	0.259	5.523	6.547
			3	6.895	0.171	6.557	7.232
			4	7.140	0.169	6.806	7.474

Table G.2: Results of the 2-way ANOVAs for Appropriateness ratings with the subjects split across the subject gender factor (refer to section 9.2.1.2). The Tables shows the Mean Values, Standard Errors and 95% Confidence Intervals. See figure 9.5.

Gender	Familiarity	Subjects (<i>N</i>)	Video Condition	Mean	Standard Error	95% Confidence Interval	
						Lower Bound	Upper Bound
Females	All	181	1	7.348	0.102	7.147	7.549
			2	6.038	0.162	5.718	6.358
			3	6.970	0.106	6.726	7.179
			4	6.977	0.110	6.760	7.194
	Unfamiliar	72	1	7.403	0.158	7.091	7.715
			2	5.792	0.251	5.296	6.288
			3	7.014	0.164	6.690	7.337
			4	7.000	0.171	6.664	7.336
	Familiar	109	1	7.294	0.128	7.040	7.547
			2	6.284	0.204	5.881	6.687
			3	6.927	0.133	6.664	7.190
			4	6.954	0.139	6.681	7.228
Males	All	89	1	7.525	0.133	7.261	7.789
			2	6.174	0.215	5.746	6.602
			3	6.697	0.172	6.355	7.040
			4	7.008	0.147	6.715	7.300
	Unfamiliar	32	1	7.313	0.213	6.890	7.735
			2	6.313	0.345	5.627	6.995
			3	6.500	0.276	5.952	7.048
			4	6.875	0.235	6.407	7.343
	Familiar	57	1	7.737	0.159	7.420	8.054
			2	6.035	0.258	5.522	6.548
			3	6.895	0.206	6.484	7.305
			4	7.140	0.176	6.790	7.491

Table G.3: Results of the 2-way ANOVAs for Expressiveness ratings with the subjects split across the robot familiarity factor (refer to section 9.2.2.1). The Tables shows the Mean Values, Standard Errors and 95% Confidence Intervals. See figure 9.6.

Familiarity	Gender	Subjects (<i>N</i>)	Video Condition	Mean	Standard Error	95% Confidence Interval	
						Lower Bound	Upper Bound
Unfamiliar	All	104	1	7.047	0.156	6.738	7.355
			2	5.628	0.262	5.108	6.149
			3	6.688	0.176	6.338	7.037
			4	6.960	0.160	6.644	7.277
	Females	72	1	7.000	0.173	6.658	7.342
			2	5.069	0.291	4.492	5.646
			3	6.875	0.195	6.488	7.262
			4	6.889	0.177	6.538	7.240
	Males	32	1	7.094	0.259	6.580	7.607
			2	6.188	0.436	5.322	7.053
			3	6.500	0.293	5.919	7.081
			4	7.031	0.265	6.505	7.558
Familiar	All	166	1	7.300	0.114	7.076	7.525
			2	5.913	0.164	5.588	6.237
			3	6.892	0.110	6.673	7.110
			4	6.990	0.114	6.765	7.215
	Females	109	1	7.092	0.133	6.829	7.355
			2	6.018	0.193	5.638	6.399
			3	6.853	0.129	6.598	7.109
			4	6.752	0.133	6.489	7.016
	Males	57	1	7.509	0.184	7.145	7.873
			2	5.807	0.266	5.281	6.333
			3	6.930	0.179	6.576	7.283
			4	7.228	0.185	6.864	7.592

Table G.4: Results of the 2-way ANOVAs for Expressiveness ratings with the subjects split across the subject gender factor (refer to section 9.2.2.2). The Tables shows the Mean Values, Standard Errors and 95% Confidence Intervals. See figure 9.7.

Gender	Familiarity	Subjects (<i>N</i>)	Video Condition	Mean	Standard Error	95% Confidence Interval	
						Lower Bound	Upper Bound
Females	All	181	1	7.046	0.110	6.828	7.263
			2	5.544	0.172	5.204	5.883
			3	6.864	0.109	6.649	7.079
			4	6.821	0.113	6.597	7.044
	Unfamiliar	72	1	7.000	0.171	6.662	7.338
			2	5.069	0.267	4.543	5.596
			3	6.875	0.169	6.541	7.209
			4	6.889	0.176	6.541	7.236
	Familiar	109	1	7.092	0.139	6.817	7.366
			2	6.018	0.217	5.590	6.447
			3	6.853	0.138	6.582	7.125
			4	6.752	0.143	6.470	7.035
Males	All	89	1	7.301	0.149	7.005	7.598
			2	5.997	0.226	5.547	6.447
			3	6.715	0.171	6.374	7.056
			4	7.130	0.145	6.842	7.417
	Unfamiliar	32	1	7.094	0.239	6.619	7.569
			2	6.188	0.362	5.467	6.908
			3	6.500	0.274	5.954	7.046
			4	7.031	0.231	6.571	7.491
	Familiar	57	1	7.509	0.179	7.153	7.865
			2	5.807	0.272	5.267	6.347
			3	6.930	0.206	6.521	7.339
			4	7.228	0.173	6.884	7.573

Table G.5: Results of the Univariate tests for the Preference ratings (refer to section 9.2.3). The table shows the Mean values, Standard Errors and 95% Confidence Intervals for the Main effects due to the video condition and robot familiarity factors. See figure 9.8.

