Book Section:
Gonzalez, J.A., Cheah, L.A., Green, P.D. et al. (2017) Restoring Speech Following Total Removal of the Larynx. In: Harnessing the Power of Technology to Improve Lives. Studies in Health Technology and Informatics, 242. IOS Press, pp. 314-321. ISSN: 0926-9630.

https://doi.org/10.3233/978-1-61499-798-6-314

Restoring speech following total removal of the larynx

Jose A. GONZALEZ [a,1], Lam A. CHEAH [b], Phil D. GREEN [a], James M. GILBERT [b], Stephen R. ELL [c], Roger K. MOORE [a], and Ed HOLDSWORTH [d]
[a] *Department of Computer Science, University of Sheffield, Sheffield, UK*
[b] *School of Engineering, University of Hull, Kingston upon Hull, UK*
[c] *Hull and East Yorkshire Hospitals Trust, Castle Hill Hospital, Cottingham, UK*
[d] *Practical Control Ltd, Sheffield, UK*

Abstract. By speech articulator movement and training a transformation to audio we can restore the power of speech to someone who has lost their larynx. We sense changes in magnetic field caused by movements of small magnets attached to the lips and tongue. The sensor transformation uses recurrent neural networks.

Keywords. Speech restoration, silent speech interfaces, speech synthesis, permanent magnet articulography

1. Introduction

In 2010 it was reported that, worldwide, more than 425,000 people were still alive up to 5 years after being diagnosed with laryngeal cancer [1]. This type of cancer only accounts for 1% of all cancers [2], but it has a high 5-year survival rate (around 70% according to [3]). Patients who undergo total laryngectomy as a treatment for laryngeal cancer will inevitably lose the power of speech. As speech is a vital part of human communication, post-laryngectomy patients often find themselves struggling with their daily communication, which can lead to social isolation, feelings of loss of identity and clinical depression [4-6].

Currently, there are 3 methods available for speech restoration after total laryngectomy: the electrolarynx, oesophageal speech and valved speech. The electrolarynx or artificial larynx is a handheld vibrating device which is placed against the neck to provide excitation of the vocal tract. The electrolarynx is relatively cheap and easy to use, but requires manual dexterity and produces an unnatural, mechanical voice. Oesophageal speech is a type of alaryngeal speech which does not require any instrumentation. In oesophageal speech, the person injects air into the upper oesophagus and then releases it in a controlled manner making the oesophagus to vibrate in order to create the speech sounds (i.e. it is like a controlled belch). This method, however, is difficult to learn and has a low speaking rate. In valved speech, which is considered to be the current gold standard, a one-way valve is inserted in the wall separating the trachea

[1] Jose A. Gonzalez, Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello St., S1 4DP Sheffield, United Kingdom; E-mail: j.gonzalez@sheffield.ac.uk.

and oesophagus. The valve allows air from the lungs to go into the oesophagus without food and liquids passing into the trachea. This method provides the most natural sounding voice among the three, but it is a masculine voice unpopular with female patients and it requires regular hospital visits for valve replacement due to biofilm formation. In addition to the three methods above, Alternative and Augmentative Communication (AAC) devices can be also employed to enable communication to laryngectomees, however, communication using AAC devices is normally much slower than standard spoken communication.

Silent speech interfaces (SSIs) [7] have emerged in the last decade as an alternative communication method when the acoustic speech signal is not desirable (e.g. to maintain privacy when speaking in public places) or not available (e.g. after laryngectomy). To enable speech communication, SSIs rely on non-acoustic signals generated when the person articulates speech sounds, such as electrical activity in the brain or the electrical activity driving the articulator muscles or the movement of the speech articulators. From these signals, a SSI tries to automatically recover the speech produced by the person. A human example of this is lip reading. SSIs can be used as assistive technology (AT) to restore the ability to speak to people who have lost their voices after disease or trauma.

