IberSPEECH 2018
21-23 November 2018, Barcelona, Spain

Silent Speech: Restoring the Power of Speech to People whose Larynx has been Removed

Jose A. Gonzalez[1], Phil D. Green[2], Damian Murphy[3], Amelia Gully[3], and James M. Gilbert[4]

[1]University of Malaga, Spain
[2]University of Sheffield, U.K.
[3]University of York, U.K.
[4]University of Hull, U.K.

j.gonzalez@uma.es

Abstract

Every year, some 17,500 people in Europe and North America lose the power of speech after undergoing a laryngectomy, normally as a treatment for throat cancer. Several research groups have recently demonstrated that it is possible to restore speech to these people by using machine learning to learn the transformation from articulator movement to sound. In our project articulator movement is captured by a technique developed by our collaborators at Hull University called Permanent Magnet Articulography (PMA), which senses the changes of magnetic field caused by movements of small magnets attached to the lips and tongue. This solution, however, requires synchronous PMA-and-audio recordings for learning the transformation and, hence, it cannot be applied to people who have already lost their voice. Here we propose to investigate a variant of this technique in which the PMA data are used to drive an articulatory synthesiser, which generates speech acoustics by simulating the airflow through a computational model of the vocal tract. The project goals, participants, current status, and achievements of the project are discussed below.

Index Terms: speech restoration, silent speech interfaces, permanent magnet articulography, articulatory synthesis, magnetic resonance imaging

1. Introduction

A total laryngectomy is a clinical procedure in which the voice box is surgically removed most commonly as a treatment for throat cancer. This procedure not only leaves the subject muted, but it is also known to cause social isolation, feelings of loss of identity and can lead to clinical depression [1, 2, 3]. Current available methods for speaking after a laryngectomy include the electro-larynx, a hand-held device which produces an unnatural, electronic voice; oesophageal speech, which is difficult to master, and the voice prosthesis, which is considered to be the current gold standard, but has a short life time (4 to 8 weeks) due candida growth, thus requiring regular hospital visits for valve replacement [4, 5, 6]. Other available methods such as the Alternative and Augmentative Communication (AAC) devices [7], where the user types words and the device synthesises them, are also limited by their slow manual input and, therefore, are not suitable for any other than short conversations.

As an alternative to existing speech restoration methods, we are investigating a new way to restore speech to those who are unable to speak [8, 9, 10, 11, 12]. The idea is to transform measurements of the lips and tongue movements obtained using magnetic sensing into audible speech using a speaker-dependent transformation, implemented by machine learning techniques (typically deep neural networks [13]). For capturing articulator movement we use Permanent Magnet Articulography (PMA) [14, 15, 16, 17], a technology developed by our collaborators at the University of Hull in which small magnets are attached to the lips and tongue and the magnetic field generated when the articulators move is captured by sensors close to the mouth (see Fig. 1 for a picture of the PMA system). The parameters of the transformation for converting articulator movement into speech are currently estimated from simultaneous recordings of audio and PMA signals acquired before the person loses her/his voice. Some audio samples produced by the proposed restoration method can be found at `https://www.jandresgonzalez.com/is2017`. As can be seen, the samples are mostly intelligible and the speaker identity is clearly preserved.

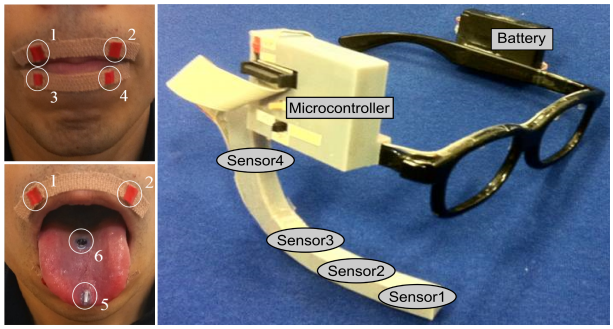

Figure 1: *Permanent Magnet Articulography (PMA) system.* ***Upper-left and lower-left****: placement of magnets used to capture the movements of the lips and tongue.* ***Right****: PMA headset consisting of micro-controller, battery and magnetic sensors for detecting the variations of the magnetic field generated by the magnets.*