Familiarity	Subjects (N)	Video Condition	Mean	Standard Error	95% Confidence Interval	
					Lower Bound	Upper Bound
All	270	1	7.436	0.087	7.265	7.608
		2	6.106	0.139	5.832	6.380
		3	6.834	0.097	6.643	7.025
		4	6.992	0.095	6.806	7.178
Unfamiliar	104	1	6.521	0.168	6.189	6.852
		2	4.785	0.266	4.261	5.308
		3	5.967	0.174	5.625	6.309
		4	6.325	0.173	5.985	6.664
Familiar	166	1	7.188	0.130	6.933	7.443
		2	5.194	0.204	4.791	5.597
		3	6.552	0.134	6.289	6.815
		4	6.615	0.133	6.354	6.876

Table G.6: Results of the Univariate tests for the Naturalness Ratings (refer to section 9.2.4). The table shows the Mean values, Standard Errors and 95% Confidence Intervals for the Interaction effect found between the video condition and subject gender factors. See figure 9.9.

Gender	Subjects (N)	Video Condition	Mean	Standard Error	95% Confidence Interval	
					Lower Bound	Upper Bound
All	270	1	7.165	0.116	6.936	7.393
		2	4.981	0.178	4.630	5.331
		3	6.586	0.109	6.372	6.800
		4	6.781	0.104	6.577	6.985
Females	181	1	7.119	0.131	6.860	7.378
		2	4.440	0.202	4.042	4.837
		3	6.549	0.123	6.306	6.791
		4	6.523	0.118	6.292	6.755
Males	89	1	7.211	0.191	6.834	7.587
		2	5.522	0.294	4.944	6.100
		3	6.623	0.179	6.271	6.976
		4	7.039	0.171	6.702	7.376

Table G.7: Results of the Univariate tests for the Like-ability Ratings (refer to section 9.2.5 and figure 9.10a) The table shows the Mean values, Standard Errors and 95% Confidence Intervals for the Interaction effect between the video condition and subject gender.

Gender	Subjects (<i>N</i>)	Video Condition	Mean	Standard Error	95% Confidence Interval	
					Lower Bound	Upper Bound
All	270	1	7.305	0.109	7.091	7.519
		2	6.462	0.145	6.176	6.748
		3	7.001	0.111	6.783	7.220
		4	7.095	0.110	6.878	7.311
Females	181	1	7.184	0.123	6.942	7.426
		2	6.077	0.165	5.753	6.640
		3	6.972	0.126	6.725	7.220
		4	6.887	0.125	6.641	7.133
Males	89	1	7.426	0.179	7.073	7.779
		2	6.846	0.239	6.375	7.318
		3	7.031	0.183	6.670	7.391
		4	7.302	0.181	6.945	7.659

Table G.8: Results of the Univariate tests for the Like-ability, Subject Gender and Robot Familiarity Interaction (refer to section 9.2.5 and figure 9.10b). The table shows the Mean values, Standard Errors and 95% Confidence Intervals for the Interaction effect found between the subject gender and robot familiarity.

Gender	Familiarity	Subjects (<i>N</i>)	Mean	Standard Error	95% Confidence Interval	
					Lower Bound	Upper Bound
Females	Unfamiliar	72	6.698	0.172	6.359	7.037
	Familiar	109	6.862	0.140	6.587	7.138
Males	Unfamiliar	32	6.680	0.258	6.171	7.189
	Familiar	57	7.623	0.194	7.242	8.004

Bibliography

- Bainbridge, W. a., Hart, J. W., Kim, E. S., and Scassellati, B. (2010). The Benefits of Interactions with Physically Present Robots over Video-Displayed Agents. *International Journal of Social Robotics*, 3(1):41–52.
- Banse, R. and Scherer, K. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, 70(3):614–636.
- Banziger, T. and Scherer, K. (2005). The role of intonation in emotional expressions. *Speech Communication*, 46(3-4):252–267.
- Bartneck, C., Kanda, T., Ishiguro, H., and Hagita, N. (2009). My Robotic Doppelgänger - A critical look at the Uncanny Valley. In *Proceedings of The 18th International Symposium on Robot and Human Interactive Communication (ROMAN 2009)*, pages 269–276, Toyama, Japan. IEEE.
- Belpaeme, T., Baxter, P., Greeff, J., Kennedy, J., Read, R., Looije, R., Neerincx, M., Baroni, I., and Zelati, M. C. (2013). Child-robot interaction: Perspectives and challenges. In Herrmann, G., Pearson, Martin, J., Lenz, A., Bremner, P., Spiers, A., and Leonards, U., editors, *Social Robotics*, volume 8239 of *Lecture Notes in Computer Science*, pages 452–459. Springer International Publishing.
- Belpaeme, T., Baxter, P., Read, R., Wood, R., Cuayáhuitl, H., Kiefer, B., Racioppa, S., Athanasopoulos, G., Enescu, V., Looije, R., Neerincx, M., Demiris, Y., Ros-Espinoza, R., Beck, A., Cañamero, L., Hiolle, A., Lewis, M., Baroni, I., Nalin, M., Cosi, P., Paci, G., Tesser, F., Somlavilla, G., and Humbert, R. (2012). Multimodal Child-Robot Interaction: Building Social Bonds. *Journal of Human-Robot Interaction*, 1(2):33–53.
- Bimler, D. and Kirkland, J. (2001). Categorical perception of facial expressions of emotion: Evidence from multidimensional scaling. *Cognition and Emotion*, 15(5):633–658.
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.
- Bornstein, M. H., Kessen, W., and Weiskopf, S. (1976). Color Vision and Hue Categorization in Young Human Infants. *Journal of experimental psychology. Human perception and performance*, 2(1):115–129.
- Breazeal, C. (2001a). Affective interaction between humans and robots. In *Proceedings of the 6th European Conference on Advances in Artificial Life (ECAL '01)*, pages 582 – 591, Prague, Czech Republic. Springer-Verlag.

