MACHINE LEARNING ANALYSIS OF SPEECH AND EGG FOR THE DIAGNOSIS OF VOICE PATHOLOGY

Ian S. Howard[1], Julian McGlashan[2] & Adrian J. Fourcin[3]

[1]Centre for Robotics and Neural Systems, University of Plymouth, Plymouth, PL4 8AA UK

[2]ENT Department, Nottingham University Hospitals UK

[3]Laryngograph Ltd, Wallington UK, Emeritus Professor at UCL, UK

ian.howard@plymouth.ac.uk

ABSTRACT: Current approaches to voice diagnosis involve a clinician examining the patient, listening to their voice and in some cases, using additional measurements of the larynx such as EGG. Here we train a feedforward convolutional neural network on a database of normal healthy drama students recorded speaking passages in English, to reconstruct the associated EGG (Lx) waveform. We then use the network to predict the EGG from the acoustic speech signal on a different set of speakers, including ones that exhibit laryngeal pathologies. We show the predicted EGG is very similar to the actual recorded EGG and, as such, can provide a useful indication of voice pathology. Importantly, the network is able predict the pathological EGG waveforms even though it was never trained on pathological speech.

1 Introduction

1.1 Electroglottography

Electroglottography (EGG) is an electrical impedance-based measurement technique that yields physiological information on vocal fold contact and vibratory patterns using electrodes placed externally on the neck over the patient's larynx. At present in clinical environments, access to such information can only be obtained from EGG since systems that reliably obtain EGG estimates or even explicit estimates of voice pathology from the acoustic signal are not available in commercial products. Consequently, EGG is employed worldwide to assist the diagnosis of laryngeal pathologies. However, EGG cannot be used with all patients or in a telemedicine setting.

1.2 Previous work

Early work [1]-[2] was the first to show that it is possible to train a neural network on data labelled using the associated EGG waveform [3] to determine glottal closure directly from the speech signal. Based on this early work, real-time estimation of the fundamental period of vocal fold oscillation was also possible [4]. Due to advances made in deep neural networks, it has recently been shown that it is now possible to accurately reconstruct the entire EGG signal directly from the acoustic speech signal [5]-[6]. Similar work using a shallow neural network was also carried out to extract vocal fold closure points and also reconstruct the EGG [7].

To assess the condition of the larynx, patients experiencing voice pathologies are generally asked to read specially developed texts and utterance, in which pathological effects manifest themselves. Here we build and test a system to estimate EGG from the acoustic speech signal on one such text using a convolutional neural network. We show that it is possible to reconstruct EGGs for several normal and pathological speakers. For the pathological speakers, we show that the predicted EGG is indicative of the voice pathology that is present.

2 Methods

2.1 Dataset

We first consolidated currently available normal and pathological speech and EGG datasets from normal healthy and clinical populations. The normal healthy speech consisted of high-quality 16-bit recording of first-year undergraduate drama students reading the phonetically balanced "Arthur the Rat" passage in British English [8]. The speech signal was recorded using a Bruel and Kjaer condenser microphone and the speech and Laryngograph signals (EGG using guard ring electrodes) were digitized in 16-bit resolution at a sampling rate of 32kHz. The clinical data consisted of 3 normal healthy speakers and 5 that exhibited voice pathologies, recorded in a normal clinical environment at either 48 kHz or 16 kHz.

2.2 Preprocessing the data

Firstly, Gaussian noise was added to the training data to achieve a SNR of 20 dB, in order to make the anechoic data more representative of real-world speech signals. The speech and Laryngograph data were then down-sampled to 4kHz. This was performed to reduce the processing load needed to train the network and also to ensure all datasets were processed at the same sampling rate. The data was subsequently high-pass filtered at 25Hz with a bidirectional 4-pole Butterworth filter, to remove low-frequency components in the EGG signal arising from electrode movement. Clearly such effects are not observable from the speech signal and would only constitute a hindrance to EGG reconstruction and introduce an offset that cannot be reconstructed. Similarly, slowly changing pressure measurements from the microphone were removed by this pre-processing. Finally, the means of each participant's speech and EGG signals were subtracted and signal amplitude was divided by signal's standard deviation. In total 49 speakers from the normal drama student dataset (33 male, 16 female) were used to train the network. This constituted about 100 minutes of training data in total. The network was evaluated on an additional 4 participants (male) from the same drama student dataset, as well as all 8 of a clinical dataset speaker: 3 normal (male) and 5 pathological (2 male, 3 female).