In this paper, a SSI system aimed at speech restoration after total laryngectomy is described. The two pillars of the proposed system are (i) a device for capturing the movement of the articulators while the person articulates words and (ii) a speech synthesis technique driven by the captured articulatory data. Articulatory data is acquired using a technique known as Permanent Magnet Articulography (PMA) [8-11]. In PMA, a set of small magnets are attached to the articulators (typically the lips and tongue) and the variations of the magnetic field generated by the magnets during speech articulation are captured by sensors located close to the mouth. To synthesise speech from PMA data, an artificial neural network [12,13] is trained to convert sensor data into acoustics. The neural network is trained with simultaneous recordings of PMA and speech data made by the person before she/he loses the voice. This method is suitable for real-time processing and, because it is trained with recordings of the person's own voice, retain the speaker's vocal identity: to approximate their own voice.

To evaluate the potential of the proposed SSI for speech restoration, some preliminary results are reported here for normal speakers. Both speech and PMA data were simultaneously recorded for two non-impaired subjects and, then, the SSI system was used to predict the speech acoustics from the captured articulatory data. A qualitative comparison between the original and predicted speech signals along with some preliminary results on the intelligibility of speech produced by the SSI are reported in this work.

2. Methods

2.1. Articulator motion capture

To capture the movement of the vocal tract during speech articulation, a magnetic sensing technique known as Permanent Magnet Articulography (PMA) [8-11] is used in this work. As illustrated in Fig. 1b, in the current PMA setup a total of six cylindrical Neodymium Iron Boron (NFeB) permanent magnets are attached to the articulators whose movement want to be monitored: four are attached to the lips (ø1mm×5mm), one

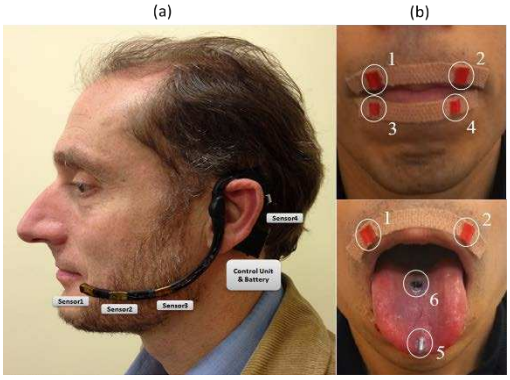

Figure 1. External PMA device. (a) Wearable PMA headset with control unit, battery and 4 tri-axial magnetic sensors. (b) Placement of the magnets.

to the tongue tip (ø2mm×4mm) and one to the tongue blade (ø5mm×1mm). These magnets are currently attached using Histoacryl surgical tissue adhesive (Braun, Melsungen, Germany) during the experimental trials, but will be surgically implanted for long term usage. As shown in Fig. 1a, the rest of the PMA system comprises four tri-axial Anisotropic Magnetoresistive (AMR) magnetic sensors mounted on the wearable headset, which capture the magnetic field generated by the magnets during articulation, a control unit, a rechargeable battery and a processing unit (e.g. computer/tablet PC). Compared to other techniques for the capture of articulator movement, such as electromagnetic articulography (EMA) [13], surface electromyography (sEMG) [15] or electropalatography (EPG) [16], the PMA system has the potential advantage of being unobtrusive, since there are no wires coming out of the mouth or electrodes attached to the skin, which may cause unwanted attention in public. Moreover, as shown in Fig. 1a, the PMA system is also relatively lightweight and highly portable.

The PMA device in Fig. 1a is the result of an iterative engineering process. Earlier PMA-based prototypes [9] demonstrated acceptable speech reconstruction performance, but were less dissatisfactory in terms of their appearances, comfort and ergonomic factors for the users. To address these challenges, the current prototype in Fig. 1a was developed accordingly to the feedback from user questionnaires and through discussion with stakeholders including clinicians, potentials users and their families [10]. As a result, the appearance and comfort of the device was extensively improved without compromising the speech performances to its predecessors.

