2020-03-31

Speech Fundamental Period Estimation using a Neural Network

Howard, Ian

Studientexte zur Sprachkommunikation Band 95: Elektronische Sprachsignalverarbeitung 2020 Conference proceedings of the 31st conference in Magdeburg with 38 contributions. ISBN: 978-3-959081-93-1

SPEECH FUNDAMENTAL PERIOD ESTIMATION USING A NEURAL NETWORK

Ian S. Howard
Centre for Robotics and Neural Systems, University of Plymouth, Plymouth, PL4 8AA, UK.
UK Email: ian.howard@plymouth.ac.uk

Abstract: Here we extend previous work for the estimation of the time of excitation (Tx) from the speech signal using a shallow neural network. We make use of a dataset that consists of the simultaneously recorded speech and Laryngograph signals from drama students speaking a phonetically balanced passage. We first use the Laryngograph signal to estimate the location of vocal fold closures as a function of time. Then, by considering the problem as a supervised learning task, we train a multi-layer perceptron to map between raw speech samples, selected using a sliding input window, to a single output target sample that represents the presence or absence of an excitation point. We present result of operation across several male speakers and also demonstrate that it is possible to reconstruct the Laryngograph directly from the speech signal.

1 Introduction

1.1 Voiced speech

Voicing represents an important aspect of speech production and arises from the vibration of the vocal folds, which periodically modulate airflow from the lungs, leading to an acoustic excitation to the vocal tract. Voicing encodes both segmental as well as prosodic information in the speech signal. The latter includes intonation corresponds to the pitch the speech utterance and relates to the frequency of vibration of the vocal folds. Analysis of vocal fold activity is consequently of interest to phoneticians as an academic pursuit, as well as providing a useful feature in machine analysis of emotion and emphasis in speech utterances.

1.2 Previous work

Voiced speech is often analyzed to provide a frequency contour representation of vocal fold activity. However, such an estimate cannot capture the irregularity present in some voicing conditions, such as in creaky voice, or in pathological cases. A more complete description of voicing can be obtained by extracting temporal markers corresponding to vocal fold closure.

Previous work in the extraction of vocal fold closure was made use of a multi-layer perceptron [1]-[3], and we previously referred to the resulting algorithms as MLP-Tx. One configuration was based on a pre-processing of the speech pressure waveform using a wide-band filter bank analyser. This gave an input to the classifier which consisted of a set of adjacent time frames from the output of the filter bank. Another configuration used an input window operating directly on the speech sample inputs, much akin to the operation of a FIR filter. In both cases, the output classifier was defined as being in one of two classes. Either there was a period epoch marker at a given output frame, or there was not. To train this classifier, a good ground-truth of vocal fold closure was made from an analysis of vocal fold contact directly obtained from a Laryngograph signal [4]. Recently, we note that additional methods have be applied to this problem [5], [6].

Here we examine the operation of the MLP-Tx approach using direct operation on the speech waveform in more detail.

