A Real-Time Silent Speech System for Voice Restoration after Total Laryngectomy

Jose. A. Gonzalez [a1] and Phil D. Green [a]

[a] Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello St, Sheffield S1 4DP, UK

Email: j.gonzalez@uma.es; p.green@sheffield.ac.uk

Corresponding author: Jose A. Gonzalez (j.gonzalez@uma.es), Departamento de Lenguajes y Ciencias de la Computación, Universidad de Málaga, Complejo Tecnológico, Campus de Teatinos, 29071 Málaga, España.

Funding

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.

Declaration of interest

The authors are responsible for the reported research, and have participated in the concept and design, analysis and interpretation of data, drafting or revising of the manuscript, and have approved the manuscript as submitted. The authors have no conflict of interests that might be interpreted as influencing the research.

Acknowledgements

We thank Prof James M. Gilbert and Dr Lam A. Cheah for their invaluable help with the PMA data capture process and designing and developing the PMA system.

[1] **Current address**: Departamento de Lenguajes y Ciencias de la Computación, Universidad de Málaga, Complejo Tecnológico, Campus de Teatinos, 29071 Málaga, España.

A Real-Time Silent Speech System for Voice Restoration after Total Laryngectomy

Resumen

Antecedentes y objetivo: Aquellas personas que han perdido su voz después de una laringectomía se ven limitadas irremediablemente en su comunicación diaria. A pesar de existir en la actualidad métodos para recuperar el habla tras la laringectomía, todos ellos presentan limitaciones. El objetivo de este trabajo es explorar un método alternativo para hablar tras la laringectomía en el que el movimiento de los órganos de la voz se transforma en una señal acústica utilizando técnicas de aprendizaje automático.

Materiales y métodos: En esta investigación participaron 6 adultos sanos. Para cada sujeto se grabaron simultaneamente tanto su voz como los movimientos de sus labios y lengua. Los movimientos de los órganos del habla fueron capturados usando una técnica conocida como Articulografía de Imán Permanente (PMA), en la cual pequeños imanes se colocan sobre estos órganos y el campo magnético generado por los imanes se captura usando unos sensores sensibles al campo magnético. Se utilizaron redes neuronales artificiales profundas para modelar la transformación entre los datos de los sensores y la acústica de la voz.

Resultados: El sistema de habla silenciosa propuesto es capaz de generar voz que suena natural, se asemeja a la propia voz del sujeto y es inteligible (hasta un 92% de inteligibilidad para algunos sujetos).

Conclusiones: El sistema propuesto podría ser en el futuro una opción viable para restaurar la voz tras una laringectomía total.

Palabras clave: Laringectomía; interfaces del habla silenciosa; rehabilitación de la voz; síntesis de voz; articulografía por imanes permanentes.

Abstract

Background and aim: Individuals who have lost their voice following a laryngectomy as a treatment for cancer will inevitably struggle with their daily communication. Unfortunately, current existing methods for speaking after laryngectomy all have limitations, either because of the poor acoustics generated by these methods or because they can be potentially harmful. The aim of this work is thus to explore an alternative method for post-laryngectomy voice restoration in which the movement of the intact articulators is captured and then converted into audible speech using machine learning techniques.

Materials and methods: To demonstrate the feasibility of speech generation from captured articulator movement, 6 healthy adults were recruited. For each subject, both the speech acoustics and the subject's articulator movements were recorded simultaneously. Articulator movements were captured using a technique known as Permanent Magnet Articulography (PMA), in which small magnets are attached to the articulators (typically tongue and lips) and the magnetic field generated by the magnets is captured with sensors located close to the mouth. Deep artificial neural networks were then used to model the mapping between the sensor data and the speech acoustics, thus, enabling the synthesis of speech from captured articulatory data.

Results: The proposed silent speech system is able to generate speech that sounds natural, resembles the subject's own voice and is fairly intelligible (up to 92% intelligibility for some speakers on a phonetically-rich corpus).

Conclusions: With further research, the proposed system could be in future a real option to restore lost voice after laryngectomy.

Keywords: Laryngectomy; silent speech interfaces; speech rehabilitation; speech synthesis; permanent magnet articulography.