- Breazeal, C. (2001b). Emotive Qualities in Robot Speech. In *Proceedings of the International Conference on Intelligent Robots and Systems (IROS 2001)*, pages 1388 – 1394, Maui, HI, USA. IEEE/RSJ.
- Breazeal, C. (2002). *Designing Sociable Robots*. The MIT Press, Cambridge, M.A., U.S.A.
- Breazeal, C. (2003a). Emotion and sociable humanoid robots. *International Journal of Human-Computer Studies*, 59(1-2):119–155.
- Breazeal, C. (2003b). Toward sociable robots. *Robotics and Autonomous Systems*, 42(3-4):167–175.
- Breazeal, C. (2004a). Function meets style: insights from emotion theory applied to HRI. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 34(2):187 – 194.
- Breazeal, C. (2004b). Social interactions in HRI: The robot view. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 34(2):181–186.
- Breazeal, C. and Brooks, R. (2005). Robot Emotion: A Functional Perspective. In Fellous, J.-M. and Arbib, M. A., editors, *Who Needs Emotions? The Brain Meets the Robot*, pages 271 – 310. Oxford University Press, Oxford, UK.
- Breazeal, C., Depalma, N., Orkin, J., and Chernova, S. (2013). Crowdsourcing Human-Robot Interaction : New Methods and System Evaluation in a Public Environment. *Journal of Human-Robot Interaction*, 2(1):82–111.
- Broekens, J. and Brinkman, W.-P. (2009). AffectButton : Towards a Standard for Dynamic Affective User Feedback Dominance-Based Feedback. In *Affective Computing and Intelligent Interaction (ACII '09)*.
- Broekens, J. and Brinkman, W.-P. (2013). Affectbutton: a method for reliable and valid affective self-report. *International Journal of Human-Computer Studies*, 71(6):641 – 667.
- Broekens, J., Pronker, A., and Neuteboom, M. (2010). Real Time Labelling of Affect in Music Using the AffectButton. In *Proceedings of the 3rd International Workshop on Affective Interaction in Natural Environments (AFFINE 2010) at ACM Multimedia 2010*, pages 21–26, Firenze, Italy. ACM.
- Bryant, G. and Barrett, H. (2007). Recognizing intentions in infant-directed speech: Evidence for universals. *Psychological Science*, 18(8):746 – 751.
- Bryant, G. and Barrett, H. (2008). Vocal emotion recognition across disparate cultures. *Journal of Cognition and Culture*, 8, 1(2):135–148.
- Burkhardt, F. and Sendlmeier, W. (2000). Verification of acoustical correlates of emotional speech using formant-synthesis. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, pages 151–156. Citeseer.
- Cahn, J. (1990). The generation of affect in synthesized speech. *Journal of the American Voice I/O Society*, 8:1–19.

- Castellano, G., Leite, I., Pereira, A., Martinho, C., Paiva, A., and Mcowan, P. W. (2013). Multimodal Affect Modelling and Recognition for Empathic Robot Companions. *International Journal of Humanoid Robotics*, 10(1):1350010.
- Chao, C. and Thomaz, A. (2013). Controlling Social Dynamics with a Parametrized Model of Floor Regulation. *Journal of Human-Robot Interaction*, 2(1):4 – 29.
- Cheal, J. L. and Rutherford, M. D. (2011). Categorical perception of emotional facial expressions in preschoolers. *Journal of Experimental Child Psychology*, 110(3):434–43.
- Chernova, S., DePalma, N., Morant, E., and Breazeal, C. (2011). Crowdsourcing Human-Robot Interaction: Application from Virtual to Physical Worlds. In *Proceedings of the 20th International Symposium on Robot and Human Interactive Communication*, pages 21–26, Atlanta, U.S.A. IEEE.
- Coradeschi, S., Ishiguro, H., Asada, M., Shapiro, S., Thielscher, M., Breazeal, C., Mataric, M., and Ishida, H. (2006). Human-Inspired Robots. *Intelligent Systems*, 21(4):74–85.
- Cowie, R. and Cornelius, R. (2003). Describing the emotional states that are expressed in speech. *Speech Communication*, 40(1-2):5–32.
- Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., and Schröder, M. (2000). 'FEELTRACE': An instrument for recording perceived emotion in real time. In *Proceedings of the ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, pages 19–24, Newcastle, Northern Ireland, United Kingdom.
- Damasio, A. R. (1994). *Descartes' Error: Emotion, Reason, and the Human Brain*. Harper Perennial.
- Darwin, C. (1872/2009). *The Expression of the Emotions in Man and Animals*. Harper Perennial, 200th anniversary edition.
- Delaunay, F., de Greeff, J., and Belpaeme, T. (2009). Towards retro-projected robot faces: An alternative to mechatronic and android faces. In *Proceedings of the 18th International Symposium on Robot and Human Interactive Communication (ROMAN 2009)*, pages 306–311, Toyama, Japan. IEEE.
- Delaunay, F., de Greeff, J., and Belpaeme, T. (2010). A study of a retro-projected robotic face and its effectiveness for gaze reading by humans. In *Proceedings of the 5th International Conference on Human-Robot Interaction (HRI'10)*, pages 39 – 44, Osaka, Japan. ACM/IEEE.
- Duffy, B. R. (2003). Anthropomorphism and the social robot. *Robotics and Autonomous Systems*, 42(3-4):177–190.
- Duffy, B. R. and Zawieska, K. (2012). Suspension of Disbelief in Social Robotics. In *Proceedings of The 21st International Symposium on Robot and Human Interactive Communication (RO-MAN 2012)*, pages 484–489, Paris, France. IEEE.
- Ekman, P. (1992). An Argument for Basic Emotions. *Cognition & Emotion*, 6(3/4):169–200.