Despite the improvements made on the external PMA prototype in Fig. 1, it is not without drawbacks: 1) issue with stability under exaggerated movement, 2) uncomfortable over a long period of time and 3) undesirable appearance for some users. To alleviate these limitations, an intraoral version of the PMA prototype was developed in [11] that fits under the palate inside the user's mouth in a form of a dental retainer, as shown in Fig. 2. Although the operational of the device remained similar to the external version, the intraoral circuitry has drastically reduced in size. Moreover, due to the proximity of the sensors to the magnetic markers, smaller magnets are needed. Since the denture retainer is completely hidden inside the user's oral cavity, thus eliminating any unwanted public attention. Previous studies suggested that the appearance is one of the most critical factors that affect the acceptability of any AT by their potential users [14,15].

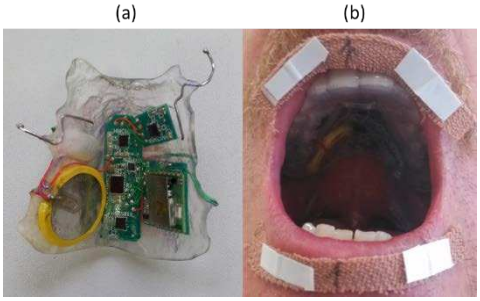

Figure 2. Intraoral version of the PMA capturing device. (a) Intraoral PMA device is prototyped in a form of a denture retainer. (b) View of the device when worn by the user.

2.2. Speech synthesis procedure

Fig. 3 shows a diagram of the procedure used to synthesise speech from captured articulator movement. As can be seen, the procedure consists of two phases: training and conversion. The aim of the training phase is to obtain a statistical model (an artificial neural network in this case) for mapping sensor data into acoustics. The parameters of this model are learned from a set of synchronous recordings with PMA and speech signals made by the person before the laryngectomy (around 30 minutes of those recordings are required in the current system). To facilitate automatic learning, the artificial neural network is trained using a set of parameters (features) extracted from the speech and PMA signals rather than with the raw signals. The speech signals are parameterised to 32-dimensional feature vectors extracted every 25 ms using the STRAIGHT vocoder [19]: 25 of those parameters are used to represent the vocal tract filter as Mel-frequency cepstral coefficients (MFCCs) [20] and the 7 remaining parameters represent the source signal by aperiodicity values in 5 bands and a fundamental frequency (F0) with explicit voicing decision. For the PMA signals, features are extracted by applying the principal component analysis (PCA) technique for dimensionality reduction over short windows spanning 25 ms of sensor samples. Finally, both the PMA and speech features are normalised to have zero mean and unit variance.

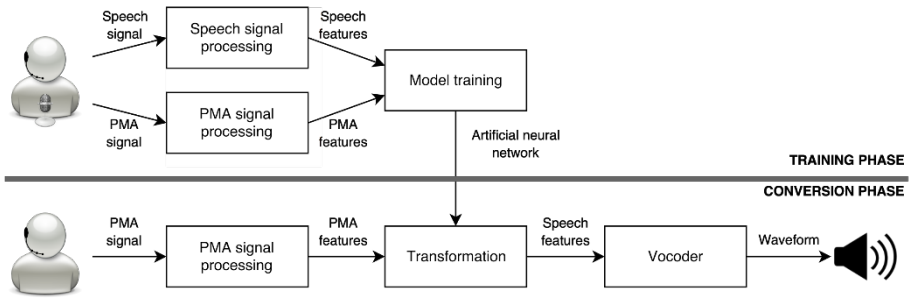

Figure 3. Flow diagram of the training and conversion stages of the speech synthesis procedure.

The artificial neural network obtained at the end of the training phase is then employed to restore the person's speech following laryngectomy. This is what is shown in the conversion phase of Fig. 3. Thus, the neural network is used to map the features computed from the sensor data into a sequence of acoustic parameters (speech features), from which a waveform is finally synthesised and then played back to the user via loudspeakers. The STRAIGHT vocoder is used again to synthesise a time-domain signal from the sequence of speech features predicted by the neural network. Provided that the latency of the conversion process is less than 50 ms (i.e. the delay between an articulatory gesture and the acoustics generated by the system), it will be possible to restore the auditory feedback without inducing mental stress or causing disfluencies to the subject [21]. There is also the possibility that real-time auditory feedback might enable the user to learn to produce better speech (like learning to play an instrument).