Introduction

Individuals who undergo total laryngectomy to treat throat cancer often find themselves struggling with oral communication after losing their voice after the laryngectomy. Although larynx cancer only accounts for ~1% of all cancers ('Laryngeal cancer statistics', 2015), according to a recent study (Ferlay Jacques et al., 2010) it has a high survival rate: around 70% of the patients live 5 years or more after the laryngectomy. Worldwide, this accounts for more than 425,000 individuals (Jones, De, Foran, Harrington, & Mortimore, 2016). After losing their voice, many laryngectomies report feelings of loss of identity and clinical depression as well as social isolation (Braz, Ribas, Dedivitis, Nishimoto, & Barros, 2005; Byrne, Walsh, Farrelly, & O'Driscoll, 1993; Danker et al., 2010).

To speak again, laryngectomees can use any of the following three voice restoration methods (Jassar, England, & Stafford, 1999): valved speech, oesophageal speech and the electrolarynx. Valved speech is the most popular and preferred method. In this method, a one-way valve allowing air from the lungs to pass into the oesophagus without food and liquids passing into the trachea is inserted into the tissue separating the trachea and oesophagus. Although valved speech provides the most natural voice among the three methods, it has to be replaced frequently and its voice sounds masculine, thus disliked by many females. Oesophageal speech is a type of alaryngeal speech which does not require any instrumentation. To speak in this way, you move air down to the upper oesophagus and then release it in a controlled manner making the oesophagus to vibrate. Oesophageal speech, however, is difficult to learn and has a low speaking rate. The third method, the electrolarynx, is a handheld vibrating machine that produces sound for you to create a voice. To speak using the electrolarynx, it has to be hold against the neck, then the device is activated and it will inject sound into the vocal tract apparatus for you to form words. The electrolarynx is relatively cheap and easy to use, but requires manual dexterity and produces a robotic voice.

Recently, the use of silent speech interfaces (SSIs) (Denby et al., 2010) has gained increased interest as a real alternative to restore speech communication. SSIs are devices enabling oral communication 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) by exploiting other non-audible biosignals generated during speech production. Examples of these biosignals are the electrical activity in the brain or the movement of the speech articulators. From these signals, a SSI recover the speech the user wished to produce, either by performing automatic speech recognition on the biosignals or by directly converting them into speech, as we investigate in this work.

The present works describes a SSI system which can be used for speech restoration after laryngectomy. Our system is able to generate audible speech from captured movement of the speech articulators. In particular, the movements of the lips and tongue of the patient is captured using a technique known as Permanent Magnet Articulography (PMA) (Cheah et al., 2015, 2016; Fagan, Ell, Gilbert, Sarrazin, & Chapman, 2008; Gilbert et al., 2010; Hofe et al., 2013), in which small magnets are attached to the articulators and the variations of the magnetic field generated by those magnets when the person articulates words are captured by sensors located close to the mouth. To transform PMA data into speech, we use deep learning techniques (Goodfellow, Bengio, & Courville, 2016; LeCun, Bengio, &

Hinton, 2015) trained with simultaneous recordings of articulatory and speech data made by the patient before she/he loses her/his voice. The proposed system is implemented as an *app* running on a smartphone. The *app* can generate speech in real time and, because the deep learning techniques are trained with recordings of the person's own voice, it generates speech resembling the person's original voice.

To evaluate the feasibility of the proposed voice restoration method, preliminary results are reported here for healthy subjects. Speech generated by our system from the captured PMA data is evaluated in term of naturalness, quality and intelligibility.

Materials and methods

Participants and database

Although the aim of our research is to help laryngectomees to recover their voice, for the present study our aim is to determine whether speech of sufficient quality can be synthesised from articulator movement. Thus, 6 healthy British subjects were recruited: 4 men (M1 to M4) and 2 women (F1 and F2). None of the subjects had any record of speech and language disorders. For each subject, articulatory data was recorded in synchrony with her/his speech for a random subset of the CMU Arctic corpus of phonetically-rich sentences (Kominek & Black, 2004). The amount of data recorded by the subjects is shown in Table 1.