- Feil-seifer, D. and Skinner, K. (2007). Benchmarks for evaluating socially assistive robotics. *Interaction Studies*, 8(3):423 – 439.
- Fernald, A. (1989). Intonation and communicative intent in mothers' speech to infants: is the melody the message? *Child Development*, 60(6):1497–1510.
- Fontaine, J. R. J., Scherer, K., Roesch, E. B., and Ellsworth, P. C. (2007). The world of emotions is not two-dimensional. *Psychological Science*, 18(12):1050–7.
- Franklin, A. and Davies, I. R. (2004). New evidence for infant colour categories. *British Journal of Developmental Psychology*, 22(3):349–377.
- Gerrits, E. and Schouten, M. (2004). Categorical Perception Depends on the Discrimination Task. *Perception & Psychophysics*, 66(3):363–376.
- Grandjean, D., Bänziger, T., and Scherer, K. R. (2006). Intonation as an interface between language and affect. *Progress in Brain Research*, 156:235 – 247.
- Hackett, C. (1960). The Origin of Speech. *Scientific American*, 203:88 – 96.
- Hagan, M. and Menhaj, M. (1994). Training feedforward networks with the Marquardt algorithm. *IEEE Transactions on Neural Networks*, 5(6):2–6.
- Harnad, S., editor (1987). *Categorical Perception: The Groundwork of Cognition*. Cambridge University Press, Cambridge UK.
- Hayes, A. and Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1(1):77–89.
- Holekamp, K. E. (2007). Questioning the social intelligence hypothesis. *Trends in Cognitive Sciences*, 11(2):65–69.
- Hwang, J., Park, T., and Hwang, W. (2013). The effects of overall robot shape on the emotions invoked in users and the perceived personalities of robot. *Applied ergonomics*, 44(3):459–71.
- Iida, A., Campbell, N., Higuchi, F., and Yasumura, M. (2003). A corpus-based speech synthesis system with emotion. *Speech Communication*, 40(1-2):161–187.
- Isomursu, M., Tähti, M., Väinämö, S., and Kuutti, K. (2007). Experimental evaluation of five methods for collecting emotions in field settings with mobile applications. *International Journal of Human-Computer Studies*, 65(4):404–418.
- Izard, C. E. (2007). Basic Emotions, Natural Kinds, Emotion Schemas, and a New Paradigm. *Perspectives on Psychological Science*, 2(3):260–280.
- James, W. (1890). *The Principles of Psychology, Vol. 1*. Henry Holt, New York, USA.
- Jee, E., Jeong, Y., Kim, C., and Kobayashi, H. (2010). Sound design for emotion and intention expression of socially interactive robots. *Intelligent Service Robotics*, 3:199–206.

- Jee, E.-S., Kim, C. H., Park, S.-Y., and Lee, K.-W. (2007). Composition of Musical Sound Expressing an Emotion of Robot Based on Musical Factors. In *Proceedings of the 16th International Symposium on Robot and Human Interactive Communication (RO-MAN 2007)*, pages 637–641, Jeju Island, Korea. IEEE.
- Jee, E.-S., Park, S.-Y., Kim, C. H., and Kobayashi, H. (2009). Composition of musical sound to express robot’s emotion with intensity and synchronized expression with robot’s behavior. In *Proceedings of the 18th International Symposium on Robot and Human Interactive Communication (RO-MAN 2009)*, pages 369–374, Toyama, Japan. IEEE.
- Juslin, P. N. and Laukka, P. (2003). Communication of emotions in vocal expression and music performance: different channels, same code? *Psychological bulletin*, 129(5):770–814.
- Juslin, P. N. and Scherer, K. R. (2005). Vocal expression of affect. In Harrigan, J., Rosenthal, R., and Scherer, K. R., editors, *The New Handbook of methods in nonverbal behavior research*, chapter 3, pages 65 – 135. Oxford University Press, Oxford, UK.
- Kaliouby, R. E. and Robinson, P. (2004). Real-time inference of complex mental states from facial expressions and head gestures. In *Proceedings of the International on Computer Vision & Pattern Recognition*, Washington, USA.
- Kelley, J. F. (1984). An Iterative Design Methodology for User-Friendly Natural Language Office Information Applications. *ACM Transactions on Information Systems*, 2(1):26–41.
- Knoll, M., Uther, M., and Costall, A. (2009). Effects of Low-Pass Filtering on the Judgment of Vocal Affect in Speech Directed to Infants, Adults and Foreigners. *Speech Communication*, 51(3):210 – 216.
- Komatsu, T. (2005). Toward Making Humans Empathize with Artificial Agents by Means of Subtle Expressions. In *1st International Conference on Affective Computing and Intelligent Interaction (ACII2005)*, pages 458 – 465, Beijing, China.
- Komatsu, T. and Kobayashi, K. (2012). Can Users Live with Overconfident or Unconfident Systems?: A Comparison of Artificial Subtle Expressions with Human-Like Expression. In *Proceedings of Conference on Human Factors in Computing Systems (CHI 2012)*, pages 1595–1600, Austin, Texas.
- Komatsu, T., Kurosawa, R., and Yamada, S. (2011). How Does the Difference Between Users’ Expectations and Perceptions About a Robotic Agent Affect Their Behavior? *International Journal of Social Robotics*, 4(2):109–116.
- Komatsu, T. and Yamada, S. (2007). How appearance of robotic agents affects how people interpret the agents’ attitudes. *Proceedings of the international conference on Advances in computer entertainment technology - ACE ’07*, page 123.