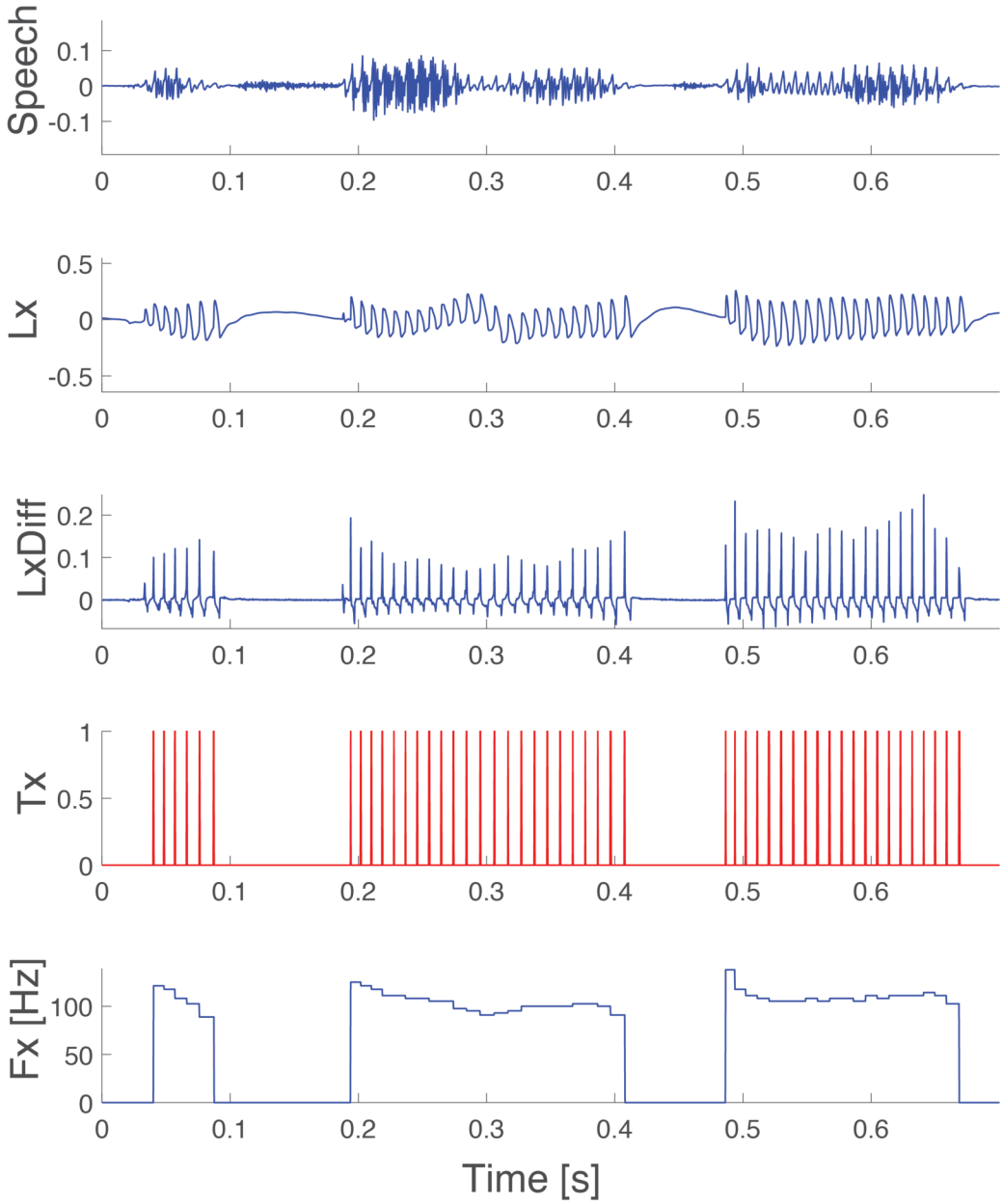

Figure 1. Example training data. Panels from the top show the speech waveform, the simultaneously recorded Laryngograph signal (Lx), the time derivative of the Laryngograph signal (LxDiff), the corresponding time of excitation markers derived from the Laryngograph signal (Tx) and the corresponding frequency contour derived from the Tx markers (Fx).

2 Methods

2.1 Dataset

We make use of a pre-recorded speech (Sp) and Laryngograph (Lx) dataset, which has been described in detail in previous publications [7], [8].