2.3 Network structure

To map between the acoustic speech signal and the EGG (Laryngograph waveform), a convolutional neural network (CNN) was used to implement non-linear regression. The network consisted of an input window of 81 adjacent speech samples, 2 convolutional layers of 50 nodes with input width 20 and ReLu output activations, and finally a fully connected layer with a regression output. It was trained using the Adam optimizer using a minibatch size of 1024. This network architecture was an initial guess at a suitable network structure and better architectures for this task almost certainly exist. We note that a fully connected network with 81 inputs, 2 fully connected layers with 50 hidden nodes and ReLu activations functions and an output regression layer generated comparable results, although they are not shown in this paper.

2.4 Implementation

The CNN was implemented in MATLAB within the Deep Neural Network Toolbox. Training was performed on a Windows 10 PC fitted with an NVIDIA GEFORCE RTX 2080Ti graphics card. Training was run for 4 hours, although longer training may have improved performance.

3 Results

The results presented here are representative of the testing dataset.

3.1 Laryngograph signal reconstruction on normal speech

Fig. 1. Shows testing carried out on a normal voice male speaker from the student dataset, who was not used for network training. It can be seen that the Lx signal in the middle of the second part of the signal indicated that contact substantially reduced, even though voiced excitation clearly continues.

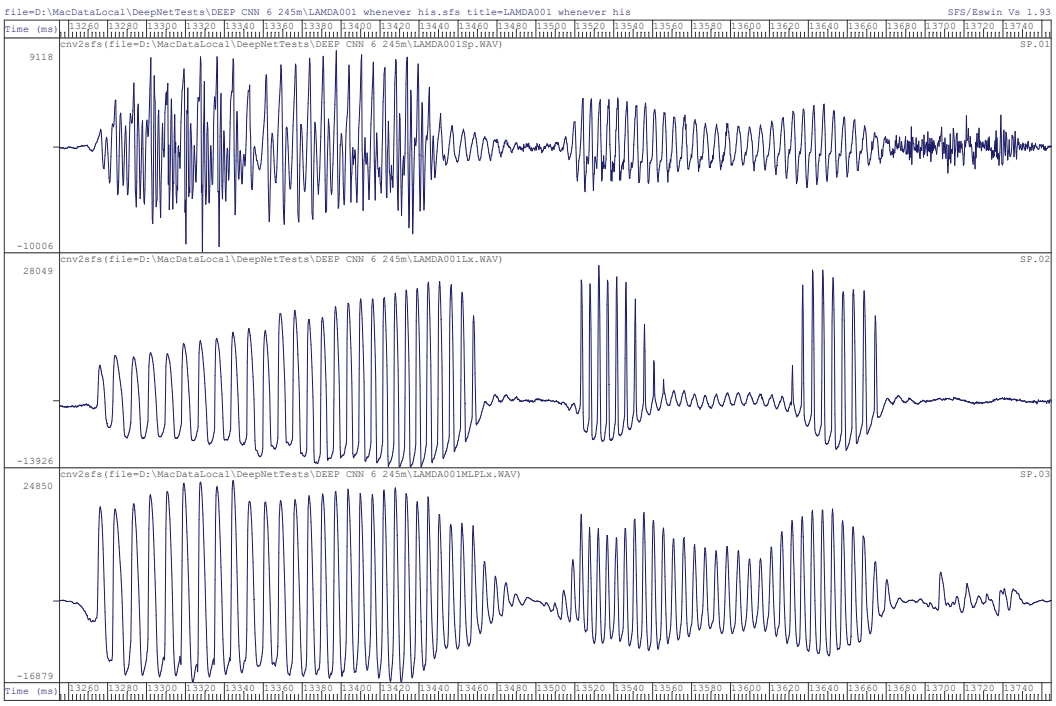

Figure 1. Example testing for normal male speaker from the student dataset who was not used to train the CNN. Panels from the top show the speech waveform, the simultaneously recorded Laryngograph signal (Lx), and the Lx signal estimated using a CNN. It can be seen that the Lx signal in the second part of the signal indicated that contact has substantially reduced even though voiced excitation continues. However the CNN estimate generalizes and generates an output in this region and shows offset and onset.