- Komatsu, T. and Yamada, S. (2008). How does appearance of agents affect how people interpret the agents' attitudes: An Experimental investigation on expressing the same information from agents having different appearance. In *IEEE Congress on Evolutionary Computation*, pages 1935–1940.
- Komatsu, T. and Yamada, S. (2011). How Does the Agents' Appearance Affect Users' Interpretation of the Agents' Attitudes: Experimental Investigation on Expressing the Same Artificial Sounds From Agents With Different Appearances. *International Journal of Human-Computer Interaction*, 27(3):260–279.
- Komatsu, T., Yamada, S., Kobayashi, K., Funakoshi, K., and Nakano, M. (2010). Artificial Subtle Expressions: Intuitive Notification Methodology of Artifacts. In *Proceedings of the 28th International Conference on Human Factors in Computing Systems (CHI'10)*, pages 1941–1944, New York, New York, USA. ACM.
- Kozima, H., Michalowski, M. P., and Nakagawa, C. (2009). Keepon: A Playful Robot for Research, Therapy, and Entertainment. *International Journal of Social Robotics*, 1(1):3–18.
- Kuhl, P. K. (1991). Human adults and human infants show a “perceptual magnet effect” for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics*, 50(2):93–107.
- Kuratate, T., Matsusaka, Y., Pierce, B., and Cheng, G. (2011). “Mask-bot”: A life-size robot head using talking head animation for human-robot communication. In *Proceedings of the 11th IEEE-RAS International Conference on Humanoid Robots (Humanoids 2011)*, pages 99–104, Bled, Slovenia. IEEE.
- Lang, P. and Bradley, M. (1994). Measuring Emotion: The Self-Assessment Manikin and the Semantic Differential. *Journal of Behaviour Therapy & Experimental Psychiatry*, 25(1):49–59.
- Laukka, P. (2005). Categorical Perception of Vocal Emotion Expressions. *Emotion*, 5(3):277–295.
- Laukka, P., Juslin, P., and Bresin, R. (2005). A Dimensional Approach to Vocal Expression of Emotion. *Cognition & Emotion*, 19(5):633–653.
- Le Sourn-Bissaoui, S., Aguert, M., Girard, P., Chevreuil, C., and Laval, V. (2013). Emotional Speech Comprehension in Children and Adolescents with Autism Spectrum Disorders. *Journal of Communication Disorders*, 46(4):309 – 320.
- Lee, M. K., Kiesler, S., Forlizzi, J., Srinivasa, S., and Rybski, P. (2010). Gracefully Mitigating Breakdowns in Robotic Services. In *Proceedings of the 5th International Conference on Human-Robot Interaction (HRI'10)*, pages 203–210, Osaka, Japan. ACM/IEEE.
- Leite, I., Martinho, C., and Paiva, A. (2013a). Social Robots for Long-Term Interaction: A Survey. *International Journal of Social Robotics*, 5(2):291–308.
- Leite, I., Pereira, A., and Mascarenhas, S. (2013b). The Influence of Empathy in Human-Robot Relations. *International Journal of Human-Computer Studies*, 71(3):250 – 260.

- Leyzberg, D., Spaulding, S., Toneva, M., and Scassellati, B. (2012). The Physical Presence of a Robot Tutor Increases Cognitive Learning Gains. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society (CogSci 2012)*, pages 1882 – 1887, Sapporo, Japan.
- Liberman, A., Harris, K., and Hoffman, H. (1957). The Discrimination of Speech Sounds Within and Across Phoneme Boundaries. *Journal of Experimental Psychology*, 54(5):358–368.
- Lison, P. and Kruiff, G.-J. (2009). Robust processing of situated spoken dialogue. In *Proceedings of the 32nd annual German conference on Advances in Artificial Intelligence (KI'09)*, pages 241–248, Paderbronn, Germany. Springer-Verlag.
- Lohse, M., Hanheide, M., Wrede, B., Walters, M., Koay, K. L., Syrdal, D., Green, A., Huttenrauch, H., Dautenhahn, K., Sagerer, G., and Severinson-Eklundh, K. (2008a). Evaluating Extrovert and Introvert Behaviour of a Domestic Robot Video Study. In *Proceedings of the 17th International Symposium on Robot and Human Interactive Communication (RO-MAN 2008)*, pages 488–493, Munich, Germany. IEEE.
- Lohse, M., Rohlfing, K. J., Wrede, B., and Sagerer, G. (2008b). “Try something else!”, When users change their discursive behaviour in human-robot interaction. In *Proceedings of the International Conference on Robotics and Automation (ICRA 2008)*, pages 3481–3486, Pasadena, California, U.S.A. IEEE.
- MacDorman, K. F. and Ishiguro, H. (2006). The uncanny advantage of using androids in cognitive and social science research. *Interaction Studies*, 7(3):297–337.
- Marsland, S. (2009). *Machine Learning: An Algorithmic Perspective*. Chapman & Hall/CRC, 1st edition.
- Maurer, D., Pathman, T., and Mondloch, C. J. (2006). The shape of boubas: sound-shape correspondences in toddlers and adults. *Developmental Science*, 9(3):316–22.
- McCartney, J. (2002). Rethinking the Computer Music Language: SuperCollider. *Computer Music Journal*, 26(4):61– 68.
- Mccoll, D. and Nejat, G. (2013). Meal-Time with a Socially Assistive Robot and Older Adults at a Long-term Care Facility. *Journal of Human-Robot Interaction*, 2(1):152 – 171.
- Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1 edition.
- Mitchell, W. J., Szerszen, K. a., Lu, A. S., Schermerhorn, P. W., Scheutz, M., and Macdorman, K. F. (2011). A mismatch in the human realism of face and voice produces an uncanny valley. *i-Perception*, 2(1):10–2.
- Moore, R. K. (2012). A Bayesian explanation of the ‘Uncanny Valley’ effect and related psychological phenomena. *Scientific Reports*, 2:864.
- Mori, M. (1970). The Uncanny Valley. *Energy*, 7:33–35.