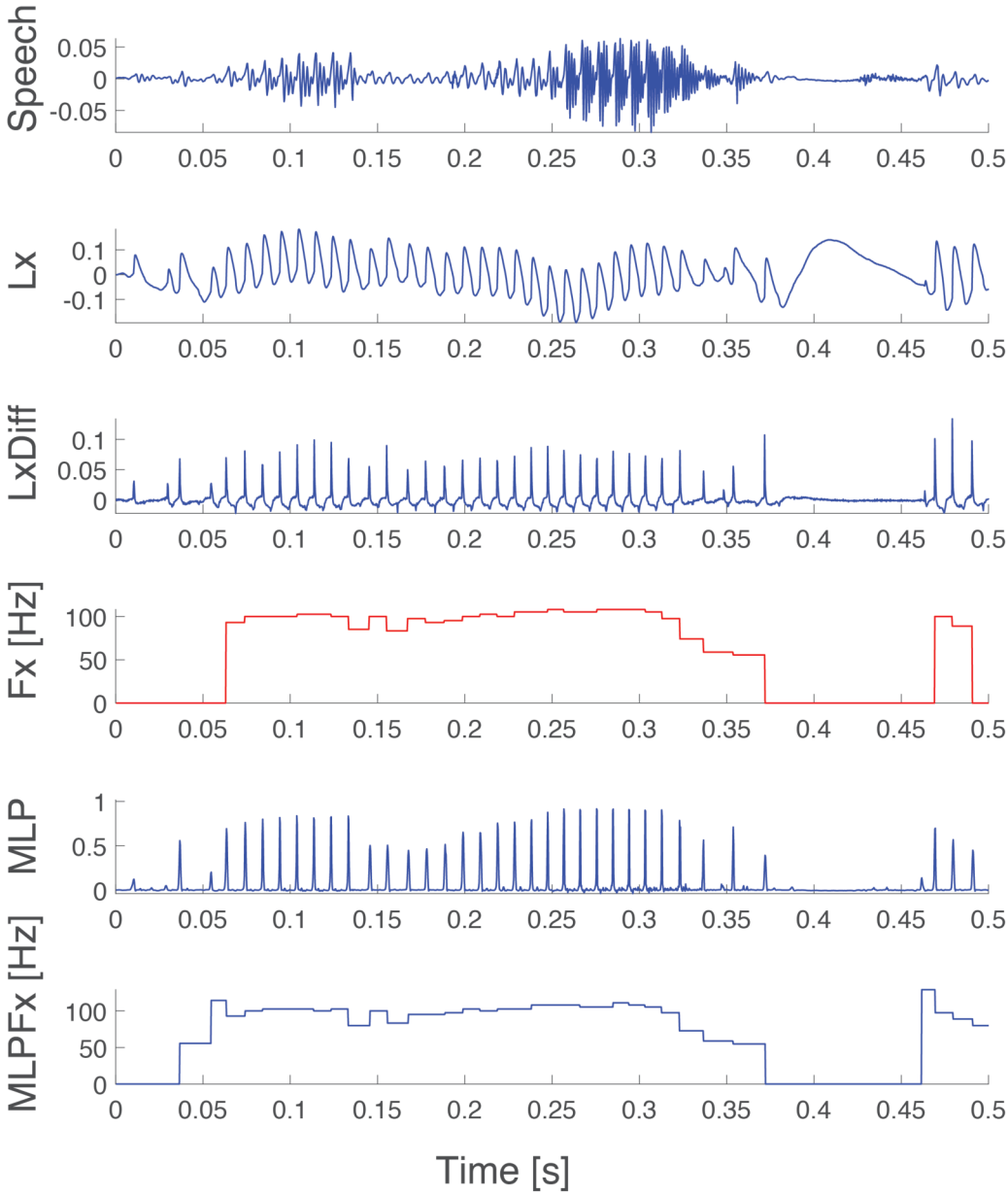

Figure 2. Example testing data and MLP-Tx algorithm output. The detector was trained at 0dB SNR with a window length of 81 samples. Panels from the top show the speech waveform, the simultaneously recorded Laryngograph signal (Lx), the time derivative of the Lx signal (LxDiff) and its corresponding frequency contour trace (Fx). Next is the raw estimated time of excitation output from the MLP-Tx algorithm (MLP) and its corresponding frequency contour trace (MLPFx).

The dataset consisted of anonymized normal voice recordings made by students reading the phonetically balanced "Arthur the Rat" passage in British English, in an anechoic chamber. Such recordings typically provided 2 minutes of data per participant. The speech signal was recorded using a Bruel and Kjaer condenser microphone. Both the raw speech and Laryngograph signals were digitized in 16-bit resolution at a sampling rate of 32kHz.

2.2 Ethics statement

The participants were all undergraduates at the London Academy of Music and Dramatic Art (LAMDA) at the beginning of their first year of study. They provided written informed consent prior to the commencement of the recording session. The experimental protocol was carried out in line with the requirements of the UCL Research Ethics Committee.