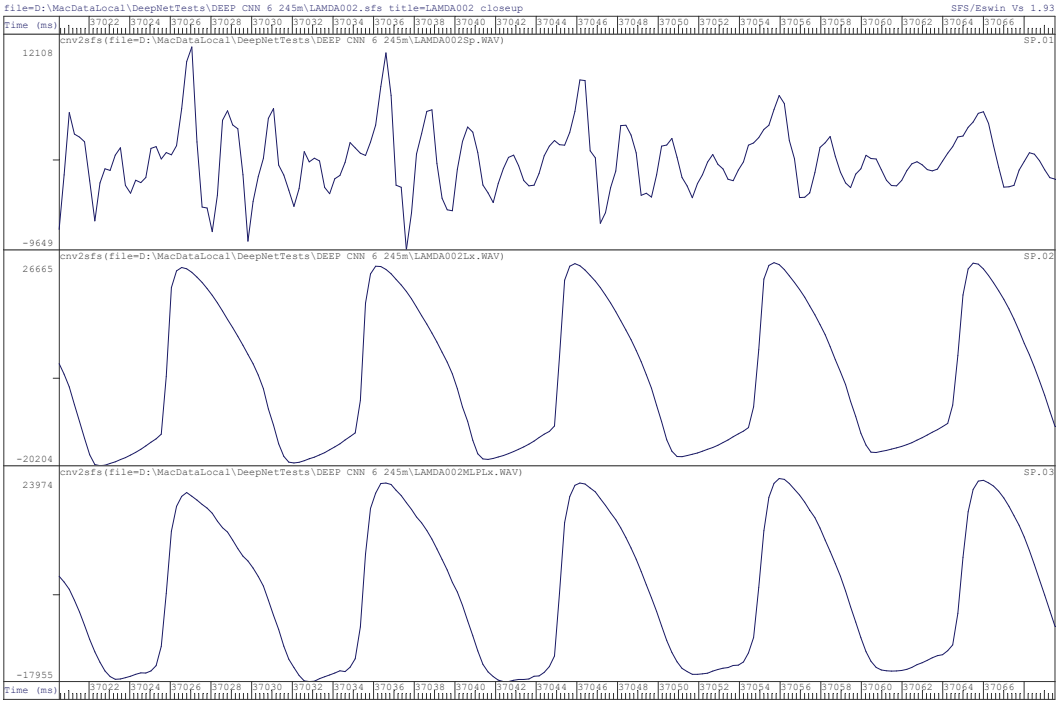

Figure 2. Example from another normal male speaker from the student dataset who was not used to training the CNN. Panels from the top show the speech waveform, the simultaneously recorded Laryngograph signal (Lx), and the Lx signal estimated using a CNN. Close-up shows that the Lx and CNN Lx estimate waveform shapes are very similar.