- Mubin, O., Bartneck, C., and Feijs, L. (2009). What you say is not what you get: Arguing for Artificial Languages Instead of Natural Languages in Human Robot Speech Interaction. In *the Spoken Dialogue and Human-Robot Interaction Workshop at IEEE RoMan 2009*, Toyama, Japan.
- Mubin, O., Bartneck, C., and Feijs, L. (2010a). Towards the Design and Evaluation of ROILA: a speech recognition friendly artificial language. In *Proceedings of the 7th International Conference on Advances in Natural Language Processing (IceTAL'10)*, pages 250–256, Reykjavik, Iceland. Springer-Verlag.
- Mubin, O., Bartneck, C., Leijs, L., Hooft van Huysduynen, H., Hu, J., and Muelver, J. (2012). Improving Speech Recognition with the Robot Interaction Language. *Disruptive Science and Technology*, 1(2):79–88.
- Mubin, O., Shahid, S., van de Sande, E., Krahmer, E., Swerts, M., Bartneck, C., and Feijs, L. (2010b). Using child-robot interaction to investigate the user acceptance of constrained and artificial languages. In *Proceedings of the 19th International Symposium in Robot and Human Interactive Communication (RO-MAN 2010)*, pages 588–593, Viareggio, Italy. IEEE.
- Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. The MIT Press, Cambridge, Massachusetts, USA.
- Murray, I. and Arnott, J. (1993). Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion. *The Journal of the Acoustical Society of America*, 93(2):1097–1108.
- Murray, I. and Arnott, J. (1996). Synthesizing emotions in speech: is it time to get excited? In *Proceeding of Fourth International Conference on Spoken Language Processing (ICSLP 96)*, pages 1816–1819. IEEE.
- Nagel, F., Kopiez, R., Grewe, O., and Altenmüller, E. (2007). EMuJoy: software for continuous measurement of perceived emotions in music. *Behavior Research Methods*, 39(2):283–290.
- Nass, C. and Brave, S. (2005). *Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship*. MIT Press, Cambridge, MA.
- Nuckolls, J. (1999). The Case for Sound Symbolism. *Annual Review of Anthropology*, 28(1999):225 – 252.
- Ortony, A. and Turner, T. (1990). What’s Basic About Basic Emotions? *Psychological Review*, 97(3):315 – 331.
- Oudeyer, P.-Y. (2003). The production and recognition of emotions in speech: features and algorithms. *International Journal of Human-Computer Studies*, 59(1-2):157–183.
- Ozturk, O., Krehm, M., and Vouloumanos, A. (2013). Sound symbolism in infancy: evidence for sound-shape cross-modal correspondences in 4-month-olds. *Journal of Experimental Child Psychology*, 114(2):173–86.
- Paepcke, S. and Takayama, L. (2010). Judging a Bot by Its Cover: An Experiment on Expectation Setting for Personal Robots. In *Proceedings of the 5th International Conference on Human-Robot Interaction (HRI'10)*, pages 45–52, Osaka, Japan. ACM/IEEE.