A limitation of the above approach for speech restoration is that simultaneous speech-and-sensor recordings are required for estimating the mapping between articulator movement and its acoustics. Thus, this makes this method unsuitable for persons who have already lost their voice. The aim is of this project is, thus, to investigate a novel approach that would make simultaneous recordings unnecessary. The idea is to predict, in real time, the position of the speech articulators from the PMA signals. This is a non-trivial problem as the magnetic field arriving at the sensors is a composite of the fields generated by all the magnets attached to the articulators. From the estimated vocal tract shapes speech can finally be synthesised by simulating airflow propagation through the vocal tract using well-known,

10.21437/IberSPEECH.2018-33

established articulatory synthesis methods [18].

In the next sections, the detailed objectives of the project, the participants, and its current status are described in detail.

2. Project objectives

As previously mentioned, the goal of this project is to investigate and develop a new method for speech restoration based on the PMA capturing technique and machine learning, but without the need of parallel speech and sensor recordings for training the machine learning techniques. We attempt to do this by, instead, learning an alternative transformation which will map the articulatory data captured by the PMA device into a physical model of the vocal tract (e.g. 1D or 2D representation of the vocal tract). Then, we will be able to generate audible speech from the estimated vocal tract shapes by using well-known articulatory synthesis methods.

The detailed objectives of this project are:

- To train a direct transformation from PMA data to vocal tract shapes used by the articulatory synthesiser.
- To personalise the synthesiser so that the speech generated sounds like the users original voice.

With regard to the latter point, we expect to use MRI images of the subject's vocal tract to personalise the synthesiser. In this way, the acoustics generated by the synthesiser will resemble the user's original voice.

3. Partners

There are four partners involved in this project:

- University of Sheffield: Jose A. Gonzalez (principal investigator; now at the University of Malaga), Phil D. Green and Roger K. Moore.
- University of York: Damian Murphy, Helena Daffern and Amelia Gully.
- University of Hull: James M. Gilbert and Lam A. Cheah.
- University of Leeds: Andy Bulpitt and Duane Carey.

4. Current status

Firstly, we have been able to record a database consisting of parallel recordings of MRI, PMA and speech data for 4 subjects. For this pilot study, we decided to record simple material, mainly consonant-vowel (CV) and vowel-consonant-vowel syllables. At the same time, we have been working on improving the quality of speech generated by our articulatory synthesiser. In this regard, we described in [19] a dynamic 3D digital waveguide mesh (DWM) vocal tract model capable of movement to produce diphthongs. In [20], we further investigated on estimating a physical model of the vocal tract (a 2D model in this case) from the speech waveform, rather than magnetic resonance imaging data. As an advantage, this method provides a clear relationship between the model and the size and shape of the vocal tract, offering considerable flexibility in terms of speech characteristics such as age and gender. Finally, we have been also working on optimizing the appearance and usability of the PMA system.

5. Acknowledgements

This work was supported by the White Rose university consortium through a Collaboration Fund grant. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research.

6. References

[1] A. Byrne, M. Walsh, M. Farrelly, and K. O'Driscoll, "Depression following laryngectomy. A pilot study," *Brit J Psychiat.*, vol. 163, no. 2, pp. 173–176, 1993.

[2] D. S. A. Braz, M. M. Ribas, R. A. Dedivitis, I. N. Nishimoto, and A. P. B. Barros, "Quality of life and depression in patients undergoing total and partial laryngectomy," *Clinics*, vol. 60, no. 2, pp. 135–142, 2005.

[3] H. Danker, D. Wollbrück, S. Singer, M. Fuchs, E. Brähler, and A. Meyer, "Social withdrawal after laryngectomy," *Eur Arch Oto-Rhino-L*, vol. 267, no. 4, pp. 593–600, 2010.

[4] S. R. Ell, A. J. Mitchell, and A. J. Parker, "Microbial colonization of the groningen speaking valve and its relationship to valve failure," *Clin Otolaryngol Allied Sci.*, vol. 20, no. 6, pp. 555–556, 1995.

[5] S. R. Ell, "Candida: the cancer of silastic," *J Laryngol Otol*, vol. 110, no. 03, pp. 240–242, 1996.

