Restoring speech following total removal of the larynx by a learned transformation from sensor data to acoustics

James M. Gilbert
School of Engineering, University of Hull, Kingston upon Hull, United Kingdom
J.M.Gilbert@Hull.ac.uk

Jose A. Gonzalez
Department of Computer Science, University of Sheffield, Sheffield, United Kingdom
J.Gonzalez@Sheffield.ac.uk

Lam A. Cheah
School of Engineering, University of Hull, Kingston upon Hull, United Kingdom
L.Cheah@Hull.ac.uk

Stephen R. Ell
Hull and East Yorkshire Hospitals Trust, Castle Hill Hospital, Cottingham, United Kingdom
srell@doctors.org.uk

Phil Green and Roger K. Moore
Department of Computer Science, University of Sheffield, Sheffield, United Kingdom
P.Green@Sheffield.ac.uk, R.K.Moore@Sheffield.ac.uk

Ed Holdsworth
Practical Control Limited, Sheffield, United Kingdom
Ed.holdsworth@practicalcontrol.com

Abstract: Total removal of the larynx may be required to treat laryngeal cancer: speech is lost. This article shows that it may be possible to restore speech by sensing movement of the remaining speech articulators and use machine learning algorithms to derive a transformation to convert this sensor data into an acoustic signal. The resulting "silent speech," which may be delivered in real time, is intelligible and sounds natural. The identity of the speaker is recognisable. The sensing technique involves attaching small, unobtrusive magnets to the lips and tongue and monitoring changes in the magnetic field induced by their movement.

[DDO]
Date Received: October 17, 2016 **Date Accepted:** February 21, 2017

1. The clinical need for silent speech

Laryngeal cancer accounts for 1% of all cancers,[1] but has a 70% 5 year survival.[2] Current methods for restoring speech include the electro-larynx, which produces an unnatural, electronic voice, oesophageal (belching) speech, which is difficult to learn, and fistula valve speech, which is considered to be the current gold standard. Valved speech, however, requires regular hospital visits for valve replacement and produces a masculine voice unpopular with female patients. All these methods sacrifice the patient's spoken identity. Laryngeal preservation would be preferable and treatment with chemo-radiotherapy (CRT) has reduced the number of total laryngectomies,[3,4] however, subsequent salvage surgery is more demanding and more subject to complications,[5] making the use of speech valves more problematic.

2. Permanent magnet articulography

A range of techniques have been investigated which extract information about a user's speech in the absence of acoustic signals. These "silent speech" technologies include electroencephalography (EEG), electromyography (EMG), ultrasound imaging of the vocal tract, electromagnetic articulography (EMA), and electropalatography.[6] All these methods have limitations in reproducing speech accurately, and are invasive and impractical outside the laboratory. As an alternative, our permanent magnet articulography (PMA) method has distinct advantages.[7] PMA utilises a set of small magnets attached to the speech articulators and a set of magnetic sensors that detect changing field patterns as the articulators move. The magnetic field monitored at each sensor is a composite, depending on the position and orientation of all of the magnetic markers. This makes it difficult to determine the Cartesian position/orientation of the magnets and so, instead, we treat the sensor signals as a nonlinear mapping of the degrees of freedom of the articulators.

The current PMA system [Fig. 1(a)] consists of six magnets: four on the lips, one on the tip of the tongue, and one on the blade of the tongue. For reasons of participant comfort and safety, no markers are placed at the back of the tongue or on the velum. For experimental purposes the markers are temporarily attached using adhesive plasters and surgical adhesive (Histoacryl®, Braun, Melsungen, Germany), although, in patients, they will ultimately be implanted. The sensor system comprises three tri-axial anisotropic magnetoresistive (AMR) sensors mounted on a bespoke wearable headset [Fig. 1(b)]. A fourth tri-axial magnetic sensor is used to allow cancellation of the background field caused by the Earth's magnetic field. Examples of speech and sensor signals for the sentence "I had been born with no organic chemical predisposition toward alcohol" recorded by a male, non-impaired British subject are shown in Fig. 1(c).

3. Direct speech synthesis from articulator movement

The core component of our "direct synthesis" technique[8–11] is a parametric transformation which maps sensor data into speech acoustics. In our current system, the parameters of the transformation are learned from a "parallel dataset"—ideally around 30 min of simultaneous recordings of articulatory and audio signals recorded by the same person before the laryngectomy. As shown in Fig. 2, the PMA and audio signals in the parallel dataset are initially pre-processed to reduce their redundancy and to represent them as sequences of feature vectors which facilitate subsequent automatic learning steps. The features extracted from the PMA data are fed as inputs to an artificial neural network, which is trained with the audio features as targets.