- Picard, R. W. (1997). *Affective Computing*. MIT Press, Cambridge, MA, U.S.A.
- Plutchik, R. (1994). *The Psychology and Biology of Emotion*. HarperCollins College Publishers, New York, NY, U.S.A.
- Read, R. and Belpaeme, T. (2010). Interpreting Non-Linguistic Utterances by Robots : Studying the Influence of Physical Appearance. In *Proceedings of the 3rd International Workshop on Affective Interaction in Natural Environments (AFFINE 2010) at ACM Multimedia 2010*, pages 65–70, Firenze, Italy. ACM.
- Read, R. and Belpaeme, T. (2012). How to Use Non-Linguistic Utterances to Convey Emotion in Child-Robot Interaction. In *Proceedings of the 7th International Conference on Human-Robot Interaction (HRI'12)*, pages 219–220, Boston, MA, U.S.A. ACM/IEEE.
- Read, R. and Belpaeme, T. (2013a). People Interpret Robotic Non-Linguistic Utterances Categorically. In *Proceedings of the 8th International Conference on Human-Robot Interaction (HRI'13)*, pages 209–210, Tokyo, Japan. ACM/IEEE.
- Read, R. and Belpaeme, T. (2013b). Using the AffectButton to Measure Affect in Child and Adult-Robot Interaction. In *Proceedings of the 8th International Conference on Human-Robot Interaction (HRI'13)*, pages 211–212, Tokyo, Japan. ACM/IEEE.
- Read, R. and Belpaeme, T. (2014a). Non-Linguistic Utterances Should be Used Alongside Language, Rather than on their Own or as a Replacement. In *Proceedings of the 9th International Conference on Human-Robot Interaction (HRI'14)*, Bielefeld, Germany. ACM/IEEE.
- Read, R. and Belpaeme, T. (2014b). Situational Context Directs How People Affectively Interpret Robotic Non-Linguistic Utterances. In *Proceedings of the 9th International Conference on Human-Robot Interaction (HRI'14)*, Bielefeld, Germany. ACM/IEEE.
- Reeves, B. and Nass, C. (1996). *The Media Equation: How People Treat computers, Television, and New Media Like Real People and Places*. CSLI Publications, Stanford, California, U.S.A.
- Remez, R., Rubin, P., Pisoni, D., and Carrell, T. (1981). Speech Perception Without Traditional Speech Cues. *Science*, 212:947–950.
- Repp, B. (1984). Categorical Perception: Issues, Methods, Findings. *Speech and language: Advances In Basic Research and Practice*, 10:243–335.
- Riek, L. (2012). Wizard of Oz Studies in HRI: A Systematic Review and New Reporting Guidelines. *Journal of Human-Robot Interaction*, 1(1):119–136.
- Robins, B., Dickerson, P., Stribling, P., and Dautenhahn, K. (2004). Robot-mediated joint attention in children with autism: A case study in robot-human interaction. *Interaction Studies*, 5(2):161–198.
- Ros Espinoza, R., Nalin, M., Wood, R., Baxter, P., Looije, R., Demiris, Y., and Belpaeme, T. (2011). Child-robot interaction in the wild: Advice to the aspiring experimenter. In *Proceedings of the 13th International Conference on Multimodal Interfaces (ICMI'11)*, pages 335–342, Valencia, Spain. ACM.

- Rosenthal-von der Pütten, A. M., Krämer, N. C., Hoffmann, L., Sobieraj, S., and Eimler, S. C. (2012). An Experimental Study on Emotional Reactions Towards a Robot. *International Journal of Social Robotics*, 5(1):17 – 34.
- Russell, J. (1980). A Circumplex Model of Affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.
- Saerbeck, M. and Bartneck, C. (2010). Perception of affect elicited by robot motion. In *Proceedings of the 5th International Conference on Human-Robot Interaction (HRI'10)*, pages 53–60, Osaka, Japan. ACM/IEEE.
- Salter, T., Werry, I., and Michaud, F. (2008). Going into the Wild in Child-Robot Interaction Studies: Issues in Social Robotic Development. *Intelligent Service Robotics*, 1:93 – 108.
- Scassellati, B. (2002). Theory of mind for a humanoid robot. *Autonomous Robots*, 12:13–24.
- Scherer, K. (1971). Randomized splicing: a note on a simple technique for masking speech content. *Journal of Experimental Research in Personality*, 5:155–159.
- Scherer, K. (1986). Vocal affect expression: a review and a model for future research. *Psychological Bulletin*, 99(2):143–65.
- Scherer, K. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40(1-2):227–256.
- Scherer, K. (2013). The evolutionary origin of multimodal synchronization in emotional expression. *Journal of Anthropological Sciences*, 91:1–16.
- Scherer, K. and Oshinsky, J. (1977). Cue utilization in emotion attribution from auditory stimuli. *Motivation and Emotion*, 1(4):331–346.
- Scherer, K. R. (1995). Expression of emotion in voice and music. *Journal of Voice*, 9(3):235–248.
- Scherer, K. R., Koivumaki, J., and Rosenthal, R. (1972). Minimal cues in the vocal communication of affect: Judging emotions from content-masked speech. *Journal of Psycholinguistic Research*, 1(3):269–285.
- Schouten, B., Gerrits, E., and van Hessen, A. (2003). The End of Categorical Perception as We Know It. *Speech Communication*, 41(1):71 – 80.
- Schröder, M. (2001). Emotional speech synthesis: A review. In *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech 2001)*, pages 2–5, Aalborg, Denmark.
- Schröder, M. (2003a). Experimental study of affect bursts. *Speech Communication*, 40(1-2):99–116.
- Schröder, M. (2003b). *Speech and Emotion Research: An overview of research frameworks and a dimensional approach to emotional speech synthesis*. PhD thesis, Institute of Phonetics, Saarland University.