[6] J. M. Heaton and A. J. Parker, "Indwelling tracheo-oesophageal voice prostheses post-laryngectomy in sheffield, uk: a 6-year review," *Acta Otolaryngol.*, vol. 114, no. 6, pp. 675–678, 1994.

[7] M. Fried-Oken, L. Fox, M. T. Rau, J. Tullman, G. Baker, M. Hindal, N. Wile, and J.-S. Lou, "Purposes of AAC device use for persons with ALS as reported by caregivers," *Augment Altern Commun.*, vol. 22, no. 3, pp. 209–221, 2006.

[8] J. A. Gonzalez, L. A. Cheah, J. Bai, S. R. Ell, J. M. Gilbert, R. K. M. 1, and P. D. Green, "Analysis of phonetic similarity in a silent speech interface based on permanent magnetic articulography," in *Proc. Interspeech*, 2014, pp. 1018–1022.

[9] J. A. Gonzalez, P. D. Green, R. K. Moore, L. A. Cheah, and J. M. Gilbert, "A non-parametric articulatory-to-acoustic conversion system for silent speech using shared gaussian process dynamical models," in *UK Speech*, 2015, p. 11.

[10] J. A. Gonzalez, L. A. Cheah, J. M. Gilbert, J. Bai, S. R. Ell, P. D. Green, and R. K. Moore, "A silent speech system based on permanent magnet articulography and direct synthesis," *Comput Speech Lang.*, vol. 39, pp. 67–87, 2016.

[11] J. A. Gonzalez, L. A. Cheah, A. M. Gomez, P. D. Green, J. M. Gilbert, S. R. Ell, R. K. Moore, and E. Holdsworth, "Direct speech reconstruction from articulatory sensor data by machine learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2362–2374, 2017.

[12] J. A. Gonzalez Lopez, L. A. Cheah, P. D. Green, J. M. Gilbert, S. R. Ell, R. K. Moore, and E. Holdsworth, "Evaluation of a silent speech interface based on magnetic sensing and deep learning for a phonetically rich vocabulary," in *Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech*. ISCA, 2017, pp. 3986–3990.

[13] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.

[14] M. J. Fagan, S. R. Ell, J. M. Gilbert, E. Sarrazin, and P. M. Chapman, "Development of a (silent) speech recognition system for patients following laryngectomy," *Med Eng Phys.*, vol. 30, no. 4, pp. 419–425, 2008.

[15] J. M. Gilbert, S. I. Rybchenko, R. Hofe, S. R. Ell, M. J. Fagan, R. K. Moore, and P. Green, "Isolated word recognition of silent speech using magnetic implants and sensors," *Med Eng Phys.*, vol. 32, no. 10, pp. 1189–1197, 2010.

[16] L. A. Cheah, J. Bai, J. A. Gonzalez, S. R. Ell, J. M. Gilbert, R. K. Moore, and P. D. Green, "A user-centric design of permanent magnetic articulography based assistive speech technology," in *Proc. BioSignals*, 2015, pp. 109–116.

[17] L. A. Cheah, J. Bai, J. A. Gonzalez, J. M. Gilbert, S. R. Ell, P. D. Green, and R. K. Moore, "Preliminary evaluation of a silent speech interface based on intra-oral magnetic sensing," in *Proc. BioDevices*, 2016, pp. 108–116.

[18] J. Mullen, D. M. Howard, and D. T. Murphy, "Waveguide physical modeling of vocal tract acoustics: flexible formant bandwidth control from increased model dimensionality," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 3, pp. 964–971, 2006.

[19] A. J. Gully, H. Daffern, and D. T. Murphy, "Diphthong synthesis using the dynamic 3d digital waveguide mesh," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 243–255, Feb 2018.

[20] A. J. Gully, T. Yoshimura, D. T. Murphy, K. Hashimoto, Y. Nankaku, and K. Tokuda, "Articulatory text-to-speech synthesis using the digital waveguide mesh driven by a deep neural network," in *Proc. Interspeech*, F. Lacerda, Ed. ISCA, 2017, pp. 234–238. [Online]. Available: http://www.isca-speech.org/archive/Interspeech_2017/abstracts/0900.html