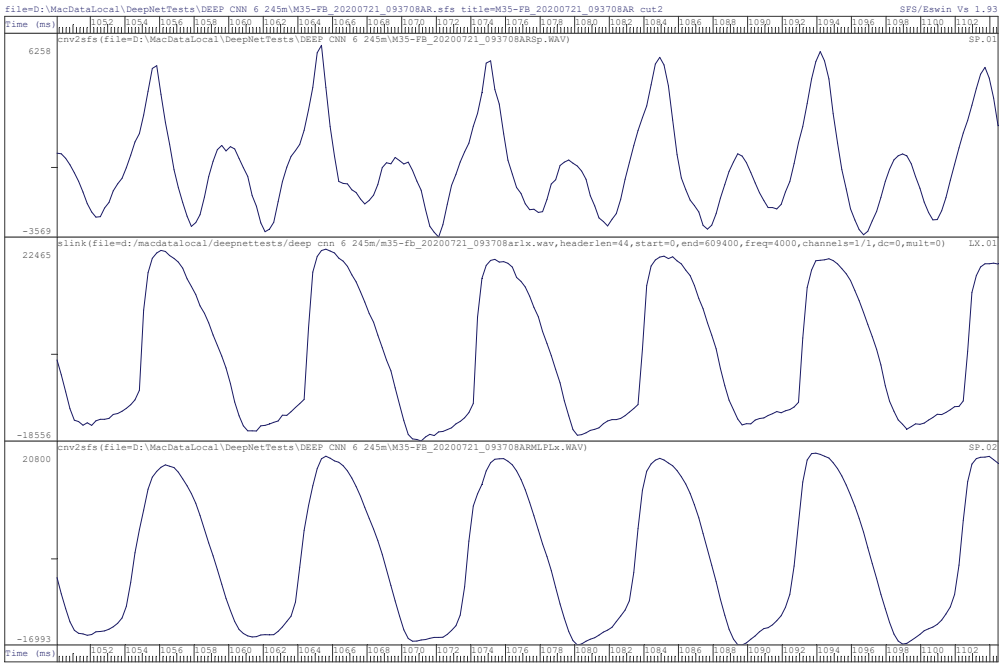

Figure 3. Example normal male speaker from the clinical dataset. Panels from the top show the speech waveform, the simultaneously recorded Laryngograph signal (Lx), and the Lx signal estimated using a CNN. Close-up shows that the Lx and CNN Lx estimate wafeform shapes are very similar. The Lx and predicted Lx cycles are very similar in shape.

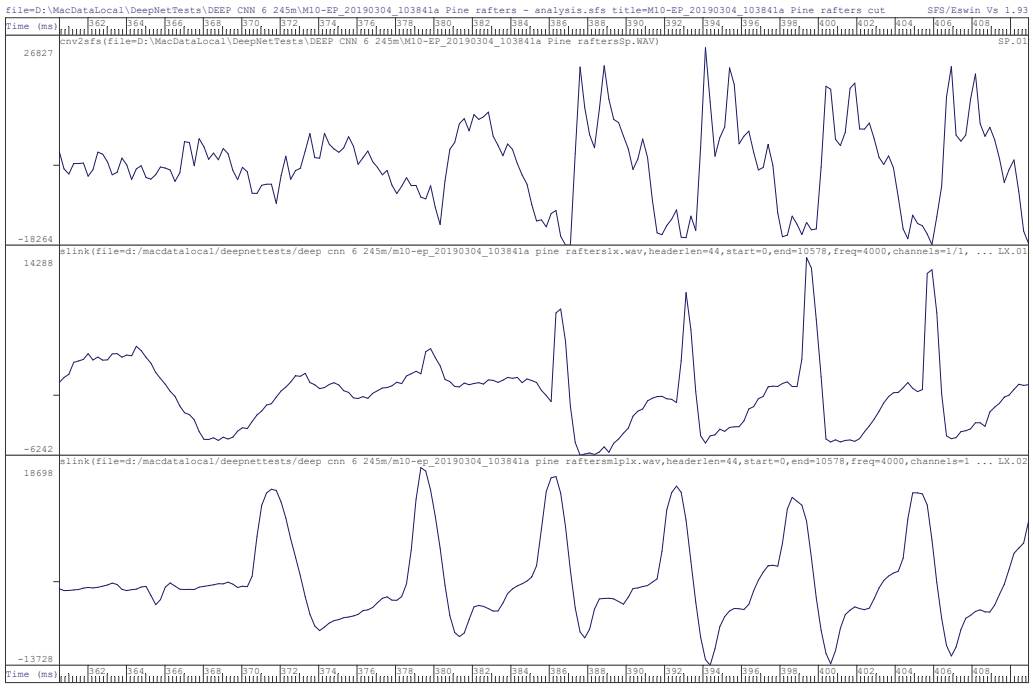

Figure 4. Irregular speech for male patient with Vocal Cord Palsy. Panels from the top show the speech waveform, the simultaneously recorded Laryngograph signal (Lx), and the Lx signal estimated using a CNN. It can be seen that there is limited contact shown in the Lx signal. Note that the predicted Lx from the CNN also captures the shape of the Lx cycles. It also accurately infers the condition where little contact is made.