For the experiments presented here, we made use of recording of 16 male participants, although more participants were present in the dataset. In total, 8 participants were used to train the algorithms and a further 8 participants were used to evaluate the algorithms (but only 4 were used for the ROC analyses). The speech and Laryngograph data were first down-sampled to 4kHz, to reduce the subsequent processing load. The downside of this is that is also reduces the precision with which the excitation point can be located.

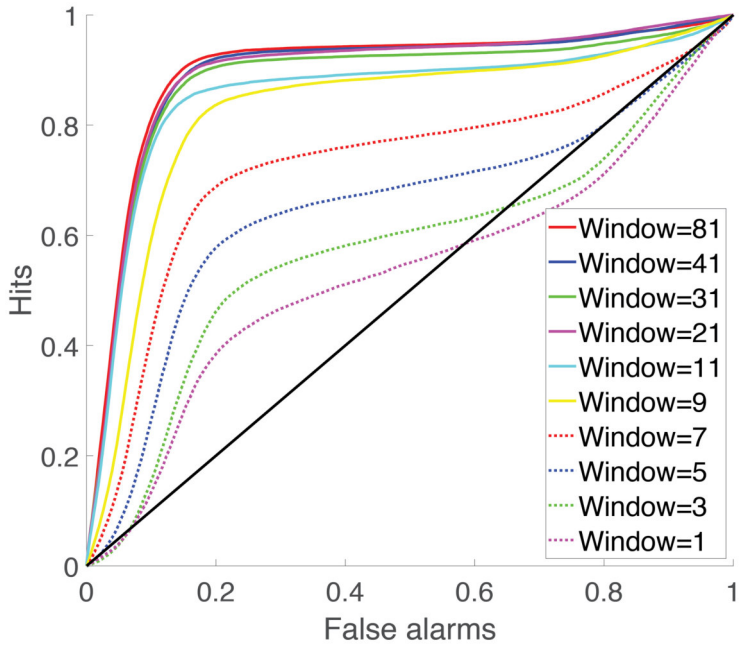

Figure 3. ROC for MLP-Tx detector as a function of window length. The detector was trained at 0dB SNR. The black diagonal line corresponds to random detector performance. The curves above it corresponds to a monatonic increase in performance as window samples increase, ranging from a window length of 1 sample for the lowest curve, to 81 for the highest curve.

2.3 Labelling the dataset with Tx markers

The speech was automatically labelled for time of excitation (Tx) using a simple algorithm that made use of the Laryngograph signal. The fundamental periods were delineated in terms of the location of closure of the vocal folds as a function of time, as defined by the location of the maximum positive differential of the Laryngograph signal. A simple threshold was then applied to the differential of the Laryngograph signal, and the exact excitation point determined by searching for the location of its maximum value. The corresponding frequency contour was computed by taking the reciprocal of the duration between subsequent Tx locations. Fig.1 illustrates all these signals on a short except of the training dataset.

The detection of Tx task from speech input is formulated as a supervised learning problem using a multi-layer perceptron. An input vector was built using a window of samples over the speech data. In the first experiment, the corresponding output target corresponded to the presence or absence of an excitation point (Tx) at sample the centre of the window. In the Laryngograph signal reconstruction experiment, the output corresponded to the value of the Laryngograph signal at sample the centre of the window. In both cases using an 81-sample input window length lead to the generation of almost 4 million training patterns.