To model the mapping between the articulatory and speech features, we use long short-term memory (LSTM) recurrent neural networks (RNNs),[12,13] a type of RNN which recently has achieved state-of-the-art performance in a variety of speech applications such as automatic speech recognition[14] and speech synthesis.[15] An LSTM-RNN is a type of neural network designed to model temporal sequences with long-term dependencies, and is thus particularly suitable for solving sequence-to-sequence mapping problems.

In the conversion stage, as illustrated in Fig. 2, the sequence of feature vectors extracted from PMA data is propagated through the trained recurrent neural network

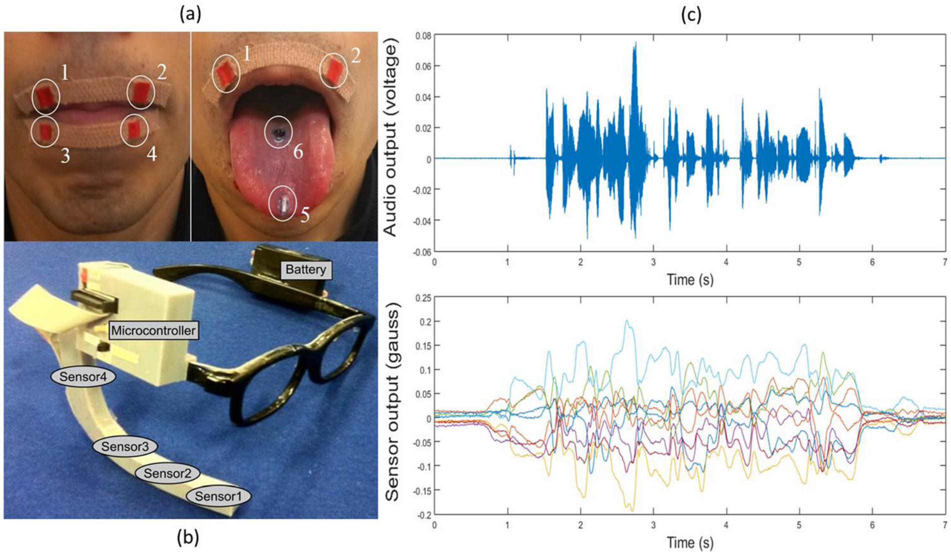

Fig. 1. (Color online) Permanent magnet articulography system. (a) Arrangement of magnetic markers on the tongue and lips. Magnets sizes are: 5 mm long by 1 mm diameter (magnets 1–4), 4 mm long and 2 mm diameter (magnet 5), and 1 mm long and 5 mm diameter (magnet 6). (b) Wearable sensor frame mounted on a pair of spectacles with four tri-axial magnetic sensors: three to measure articulator movements and one as a reference sensor to measure the Earth's magnetic field. (c) Audio (top) and nine-channel sensor data (bottom) for the sentence "I had been born with no organic chemical predisposition toward alcohol" recorded by a non-impaired British male subject.

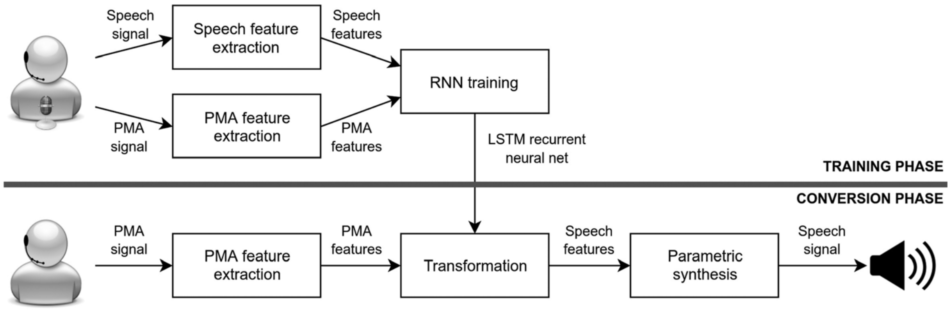

Fig. 2. Diagrams of the training and conversion phases of the direct synthesis technique. (a) In the training phase, the parameters of a LSTM recurrent neural network are optimized to minimize the prediction error of the speech features given the sequence of input feature vectors computed from the sensor data. (b) During conversion, the learned recurrent neural network is used to transform input feature vectors into speech feature vectors and, from them, a time-domain speech signal is finally synthesised.

to predict the speech features. Finally, from the sequence of predicted speech feature vectors, a time-domain signal is reconstructed and played back through a loudspeaker. Provided that the whole conversion process takes less than 250 ms (Ref. 16) we will be able to restore the normal articulation-acoustic feedback loop and therefore the user should be able to learn to improve the speech quality: like learning to play a musical instrument. Furthermore, the speech being synthesised will resemble the user's own voice, since training is based on her/his own recordings.