Data was recorded in a sound-proof room to optimize the audio quality. During the recordings, the subjects were asked to read aloud a random subset of sentences included in the CMU Arctic corpus. A visual prompt of each sentence was presented to the subject at regular intervals of 10 s. PMA and audio signals were recorded simultaneously at sampling frequencies of 100 Hz and 48 kHz, respectively. The audio was recorded using a AKG C1000S condenser microphone located about 20 cm from the subject, whereas the articulatory data was captured using the in-house PMA device described in the following section. Later, the audio signals were digitally downsampled to 16 kHz.

Articulator motion capture

Permanent Magnet Articulography (PMA) (Cheah et al., 2015, 2016; Fagan et al., 2008; Gilbert et al., 2010; Hofe et al., 2013) is used in this work to capture the movement of the speech organs. As shown in Figure 1b, six Neodymium Iron Born (NFeB) permanent magnets are used in the current setup: four of them are attached to the lips, one to the tongue tip and one to the tongue blade. During the experiments the magnets are temporarily attached to the articulators using tissue glue, which last 1-2 hours approximately, but they will be implanted for long-term usage. The magnetic field generated by the magnets is then captured by four tri-axial magnetic sensors mounted on the wearable headset shown in Figure 1a. Each sensor captures the 3-dimensional spatial components of the magnetic field at the sensor location. The headset also includes a rechargeable battery and a control unit with a Bluetooth transmitter, which sends the articulatory data to a processing unit (e.g. PC or smartphone) for further processing.

In contrast to other methods for articulator motion capture, such as electromagnetic articulography (EMA) (Schönle et al., 1987) or electropalatography (EPG) (Hardcastle, Gibbon, & Jones, 2011), the exact position of the speech organs during speech production cannot be easily derived from PMA data. In other words, it is not currently possible to determine the positions of the magnets from the captured PMA data. Rather, pattern recognition techniques are used in the current work to map the articulatory data into acoustic speech parameters from which an acoustic signal can be synthesised. As an advantage, the PMA system is potentially unobtrusive, as there are no wires coming out of the mouth (as in EMA or EPG) or electrodes attached to the skin (as in surface electromyography (Schultz & Wand, 2010)).

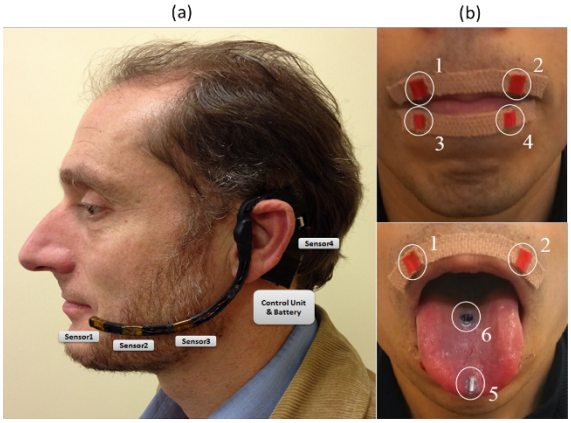

Figure 1. Permanent Magnet Articulography (PMA) capturing device. (a) Wearable PMA headset with control unit, battery and 4 tri-axial magnetic sensors. (b) Placement of six magnets on lips (pellets 1-4), tongue tip (pellet 5) and tongue blade (pellet 6).

Speech generation procedure

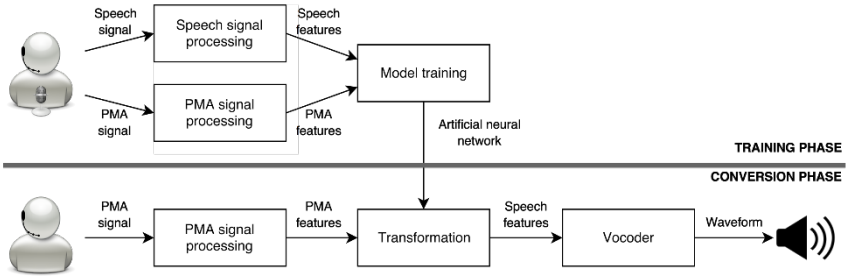

Figure 2. Block diagram of the training and conversion stages of the speech synthesis procedure. In the training phase, the parameters of an artificial neural network, which maps PMA-derived feature vectors into speech feature vectors, are learned. In the conversion phase, the neural network is employed to transform the captured articulatory data into speech.