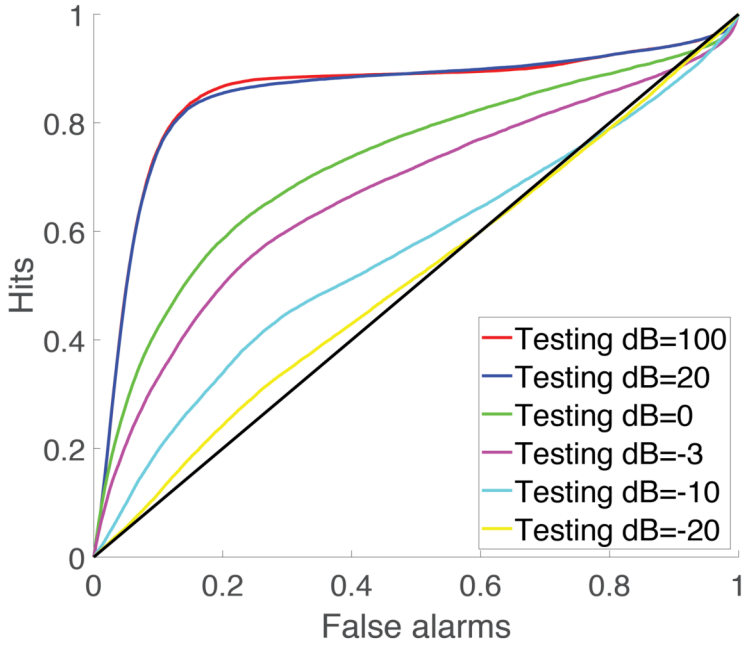

Figure 4. ROC for MLP-Tx detector as a function of signal SNR. The detector was trained at 0dB SNR with a window length of 81 samples. The black diagonal line corresponds to random detector performance. The curves above it corresponds to a monatonic increase in SNR, from -20 dB for the lowest curve to 100 dB for the highest curve.

To provide a more realistic dataset that the anechoically recorded data, Gaussian noise was added to the speech signals to give the required signal to noise ratio (SNR). The training data used a SNR of 0dB, and testing was carried out over the SNR range of -20 dB to 100 dB.

2.4 Network structure

A 2-hidden layer multi-layer perceptron was used to learn the mapping between the input speech signal and the Tx or Lx output. In both cases, after some experimentation, a network with 12 units in the 1st hidden layer and 10 in the second was used. It had a maximum input window length of 81 samples and a single linear output unit (i.e. the network structure was typically 81-12-10–1), although the window length was also a parameter that was investigated in the current study.

2.5 Quantitative comparisons

Because estimation of the time of speech excitation is formulated as the detection of an excitation point, we were able to treat the problem as a classical signal detection task. This directly lends itself to receiver operating characteristic (ROC) analysis in order to determine the fundamental performance of the MLP-Tx detector. One big advantage of this approach is that it avoids any specific threshold being used to detect Tx excitation points, and instead trades off hits against false alarms. Using this measure, good performance corresponds to achieving a large number of hits with few false alarms. In this work we made use the Matlab function roc. The main disadvantage of the ROC approach is that it required strict alignment of estimated and reference Tx locations, although alignment can be achieved using dynamic programming if a less stringent metric is required [9].

2.6 Implementation

All data processing was carried out using Matlab. The multi-layer perceptron was implemented in Matlab within the Deep Neural Network Toolbox. Training on the 8-male participant dataset was rapid: Running on a Windows 10 PC fitted with an NVIDIA GEFORCE RTX 2080Ti graphics card, training a single experimental configuration only typically took about 10 minutes.

3 Results

Fig. 2. shows an excerpt from testing data and the corresponding output of the MLP-Tx algorithm with an 81-sample input window. It can be seen that the differential of the Laryngograph signal is quite similar to the output of the multi-layer perceptron, although the latter was trained with unity targets. This can be interpreted as the MLP-Tx output giving an indication of evidence, rather than a simple binary value. In order to facilitate comparison, this figure also shows the corresponding frequency contours obtained from the laryngograph directly and also from excitation points estimated using MLP-Tx, which are quite similar.

Quantitative comparison of the MLP-Tx algorithm was made against the reference obtained directly from the Laryngograph signal using receiver operating characteristic analysis. Two sets of tests were run. Firstly, to examine the effect of window length and secondly to examine the effect of additive noise. In both cases, we ran the ROC analysis on 4 male participants not used for training. In general, we note that ROC performance on each individual participant separately was better than across all 4 taken together, suggesting different threshold values were required to optimally decode the Tx for the different participants.

3.1 Effect of window length

We first examined the effect of window length on performance. Fig. 3. shows the ROC for MLP-Tx detector as a function of window length. It can be seen that the best performance was achieved with the largest window length of 81 samples. This suggests that contextual information is required in order for the classification process to operate effectively. It seems likely that an even longer window may give even better results. However, in the current implementation, memory limitation prevented a larger window size being tested.