Considerable effort was spent on investigating the best machine learning technique for modelling the PMA-to-acoustic mapping. As a result, it was found that recurrent neural networks (RNNs) [14,22], a type of artificial neural especially suited for modelling sequential data, provide a good compromise between speech reconstruction performance and conversion latency. A RNN consists of a set of recurrently connected blocks, each one implementing a nonlinear mapping from the inputs to the outputs. During learning, the RNN parameters are iteratively optimized to minimize the error between the speech features computed from the original speech signals and the features predicted by the network from the sensor data. The RNN employed in this work has four hidden layers with 164 gated recurrent units (GRUs) [23] each. The RNN parameters are randomly initialised and optimized using the stochastic gradient descent technique with mini-batches of 50 sentences. Training is run for 100 epochs or until the error computed over a validation set start increasing.

2.3. Parallel articulatory-speech database

For this preliminary study, data was recorded by two native British-English male subjects (S1 and S2) with normal speaking ability. As the aim of this study is to demonstrate the feasibility of voice reconstruction from articulator movement, we only focus on non-impaired people in this work. Only one of the subject S1 was familiar with the PMA device and had used it prior to this study. Each subject recorded a random subset of the CMU Arctic corpus of phonetically-rich sentences [24]. This corpus was selected because it is widely used in speech synthesis research and it allows us to evaluate the full phonetic range. The total amount of data recorded by the subjects was: 470 sentences (28 minutes) by S1 and 509 sentences (26 minutes) by S2. Each recording session lasted approximately 75 minutes, including the time to fit the magnets and PMA device to the subject and the actual recording time. The recordings were conducted in an acoustically isolated room. During recording, the subject was asked to read aloud a random subset of sentences from the CMU Arctic corpus. A visual prompt of each sentence was presented to the participant at regular intervals of 10 s. PMA and audio signals were recorded simultaneously at sampling frequencies of 100 Hz and 16 kHz, respectively. The audio was recorded using a shock-mounted AKG C1000S condenser microphone via a dedicated stereo Lexicon Lambda USB-sound card. Articulatory data, was recorded using the PMA device shown in Fig. 1.

3. Results

As a qualitative evaluation of the speech quality achieved by the speech restoration system, Fig. 4 compares speech signals recorded by the subjects (Original) and the corresponding ones predicted from sensor data (SSI) for both subjects S1 and S2. The comparison is made at three levels: at the waveform level (1st row), between the spectrograms of the signals (2nd row) and, finally, between the F0 contours (i.e. evolution of the fundamental frequency across time) of the signals.

Secondly, a listening test was conducted to evaluate the intelligibility of speech generated by the SSI. In the test, listeners were asked to transcribe a random subset of 12 sentences chosen from the ones available for subjects S1 and S2 (6 sentences for each subject). A total of 21 subjects participated in the test. Listeners were allowed to replay the speech stimuli as many times as they wanted. Table 1 shows the results of the listening test. Two intelligibility measures are reported: the percentage of words correctly identified by the listeners (word correct) and the word accuracy (i.e. ratio of words correctly identified after discounting the insertion errors). For each measure, the 95% bootstrapped confidence intervals are also presented.