A block diagram of the speech synthesis procedure used to transform PMA data into speech is depicted in Figure 2. The procedure consists of two distinct stages: training and conversion. The aim of the training phase is to model the relationship between the articulatory data and the acoustics for each subject. Deep neural networks (LeCun et al., 2015), which are a class of machine learning techniques that have been shown to provide unparalleled performance in a number of real-word tasks, are used to

this end. The neural networks are trained with a set of synchronous recordings with PMA and speech signals made by the same person before the laryngectomy. This allows the neural network to learn the mapping between the PMA data and the corresponding acoustics. Rather than training the network with the raw signals, as illustrated in Figure 2, the PMA and speech signals are independently represented as sequences of feature vectors which contain less redundancy than the raw signals. The speech signals are parametrised as sequences of 32-dimensional feature vectors extracted every 5 ms from 25 ms-long analysis windows using the STRAIGHT vocoder (Kawahara, Masuda-Katsuse, & de Cheveigné, 1999): 25 parameters are employed to represent the vocal tract filter as Mel-frequency cepstral coefficients (MFCCs) (Fukada, Tokuda, Kobayashi, & Imai, 1992) and the 7 remaining parameters represent the glottal signal as 5-band aperiodicity values, fundamental frequency (F_0) and unvoiced/voiced decision. Similarly, the PMA signals as represented as sequences of features vectors extracted at the same frame rate as for the speech signals. These feature vectors are computed by applying the principal component analysis (PCA) technique for dimensionality reduction over the 25 ms signal windows. After the feature extraction procedure, the PMA and speech derived features are standardised.

Several neural network architectures were investigated for modelling the PMA-to-speech mapping (Gonzalez, Cheah, Gomez, et al., 2017; Gonzalez, Cheah, Green, et al., 2017). From this investigation it was clear that recurrent neural networks (RNNs) (Cho et al., 2014; Graves, Mohamed, & Hinton, 2013; LeCun et al., 2015), a class of neural networks especially suited for modelling sequential data, provided the best speech reconstruction performance among all the architectures. A RNN consists of a set of recurrently connected processing blocks which allow not only for the current articulatory data, but also for historic data, to be taken into account in the PMA-to-speech mapping. Because RNNs can use the historic data, it can model long-term correlations which are beneficial when modelling different aspects of speech, such as the prosody, etc.

As shown in Figure 2, the neural network obtained in the training phase is used in the conversion phase to transform PMA feature vectors into speech feature vectors. Finally, the STRAIGHT vocoder is used again to synthesise a waveform from the estimated speech feature vectors.

In summary, the proposed SSI system is able to restore the person's speech following laryngectomy. Provided that the latency of the conversion process is small enough, it will be possible to restore the normal auditory feedback to the subject (Yates, 1963).

Real-time speech synthesis

To evaluate our system for speech restoration, the conversion stage in Figure 2 was implemented as a Mobile app running on an Android smartphone. The communication between the PMA headset in Figure 1a and the *app* is done via Bluetooth. All the remaining processing steps are implemented on the *app*. In particular, the *app* is responsible for transforming the articulatory data into speech feature vectors, synthesising speech from those vectors and, finally, playing the speech waveform back to the user via loudspeakers integrated in the smartphone.

Timings of the time incurred by the processing steps indicate that the app is able to generate speech in real-time. In other words, the delay between an articulatory gesture and the corresponding acoustic feedback generated by the app is always less than 50 ms. As indicated in (Yates, 1963), this value can be considered as the upper for the maximum acceptable delay without causing disfluencies or stress to the subject.

Results

To evaluate the quality of the speech generated by our SSI, we compared the speech signals obtained from PMA data with the original signals recorded by the subjects. Then, speech quality was objectively measured by using the Mel-cepstral distortion measure (MCD) (Kubichek, 1993), which is a well-known spectrographic distortion metric. This allows us to compute the error of the synthesised speech signals with respect to the original ones recorded by the subjects. In MCD, higher values indicate a higher distortion, whereas a perfect reconstruction would yield a value of 0 dB in terms of this metric.