3.2 Effect of SNR

We then examined the effect of SNR of performance. Fig. 4. shows the ROC for MLP-Tx detector as a function of test speech SNR. It can be seen that adding additive Gaussian noise degraded performance. However, operation at 20 dB was essentially the same as at essentially noise-free 100 dB condition, illustrating the robustness of the algorithm.

3.3 Laryngograph signal reconstruction

Finally, we investigate reconstructing the laryngograph signal from the speech signal using the non-linear regression capability of the neural network. This is illustrated in Fig 5. In the small sample shown here, it can be seen that the general form of the laryngograph signal is recovered. However low frequency fluctuations absent from the reconstruction. This is not surprising since they are not relevant to acoustic output. To demonstrate the extent to which vocal fold closure dynamics are preserved in the reconstruction, we differentiated the reconstructed signal. It be can see that the differential from the reconstruction differs somewhat from that of laryngograph signal, illustrating reconstruction is not perfect.

4 Discussion

4.1 Summary

In this work we trained a multi-layer perceptron to estimate the time of excitation of a speech signal arising from vocal fold closure. The problem was formulated as a supervised learning task, with an input pattern coming from the speech signal and the output target corresponding to the absence or presence of a speech excitation point at the centre of the window. We extended previous results by first examining the effect of window size and found a window of 81 samples to give the best results. We were unable to test larger window sizes due to memory limitation, although this issue could easily be overcome in future implementations. We also examined the effects of additive noise on the performance of the algorithm. As expected, performance degraded as the signal-to-noise ratio went down. However, the algorithm proved to be quite robust and able to deal with 20 dB signal-to-noise ratio almost as well as a noise-free condition.

4.2 Conclusions

A strength of the MLP-Tx algorithm is that the fundamental period estimations are made on a cycle-by-cycle basis and therefore irregularities in vocal fold vibration can be detected by the algorithm, whereas most frequency domain algorithms tend to smooth the period values. Creaky voice can be dealt with effectively using the MLP-Tx algorithm, whereas many other algorithms treat this important larynx excitation as being unvoiced due to its intrinsic irregularity. We note also that the MLP-Tx algorithm is well suited for real-time implementation because of the simple uniform structure of the MLP, and the inherently small (here 10ms) input to output delay.

4.3 Future work

The results showing the longer analysis window proved helpful illustrates the need of contextual information in the detection of the speech excitation point. In the future we will therefore also investigate the use of recurrent networks, such as LSTM and others, since such architectures can can intrinsically deal with temporal context in an efficient manner.

The observation that different threshold values were required to optimally decode the Tx locations for different participants points to a deficiency of the current analysis, since ideally participants should result in the same output levels from the MLP-Tx algorithm. This suggests that deeper and more non-linear networks may give rise to better overall results and will be investigated in future implementations of the MLP-Tx algorithm.

In the future we will also investigate training and testing using both male and female participants.

5 Acknowledgements

We thank Adrian Fourcin at UCL for providing the LAMBDA /RADA dataset. We thank the University of Plymouth for support.

6 References

[1] WALLIKER, J. AND HOWARD, I. S.; “Real-time portable multi-layer perceptron voice fundamental-period extractor for hearing aids and cochlear implants,”, Speech Communication, 1990.

[2] HOWARD, I. S.; “Speech fundamental period estimation using a trainable pattern classifier,” *Proc Speech'88: 7th FASE Symposium*, 1988.