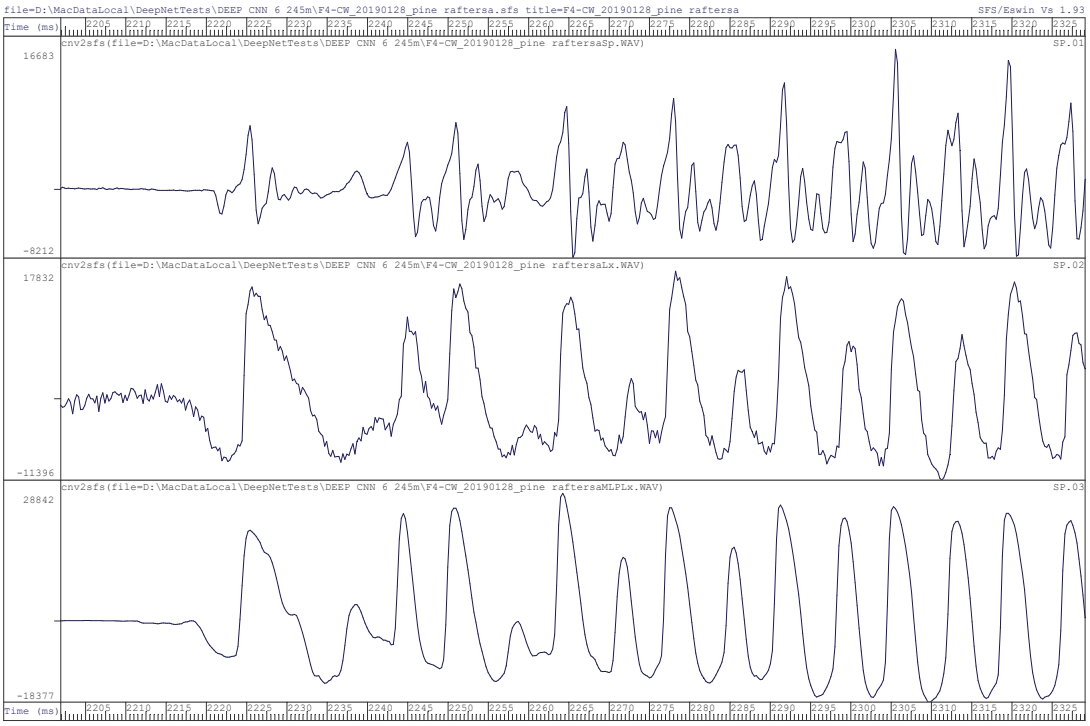

Figure 5. Speech for a female patient with Muscle Tension Dysphonia. Panels from the top show the speech waveform, the simultaneously recorded Laryngograph signal (Lx), and the Lx signal estimated using a CNN. It can be seen that there is limited contact shown in the Lx signal. Note that the lack of contact in the Lx waveform is also present in the CNN Lx prediction.

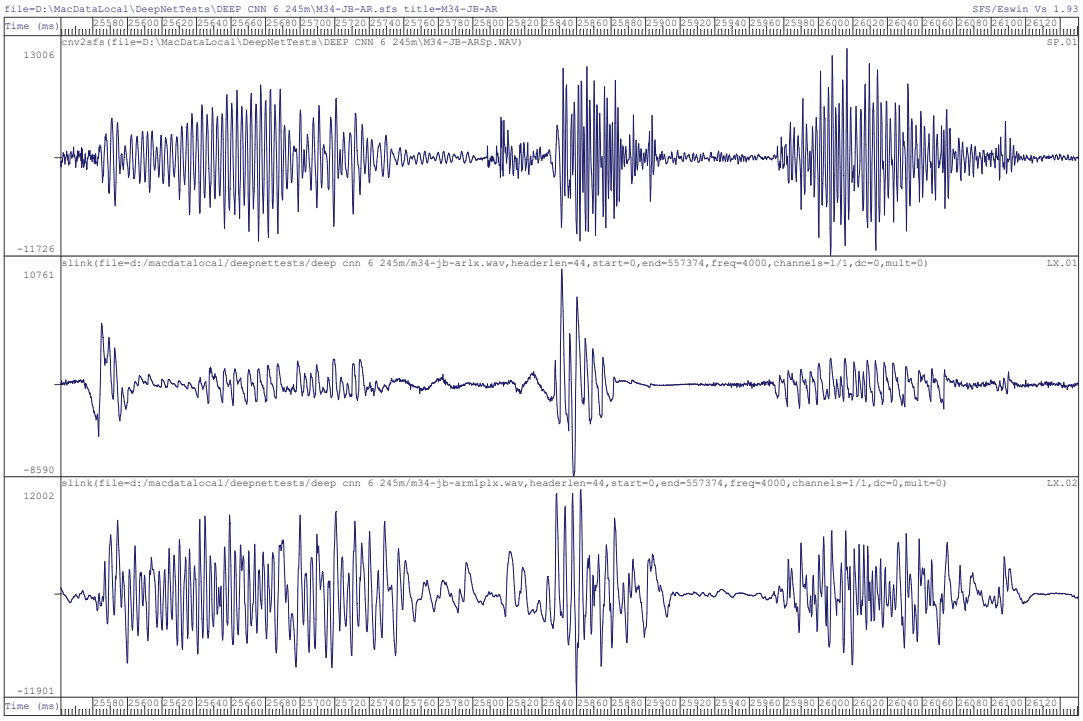

Figure 6. Speech for male patient with Adductor Spasmodic Dysphonia. Panels from the top show the speech waveform, the simultaneously recorded Laryngograph signal (Lx), and the Lx signal estimated using a CNN.