- Schröder, M. (2004). Dimensional emotion representation as a basis for speech synthesis with non-extreme emotions. *Proceedings of the Workshop on Affective Dialogue Systems*, pages 209 – 220.
- Schröder, M., Bevacqua, E., Cowie, R., Eyben, F., Gunes, H., Heylen, D., ter Maat, M., McKeown, G., Pammi, S., Pantic, M., Pelachaud, C., Schuller, B., de Sevin, E., Valstar, M., and Wollmer, M. (2012). Building Autonomous Sensitive Artificial Listeners. *Transactions on Affective Computing*, 3(2):165–183.
- Schröder, M., Burkhardt, F., and Krstulovic, S. (2010). Synthesis of emotional speech. In Scherer, K. R., Bänziger, T., and Roesch, E., editors, *Blueprint for Affective Computing*, pages 222 – 231. Oxford University Press, Oxford, UK.
- Shiwa, T., Kanda, T., Imai, M., Ishiguro, H., and Hagita, N. (2009). How Quickly Should a Communication Robot Respond? Delaying Strategies and Habituation Effects. *International Journal of Social Robotics*, 1(2):141–155.
- Singh, A. and Young, J. (2012). Animal-Inspired Human-Robot Interaction: A Robotic Tail For Communicating State. In *Proceedings of the 7th International Conference on Human-Robot Interaction (HRI'12)*, pages 237–238, Boston, USA.
- Tapus, A., Mataric, M., and Scassellati, B. (2007). The Grand Challenges in Socially Assistive Robotics. *IEEE Robotics and Automation Magazine*, 14(1):35–42.
- Tervaniemi, M., Szameitat, A. J., Kruck, S., Schröger, E., Alter, K., De Baene, W., and Friederici, A. D. (2006). From Air Oscillations to Music and Speech: Functional Magnetic Resonance Imaging Evidence for Fine-Tuned Neural Networks in Audition. *The Journal of Neuroscience*, 26(34):8647 – 8652.
- Tinwell, A., Grimshaw, M., Nabi, D. A., and Williams, A. (2011). Facial expression of emotion and perception of the Uncanny Valley in virtual characters. *Computers in Human Behavior*, 27(2):741–749.
- Torta, E., Oberzaucher, J., Werner, F., Cuijpers, R. H., and Juola, J. F. (2012). Attitudes Towards Socially Assistive Robots in Intelligent Homes: Results From Laboratory Studies and Field Trials . *Journal of Human-Robot Interaction*, 1(2):76 – 99.
- Tuuri, K., Eerola, T., and Pirhonen, A. (2011). Design and evaluation of prosody-based non-speech audio feedback for physical training application. *International Journal of Human-Computer Studies*, 69(11):741–757.
- Wada, K. and Shibata, T. (2006). Living with Seal Robots in a Care House—Evaluations of Social and Physiological Influences. In *Proceedings of the International Conference on Intelligent Robots and Systems (IROS 2006)*, pages 4940–4945, Beijing, China. IEEE/RSJ.
- Walters, M. L., Lohse, M., Hanheide, M., Wrede, B., Syrdal, D. S., Koay, K. L., Green, A., Hüttenrauch, H., Dautenhahn, K., Sagerer, G., and Severinson-Eklundh, K. (2011). Evaluating the Robot Personality and Verbal Behaviour of Domestic Robots Using Video-Based Studies. *Advanced Robotics*, 25(18):2233–2254.

- Walters, M. L., Syrdal, D. S., Dautenhahn, K., te Boekhorst, R., and Koay, K. L. (2007). Avoiding the uncanny valley: robot appearance, personality and consistency of behaviour in an attention-seeking home scenario for a robot companion. *Autonomous Robots*, 24(2):159–178.
- Weninger, F., Eyben, F., Schuller, B. W., Mortillaro, M., and Scherer, K. R. (2013). On the Acoustics of Emotion in Audio: What Speech, Music, and Sound have in Common. *Frontiers in Psychology*, 4:1–12.
- Woods, S., Walters, M., Dautenhahn, K., and Koay, K. (2006a). Comparing Human Robot Interaction Scenarios Using Live and Video Based Methods: Towards a Novel Methodological Approach. *Proceedings of the 9th International Workshop on Advanced Motion Control (AMC'06)*, pages 750–755.
- Woods, S., Walters, M., Koay, K., and Dautenhahn, K. (2006b). Methodological Issues in HRI: A Comparison of Live and Video-Based Methods in Robot to Human Approach Direction Trials. *Proceedings of the 15th International Symposium on Robot and Human Interactive Communication (RO-MAN 2006)*, pages 51–58.
- Yilmazyildiz, S., Henderickx, D., Vanderborght, B., Verhelst, W., Soetens, E., and Lefebvre, D. (2011). EMOGIB : Emotional Gibberish Speech Database for Affective Human-Robot Interaction. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII'11)*, pages 163–172, Memphis, TN, U.S.A. Springer-Verlag.
- Yilmazyildiz, S., Henderickx, D., Vanderborght, B., Verhelst, W., Soetens, E., and Lefebvre, D. (2013). Multi-modal emotion expression for affective human-robot interaction. In *Proceedings of the Workshop on Affective Social Speech Signals (WASSS 2013)*, Grenoble, France.
- Yilmazyildiz, S., Latacz, L., Mattheyses, W., and Verhelst, W. (2010). Expressive gibberish speech synthesis for affective human-computer interaction. In *Proceedings of the 13th International Conference on Text, Speech and Dialogue (TSD'10)*, pages 584–590, Brno, Czech Republic. Springer-Verlag.
- Yilmazyildiz, S., Mattheyses, W., Patsis, Y., and Verhelst, W. (2006). Expressive Speech Recognition and Synthesis as Enabling Technologies for Affective Robot-Child Communication. *Advances in Multimedia Information Processing - PCM 2006, Lecture Notes in Computer Science (LNCS)*, 4261:1–8.
- Yngve, V. H. (1970). On getting a word in edgewise. In *Papers from the 6th Regional Meeting of the Chicago Linguistic Society*, pages 567– 578.
- Zeng, Z., Pantic, M., Roisman, G. I., and Huang, T. S. (2009). A survey of affect recognition methods: audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence*, 31(1):39 – 58.
- Zhou, K., Mo, L., Kay, P., Kwok, V. P. Y., Ip, T. N. M., and Tan, L. H. (2010). Newly Trained Lexical Categories Produce Lateralized Categorical Perception of Color. *Proceedings of the National Academy of Sciences of the United States of America*, 107(22):9974–9978.