[3] HOWARD, I. S.; "Speech fundamental period estimation using pattern classification," *PhD Thesis, University of London*, Oct. 1991.
[4] FOURCIN, A. AND ABBERTON. E.; "First applications of a new laryngograph," *Medical & biological illustration*, vol. 21, pp. 172–182, 1971.
[5] MATOUSEK, J. Tihelka, D.; Interspeech, 2018, "Glottal Closure Instant Detection from Speech Signal Using Voting Classifier and Recursive Feature Elimination.," *In Interspeech (pp. 2112-2116).*
[6] MATOUSEK, J. Tihelka, D.; International, 2019, "Using Extreme Gradient Boosting to Detect Glottal Closure Instants in Speech Signal," *In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6515-6519). IEEE*
[7] GARCIA, S.; *An analysis of the effects of nine months of vocal training on the voice*. University of London, University College London 2006.
[8] FOURCIN, A.; "A note on voice timing and the evolution of connected speech.," *Logoped Phoniatr Vocol*, vol. 35, no. 2, pp. 74–80, Jul. 2010.
[9] HOWARD, I. S. AND HOWARD, D. M.; "Quantitative comparisons between time domain speech fundamental frequency estimation algorithms," presented at the Institute of Acoustics, Windemere, 1986.

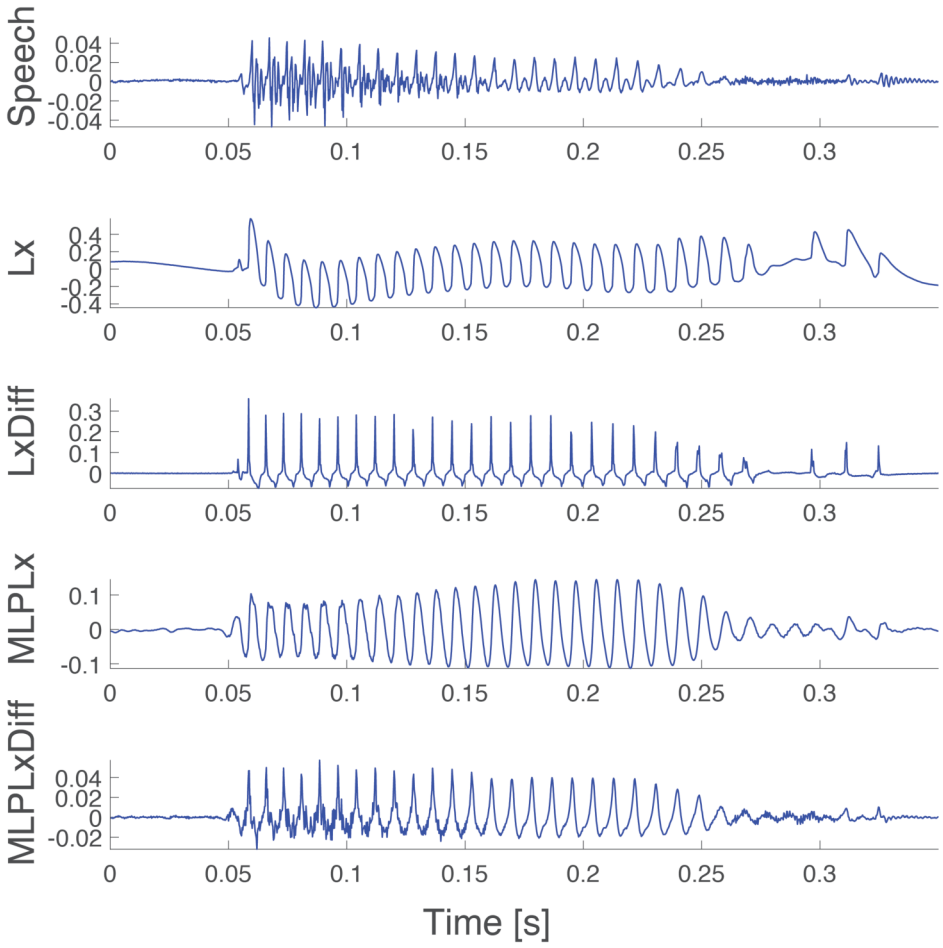

Figure 5. Example testing data and Laryngograph reconstruction using the MLP network. The detector was trained at 0dB SNR with a window length of 81 samples. Panels from the top show the speech waveform, the simultaneously recorded Laryngograph signal (Lx), time derivative of the Laryngograph signal (LxDiff), reconstructed Laryngograph signal from the MLP (MLPLx) and the differential of the reconstructed Laryngograph signal (MLPLxDiff).