Figure 3 shows the average MCD results over all subjects for different phone categories. To obtain these results, the original speech signals were first segmented into phones by force-aligning their word-level transcriptions using automatic speech recognition. The phone-level transcriptions were then used to segment the original and synthesised speech signals. Next, the MCD metric was independently computed for each phone. When considering the manner of articulation, it can be seen that the vowels and fricatives are the sounds that are more accurately synthesised, while plosive, affricate and approximant consonants are less well reconstructed due to their more complex articulation and dynamics. For the place of articulation, the phones articulated at the middle and back of the mouth (i.e. palatal, velar and glottal consonants) are systematically more poorly reconstructed. This is due to those areas of the vocal tract not being well captured by the current PMA device (Gonzalez et al., 2014; Gonzalez, Cheah, Green, et al., 2017).

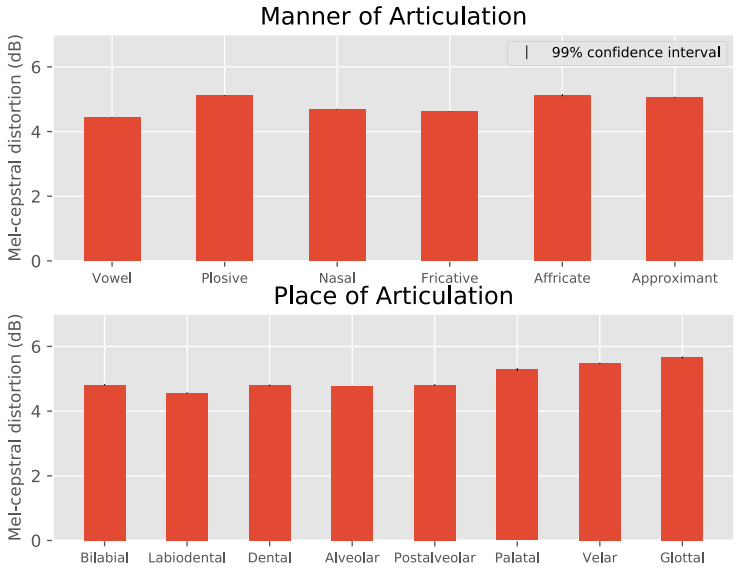

Figure 3. Objective distortion results for the speech generated by the proposed SSI for different phonetic categories. Speech distortion is measured using the Mel-cepstral distortion metric.

Next, we conducted a listening test to evaluate the intelligibility of the speech generated by the SSI. Twenty one raters participated in the listening test. Each rater was asked to manually transcribe 9 speech samples from those generated by our system. The raters were allowed to replay the speech samples as many times as they wanted. Table 2 presents the mean intelligibility results over all subjects in our database (Average) and the results for the most intelligible subject, M4 (Best). Two intelligibility measures are reported: the percentage of words correctly identified by the raters (word correct) and the word accuracy, which corresponds to the ratio of words correctly identified after discounting the insertion errors.

On average, around 75% intelligibility is obtained when considering all subjects in our study, but this value yields 92% for the subject M4.

Discussion

In this work we have described an speech synthesiser which can be controlled in real time using motion data captured from the user's speech organs. To synthesise speech from articulatory data, recurrent neural networks are trained with simultaneous recordings of articulatory and speech data from the same subject, thus effectively enabling the system to transform captured articulatory data into audible speech. To evaluate our system, experiments were conducted in which the system was used to generate speech from articulatory data recorded by several healthy subjects.

We first assessed the objective quality of the speech generated by the system. To this end, the speech signals generated by our system were represented as a series of Mel-frequency cepstral coefficients (MFFCs) and compared with the MFCCs computed from the original speech signals recorded by the subjects using the MCD distortion metric. Broadly speaking, when considering the manner of articulation of the phones, plosive and affricative consonants were consistently moor poorly reconstructed, which can be due to their more complex behaviour, while the vowels were, on average, more accurately reconstructed. For the place of articulation categories, we found that the phones articulated at the middle and back of the mouth were more poorly estimated. This coincides with similar findings from other studies using Permanent Magnet Articulagraphy (Gonzalez, Cheah, Gomez, et al., 2017; Gonzalez et al., 2014, 2016). In those investigations, it was empirically shown that the PMA system has limitations in capturing the movements of the back of the mouth since no magnet is attached to the velum or the glottis. Moreover, PMA provides very little information about the voicing because, similarly, no magnet is attached to the vocal folds.