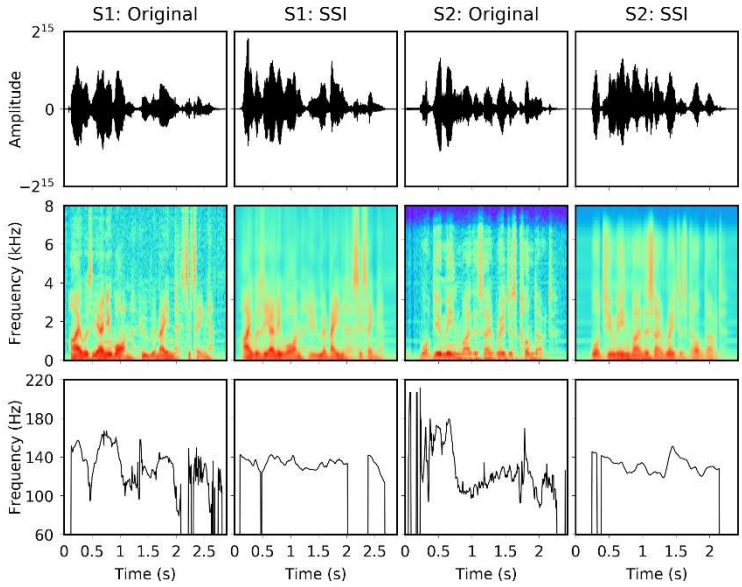

Figure 4. Examples of speech waveforms (1st row), spectrograms (2nd row) and F0 contours (3rd row) for the sentences 'I was the only one who remained sitting' and 'What do you mean by this outrageous conduct?' spoken by the subjects S1 and S2. Both original speech signals and signals predicted by the SSI are shown.

Table 1. . Speech intelligibility results for the proposed silent speech system.

Subject	Word correct (%)	Word accuracy (%)
S1	65.97±8.79	64.80±8.91
S2	65.56±8.25	63.44±8.40

4. Discussion

From the signal examples shown in Fig. 4, it can be seen that the SSI is able accurately to reproduce the speech signals originally uttered by both subjects. In particular, the speech formants in both cases are well predicted and their trajectories are sharp and stable. Other detailed characteristics of speech, however, are not accurately modelled by the current PMA device, and that is the reason that the spectrograms of the predicted signals appear smoothed compared to the originals.

It is remarkable that the SSI is able to predict F0 contours that seem natural and relatively similar to the original although PMA only provides information about the upper part of the vocal tract and very little information about voicing [25,26]. It could be that the system is learning some latent correlations between the movement of the articulators and the excitation parameters. Also, because the SSI is adapted to each particular subject, the RNN models can learn the statistics of the fundamental frequency for that subject (i.e. range, average F0 value for that subject, etc.). The problem of estimating a good excitation signal becomes especially relevant in laryngectomy patients, who no longer have vocal folds.

Finally, regarding the results in Table 1, there are several reasons why only ~65% intelligibility was obtained for both subjects. First, it is well-known that the CMU Arctic sentences are difficult material that was not written to be spoken and contains unusual words that are not in common usage. Second, the participants of the listening test did not have access to any visual clues (e.g. movement of the lips) which are of considerable help in following a speaker. These clues, however, will be normally available when the SSI is used by laryngectomees. Third, the PMA device used in this study was designed on the basis of an average head size for an adult. In this sense, a subject- tailored design is expected to improve the quality of the captured articulatory data.

5. Conclusions

In comparison to other silent speech techniques, our sensor technology is unobtrusive and we can produce speech which resembles the subject's own voice. We are about to enter a clinical trial. In this, our challenge is how to obtain the parallel sensor/speech data required to train the transformation. In many cases it will not be possible to obtain this data prior to the laryngectomy, but we may be able to have the subject mime to audio recordings once the implants are in place.

Acknowledgements

This is a summary of independent research funded by the National Institute for Health Research (NIHR)'s Invention for Innovation Programme (Grant Reference Number II-LB-0814-20007). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

References

[1] Jones, T. M., De, M., Foran, B., Harrington, K., & Mortimore, S. (2016). Laryngeal cancer: United Kingdom National Multidisciplinary guidelines. J Laryngol Otol, 130(S2), S75-S82.

[2] Cancer Research UK (2015), Laryngeal cancer statistics. Retrieved February 21, 2017, from http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/laryngeal-cancer#heading-Two

[3] Ferlay, J., Shin, H. R., Bray, F., Forman, D., Mathers, C., & Parkin, D. M. (2010). Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. Int. J. Cancer., 127(12), 2893-2917.

[4] Byrne, A., Walsh, M., Farrelly, M., & O'Driscoll, K. (1993). Depression following laryngectomy. A pilot study. Br J Psychiatry, 163(2), 173-176.

[5] Braz, D. S. A., Ribas, M. M., Dedivitis, R. A., Nishimoto, I. N., & Barros, A. P. B. (2005). Quality of life and depression in patients undergoing total and partial laryngectomy. Clinics, 60(2), 135-142.