However, in this case the CNN estimate generalized and generated an output in this region.

Fig, 2. shows testing on another normal male speaker from the student dataset, who was not used to train the network. The Lx and predicted Lx waveform cycles are very similar in shape, although the estimated Lx appears smoother.

Fig.3. shows an example of a normal male speaker from the clinical dataset that was used to test network operation. Once again, the Lx and predicted Lx waveform cycles are very similar in shape.

3.2 Laryngograph signal reconstruction on pathological speech

Fig.4. shows the incomplete vocal fold closure for a male patient with Vocal Cord Palsy and the shape of Lx is well replicated in the CNN estimate of EGG.

Fig.5. shows a lack of contact in the Lx waveform, which is also quite irregular, for a female patient with Muscle Tension Dysphonia and these features are also present in the CNN Lx prediction.

Fig.6. shows the rather chaotic excitation in a male patient with Adductor Spasmodic Dysphonia is replicated in the CNN estimate of EGG.

3.3 Laryngeal state observability

The sample-by-sample analysis of the acoustic speech signal by the CNN was able to effectively replicate EGG, even when it became very irregular. In particular, the predicted EGG also exhibited the cycle-by-cycle shape characteristics of the EGG waveform.

From a control engineering and reinforcement leaning perspective, one issue that is apparent by consideration of the speech and EGG waveforms, is that the the laryngeal system is not generally fully observable. For example, it is not possible to fully estimate the state of the vocal folds just by consideration to the EGG waveform. Firstly, EGG is a 1-dimensional signal and cannot distinguish the effect of the 2 vocal folds. Also, since EGG relates to changes in current flow and contact, when no contact is made, little modulation of the EGG signal is observed. This is illustrated in Fig. 1. Here, consideration to the speech waveform shows that voiced excitation due to vibration of the vocal folds is still present. In such conditions vocal fold vibration is unobservable in the EGG. Conversely, the EGG signal provides a very good estimation of vocal fold contact, the effects of which manifest themselves less directly in the acoustic signal.

4 Discussion

4.1 Summary

We trained a convolutional neural network (CNN) to reconstruct the EGG waveform from just the acoustic speech signal. We showed that a convolutional neural network was able to generate good estimates of EGG on normal and pathological speech that was not used in training.

4.2 Limitation of the current study

Although respectable results were achieved, the network used here was a first guess as a suitable structure. In particular the network used was quite shallow. Deeper networks and, for example, ResNets, as well as networks including recursive layers, such as LSTMs, may give even better results. In the future, both the network architecture and its hyperparameters will be investigated and further optimized.

In the current work, the network was only trained on a limited quantity of normal speech. As is well known in work on deep learning, additional data is likely to improve performance and also make training deeper network architectures viable.

The value of EGG (which would also include EGG predicted from the acoustic signal) in the diagnosis of laryngeal pathologies is well recognized, but it is currently very cumbersome to use clinically. This is because clinicians have to identify speech regions by hand in order to estimate the presence of any abnormality. Thus, although evaluating voice disorders with the assistance of EGG recordings is possible, at present the analytical capability within clinical practices are rudimentary and most clinical information is not extracted from the EGG signal. This is because manual data labelling to identify epochs that are indicative of laryngeal pathologies is clinically impractical. Such critical regions are, for example, transitions from vocalic to fricative sounds and vice-versa. Automatic detection of critical parts of speech would enable analysis metrics relating to pathology be made automatically. Therefore, a system that can identify regions of interest on the one hand, and extract useful metric directly from acoustic speech signal alone using an EGG-trained network model, would be of great value.

Future work will also make use of speech from pathological as well as normal speakers. Such an approach then has the potential to automatically detect the presence and also the locations of where speech pathology occurs in speech utterances. Indeed, such an approach would also provide a means for a neural network-based system to create specific features to extract indications of speech pathology itself directly from the speech (and also EGG) waveforms. This would avoid the need for using less effective hand-crafted features, such as those based on simplistic concepts of open quotient, and so on. Thus, training on pathological and normal speech should kill two birds with one stone in both identifying and locating pathological conditions in speech utterances.