Next, we evaluated the intelligibility of the speech generated by our system. Despite the limitations of the current PMA system for detecting some speech sounds, we found that the speech generated by our system was fairly intelligible: 75% intelligible (i.e. 75 out of 100 words are correctly identified) on average for all subjects in our database, but yielding 92% intelligible for a particular subject. It is worth pointing out that, in addition to the accuracy of the articulatory-to-speech mapping, for a normal conversation other factors come into play. For instance, contextual information is known to be beneficial to

disambiguate possible word meaning confusions. Visual clues (e.g. lip reading), which the participants in our listening test did not have access to, are also of considerable help in following a conversation.

Our results indicate the potential of the technology described in this work to be used for restoring the voice to laryngectomees. Before this could be achieved, several problems have to be addressed in future investigations. First, the magnets would need to be implanted for long-term usage as the current medical glue used to attach the magnets last 1-2 hours approximately. Second, the current system relies on the availability on parallel articulatory and speech data for training the deep models, but it may not be possible to obtain such a data always (e.g. if the patient has already lost the voice). In such a case, it might be possible to obtain the parallel data by asking the patient to mime to another voice while wearing the magnets implanted. Finally, improving the estimation of the speech prosody would also need to be addressed in future work.

Conclusions

In this work we have described a system which generates speech acoustic from lips and tongue movement. Our focus is to help laryngectomy patients to speak again. In comparison with other methods to restore speech, our technology is unobstrusive and produces speech which sounds as the person's own voice. Results for healthy subjects demonstrate the feasibility of this approach. In future work we will investigate the application of our system to real laryngectomy patients.

References

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 (Sao Paulo, Brazil)*, *60*(2), 135–142. https://doi.org//S1807-59322005000200010

Byrne, A., Walsh, M., Farrelly, M., & O'Driscoll, K. (1993). Depression following laryngectomy. A pilot study. *The British Journal of Psychiatry: The Journal of Mental Science*, *163*, 173–176.

Cheah, L. A., Bai, J., Gonzalez, J. A., Ell, S. R., Gilbert, J., Moore, R., & Green, P. (2015). A User-centric Design of Permanent Magnetic Articulography based Assistive Speech Technology. In *Biosignals* (pp. 109–116). Lisbon, Portugal. https://doi.org/10.5220/0005354601090116

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). Rome, Italy. https://doi.org/10.5220/0005824501080116

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1724–1734). Doha, Qatar: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/D14-1179

Danker, H., Wollbrück, D., Singer, S., Fuchs, M., Brähler, E., & Meyer, A. (2010). Social withdrawal after laryngectomy. *European Archives of Oto-Rhino-Laryngology: Official Journal of the European Federation of Oto-Rhino-Laryngological Societies (EUFOS): Affiliated with the German Society for*

Oto-Rhino-Laryngology - Head and Neck Surgery, *267*(4), 593–600. https://doi.org/10.1007/s00405-009-1087-4

Denby, B., Schultz, T., Honda, K., Hueber, T., Gilbert, J. M., & Brumberg, J. S. (2010). Silent speech interfaces. *Speech Communication*, *52*(4), 270–287. https://doi.org/10.1016/j.specom.2009.08.002

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. *Medical Engineering & Physics*, *30*(4), 419–425. https://doi.org/10.1016/j.medengphy.2007.05.003

Ferlay Jacques, Shin Hai-Rim, Bray Freddie, Forman David, Mathers Colin, & Parkin Donald Maxwell. (2010). Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *International Journal of Cancer*, *127*(12), 2893–2917. https://doi.org/10.1002/ijc.25516

Fukada, T., Tokuda, K., Kobayashi, T., & Imai, S. (1992). An adaptive algorithm for mel-cepstral analysis of speech. In *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing* (Vol. 1, pp. 137–140 vol.1). https://doi.org/10.1109/ICASSP.1992.225953