[6] Danker, H., Wollbrück, D., Singer, S., Fuchs, M., Brähler, E., & Meyer, A. (2010). Social withdrawal after laryngectomy. Eur Arch Otorhinolaryngol, 267(4), 593-600.

[7] Denby, B., Schultz, T., Honda, K., Hueber, T., Gilbert, J. M., & Brumberg, J. S. (2010). Silent speech interfaces. Speech Commun, 52(4), 270-287.

[8] Fagan, M.J., Ell, S.R., Gilbert, J.M., Sarrazin, E., & Chapman, P.M. (2008). Development of a (silent) speech recognition system for patients following laryngectomy. Med Eng Phys, 30(4), 419-425.

[9] Hofe, R., Ell, S. R., Fagan, M. J., Gilbert, J. M., Green, P. D., Moore, R. K., & Rybchenko, S. I. (2013). Small-vocabulary speech recognition using a silent speech interface based on magnetic sensing. Speech Commun, 55(1), 22-32.

[10] Cheah, L. A., Bai, J., Gonzalez, J. A., Ell, S. R., Gilbert, J. M., Moore, R. K., & Green, P. D. (2015). A User-centric design of permanent magnetic articulography based assistive speech technology. In BIOSIGNALS (pp. 109-116).

[11] Cheah, L. A., Bai, J., Gonzalez, J. A., Gilbert, J. M., Ell, S. R., Green, P. D., & Moore, R. K. (2016). Preliminary evaluation of a silent speech interface based on intra-oral magnetic sensing. In BIODEVICES (pp. 108–116).

[12] Bishop, C. M. (2006). Pattern recognition. Machine Learning, 128, 1-58.

[13] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

[14] Schönle, P. W., Gräbe, K., Wenig, P., Höhne, J., Schrader, J., & Conrad, B. (1987). Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract. Brain Lang, 31(1), 26-35.

[15] Schultz, T., & Wand, M. (2010). Modeling coarticulation in EMG-based continuous speech recognition. Speech Commun, 52(4), 341-353.

[16] Hardcastle, W. J., Gibbon, F. E., & Jones, W. (1991). Visual display of tongue-palate contact: Electropalatography in the assessment and remediation of speech disorders. Brit J Disorder Comm, 26(1), 41-74.

[17] Hirsch, T., Forlizzi, J., Hyder, E., Goetz, J., Kurtz, C., & Stroback, J. (2000, November). The ELDer project: social, emotional, and environmental factors in the design of eldercare technologies. In Proceedings on the 2000 conference on Universal Usability (pp. 72-79). ACM.

[18] Martin, J. L., Murphy, E., Crowe, J. A., & Norris, B. J. (2006). Capturing user requirements in medical device development: the role of ergonomics. Physiol Meas, 27(8), R49.

[19] Kawahara, H., Masuda-Katsuse, I., & De Cheveigne, A. (1999). Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. Speech Commun, 27(3), 187-207.

[20] Fukada, T., Tokuda, K., Kobayashi, T., & Imai, S. (1992). An adaptive algorithm for mel-cepstral analysis of speech. In ICASSP (pp. 137-140).

[21] Yates, A. J. (1963). Delayed auditory feedback. Psychological Bulletin, 60(3), 213-232.

[22] Graves, A., Mohamed, A. R., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In ICASSP (pp. 6645-6649).

[23] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. In Proc. Conference on Empirical Methods in Natural Language Processing (pp. 1724–1734).

[24] Kominek, J., Black, A. W., & Ver, V. (2003). CMU ARCTIC databases for speech synthesis.

[25] Gonzalez, J. A., Cheah, L. A., Bai, J., Ell, S. R., Gilbert, J. M., Moore, R. K., & Green, P. D. (2014). Analysis of phonetic similarity in a silent speech interface based on permanent magnetic articulography. In InterSpeech (pp. 1018-1022).

[26] Gonzalez, J. A., Cheah, L. A., Gilbert, J. M., Bai, J., Ell, S. R., Green, P. D., & Moore, R. K. (2016). A silent speech system based on permanent magnet articulography and direct synthesis. Comput Speech Lang, 39, 67-87.