Longer range effects that manifest themselves over lengthier timescales than individual excitation epochs and relate to prosodic and breathing aspect of speech productions are also likely to provide a source of useful information in quantifying voice pathology. Such issues will also be investigated in the future.

4.3 Practical applications

At present, EGG cannot be used with all patients. Moreover, EGG requires a patient physically visiting a specialist. Performing an automatic analysis of the patient's condition on the basis of only the acoustic signal would therefore be very beneficial. Speech-only operation will also widen the area of application to include other fields where EGG devices cannot be accessed. For example, in occupational settings for professional voice users, such as class-room teachers, or to enable longitudinal studies to be run with repeat measurements over a period of time. Speech-only analysis of voice also opens-up the possibility of remote diagnosis, since speech input could potentially also be recorded using smartphone technology.

4.4 Sensory integration

The fact that EGG can be predicted from the speech signal indicates that much information obtained from the EGG is present in the speech signal. This suggest that there is now a strong opportunity of building commercial systems for estimating EGG in a clinical setting without actually recording it directly. Since much EGG information is present in the speech signal, it also indicates that measures of voice quality and voice pathology that currently make use of EGG can almost certainly be directly estimated from the acoustic signal.

However, taking a sensory-integration perspective, we suggest that providing a suitable processing structure is used, using both speech and EGG, is likely to give a better estimate of vocal condition than using just a single signal. Additional measurements, even when noisy or intermittently uninformative, are still useful. Indeed, incorporating a video stream of vocal fold

vibration could also be integrated into such a system, further improving the estimate of what is going on at the vocal folds and indeed provide a direct estimate of pathology. Using commercially available neural network frameworks, such as those integrated in MATLAB, designing networks to realize such multi-sensory integration across modalities would be a simple task. Having generated a more complete estimate of vocal fold activity and state using such an approach, it may then subsequently be possible to estimate it directly too from the speech signal by means of another neural network.

4.5 Model free versus model-based analysis

Neural networks represent a model-free black box approach to solving problems in regression and classification. They do not assume any particular data model but rather constitute an empirical model that fits the available data by optimizing parameters in their computational architectures. In contrast, it's also possible to adopt a model-based approach to data analysis. The latter can involve building a physical model of the system and using optimization or reinforcement learning approaches to fit the model's parameters to the data. One advantage of the latter approach is that it can make the operation more understandable. Indeed, understandable AI has now become a big theme because it is now widely recognized that it is not only important to make decision or assessments of decisions made on the data. It is also important to justify why decisions were made. Consequently, future work will also investigate the use of model-based approaches to fit models of the vocal fold and larynx to the data, and identify pathologies when parameter values exceed those seen in models of normal phonation.

5 Acknowledgements

We thank Professor Mark Huckvale (University College London) and Professor Reza Nouraei (University of Southampton) for valuable discussion on the themes and ideas addressed in this paper. We also thank the University of Plymouth for supporting Dr Ian Howard.

6 References

[1] HOWARD, I.S., Speech fundamental period estimation using a trainable pattern classifier, Proc Speech'88: 7th FASE Symposium, 1988.
[2] HOWARD, I.S. "Speech fundamental period estimation using pattern classification," PhD Thesis, University of London, Oct. 1991.
[3] FOURCIN A. J. & ABBERTON E,, First applications of a new laryngograph, Medical & biological illustration, vol. 21, pp. 172–182, 1971.
[4] WALLIKER J.R. AND HOWARD I.S., "Real-time portable multi-layer perceptron voice fundamental-period extractor for hearing aids and cochlear implants," 1990.
[5] REDDY, M. G. & T. M. K. S. RAO: Glottal closure instants detection from pathological acoustic speech signal using deep learning. In Proceedings of the Machine Learning for Health Workshop. 2018.
[6] PRATHOSH, A.P. SRIVASTAVA, V, & MISHRA, M. Adversarial approximate inference for speech to electroglottograph conversion. IEEE/ACM Transactions on Audio, Speech, and Language, 2019
[7] HOWARD, I.S, Speech fundamental period estimation using a neural network, ESSV, Magdeburg, Germany, 2020
[8] GARCIA, S.; An analysis of the effects of nine months of vocal training on the voice. University of London, University College London 2006.