Gilbert, J. M., Rybchenko, S. I., Hofe, R., Ell, S. R., Fagan, M. J., Moore, R. K., & Green, P. (2010). Isolated word recognition of silent speech using magnetic implants and sensors. *Medical Engineering & Physics*, *32*(10), 1189–1197. https://doi.org/10.1016/j.medengphy.2010.08.011

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. *Computer Speech & Language*, *39*, 67–87. https://doi.org/10.1016/j.csl.2016.02.002

Gonzalez, J. A., Cheah, L. A., Gomez, A. M., Green, P. D., Gilbert, J. M., Ell, S. R., … Holdsworth, E. (2017). Direct Speech Reconstruction From Articulatory Sensor Data by Machine Learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*(12), 2362–2374. https://doi.org/10.1109/TASLP.2017.2757263

Gonzalez, J. A., Cheah, L., Bai, J., Ell, S. R., Gilbert, J., Moore, R., & Green, P. (2014). Analysis of Phonetic Similarity in a Silent Speech Interface based on Permanent Magnetic Articulography. In *Interspeech* (pp. 1018–1022). Singapore.

Gonzalez, J. A., Cheah, L., Green, P., Gilbert, J., R. Ell, S., Moore, R., & Holdsworth, E. (2017). Evaluation of a Silent Speech Interface Based on Magnetic Sensing and Deep Learning for a Phonetically Rich Vocabulary. In *Interspeech* (pp. 3986–3990). Stockholm, Sweden. https://doi.org/10.21437/Interspeech.2017-802

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Graves, A., Mohamed, A. r, & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 6645–6649). https://doi.org/10.1109/ICASSP.2013.6638947

Hardcastle, J. W., Gibbon, F. E., & Jones, W. (2011). Visual display of tongue-palate contact: Electropalatography in the assessment and remediation of speech disorders. *International Journal*

of Language & Communication Disorders, *26*(1), 41–74. https://doi.org/10.3109/13682829109011992

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 Communication*, *55*(1), 22–32. https://doi.org/10.1016/j.specom.2012.02.001

Jassar, P., England, R. J. A., & Stafford, N. D. (1999). Restoration of Voice after Laryngectomy. *Journal of the Royal Society of Medicine*, *92*(6), 299–302. https://doi.org/10.1177/014107689909200608

Jones, T. M., De, M., Foran, B., Harrington, K., & Mortimore, S. (2016). Laryngeal cancer: United Kingdom National Multidisciplinary guidelines. *The Journal of Laryngology and Otology*, *130*(Suppl 2), S75–S82. https://doi.org/10.1017/S0022215116000487

Kawahara, H., Masuda-Katsuse, I., & de Cheveigné, 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 sounds1Speech files available. See http://www.elsevier.nl/locate/specom1. *Speech Communication*, *27*(3), 187–207. https://doi.org/10.1016/S0167-6393(98)00085-5

Kominek, J., & Black, A. W. (2004). The CMU Arctic speech databases. *SSW5-2004*.

Kubichek, R. (1993). Mel-cepstral distance measure for objective speech quality assessment. In *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing* (Vol. 1, pp. 125–128 vol.1). https://doi.org/10.1109/PACRIM.1993.407206

Laryngeal cancer statistics. (2015, May 14). Retrieved 15 April 2018, from http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/laryngeal-cancer

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

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 and Language*, *31*(1), 26–35. https://doi.org/10.1016/0093-934X(87)90058-7

Schultz, T., & Wand, M. (2010). Modeling coarticulation in EMG-based continuous speech recognition. *Speech Communication*, *52*(4), 341–353. https://doi.org/10.1016/j.specom.2009.12.002

Yates, A. J. (1963). Delayed auditory feedback. *Psychological Bulletin*, *60*(3), 213–232. https://doi.org/10.1037/h0044155

Subject	No. of sentences	Minutes of data
F1	353	20
F2	432	22
M1	420	22
M2	470	28
M3	509	26
M4	519	35

Table 1. Amount of data recorded by each subject.

	Word correct (%)	Word accuracy (%)
Average	74.81 ±4.54	73.49±4.66
Best	92.00±3.41	91.53±3.57

Table 2. Results of the listening test for speech intelligibility. Average denotes the average intelligibility results over all subject and best denotes the results obtained for the most intelligible subject (M4). 95% confidence intervals are presented for each measure.