The training procedure described above relies on the availability of parallel data for each patient to train the neural networks. However, it may not always be possible to record parallel data directly, either because the patient has already lost their voice, or because of the short notice between diagnosis and laryngectomy prohibits arranging the recording. In this case our plan is to make audio recordings before the operation and then, after the PMA magnets have been implanted, obtain the sensor data by asking the patient to mime to these recordings. A "donor" voice from another person, perhaps a relative, could be used in the same way. Because the synchrony between modalities (PMA and speech) is not guaranteed in this case, both streams have to be time-aligned in order to obtain the parallel data.

4. Results

Before attempting to help laryngectomees, it was necessary to evaluate speech reconstruction performance on data from normal speakers. We made simultaneous speech and PMA data recordings for four non-impaired subjects: three men and one woman. In each recording session, the participant was fitted with the magnets and PMA device and parallel data were recorded for a given vocabulary. Here, the material is the CMU Arctic corpus:[17] sentences taken from novels and chosen for phonetic balance, a standard task for speech synthesis. This material is difficult: the vocabulary is unlimited and the sentences are meant to be read, not spoken. The amount of data recorded by the subjects was: 420 sentences (21.62 min), 509 sentences (26.10 min), 519 sentences (35 min), and 439 sentences (22.07 min) for the three men and the woman, respectively.

To extract the PMA and speech features, an analysis window spanning 25 ms of data with 5 ms increment is used. Speech signals are parameterised to 32-dimensional feature vectors using the STRAIGHT vocoder:[18] 25 Mel-frequency cepstral coefficients[19] (MFCCs) represent the spectral envelope and the seven remaining parameters represent the excitation signal [5-band aperiodicity values, continuous fundamental frequency (F_0) value on a logarithmic scale and binary voicing decision]. During synthesis, the vocal tract and excitation parameters estimated from the articulatory data are used to synthesise the final acoustic signal. PMA signals are first preprocessed to remove the effects of the Earth's magnetic field and involuntary head movements on the captured articulatory data. Next, data frames are computed each 5 ms from segments spanning 25 ms of data. Principal component analysis (PCA) is applied retaining the 99% of the total variance. Finally, both the PMA and speech features are mean-and-variance normalised.

Two separate LSTM-RNN models are trained for each subject from her/his parallel data: one for predicting the continuous speech features (MFCCs, five-band aperiodicities and logF_0 values) and other for the unvoiced/voiced decision. In both cases, a RNN with four hidden layers with 128, 256, 256, and 128 LSTM units is employed. During training, the network parameters are optimised with the stochastic gradient descent technique using mini-batches of 50 sentences. The networks are trained for 100 epochs or until the error computed over a validation set starts to rise.

A 10-fold cross-validation scheme is used to assess the system's performance: parallel data corresponding to 90% of the recorded sentences was used for training purposes and the remaining 10% was used for evaluation. This was repeated ten times. In each round, PMA data in the evaluation subset was processed by the recurrent neural networks trained in that round to predict the corresponding speech features. For objective evaluation of speech reconstruction accuracy, we compare the speech features predicted from PMA data with those extracted from the signals recorded by the subjects. Mel-cepstral distortion[20] (MCD) metric, a standard metric in speech coding and synthesis, is used to evaluate the accuracy of MFCC estimation. In MCD, lower values indicate better results. Roughly speaking, an MCD of four decibels (dB) is considered to be high spectral accuracy, while 5.5 dB is moderate and usually tolerable.[21] For the voicing parameter we compute the error rate and, for voiced sounds, the correlation between the actual and predicted $\log F_0$ values.

The results for the four subjects are shown in Fig. 3. In standard recurrent neural networks, the output at each time instant is computed from the current and past inputs. For voice reconstruction, the additional use of future input samples could improve the system's performance by taking into account more information about the articulators' dynamics, at the expense of introducing a certain delay. Provided this delay is less than 250 ms (Ref. 16) it will not be noticed by the user. In Fig. 3, the performance with varying delays is shown. Encouragingly, the best performance for spectral estimation is obtained with delays between 50 and 100 ms.

Since PMA technology only captures information about the upper-part of the vocal tract (lips and tongue, specifically), it is not surprising that the performance for voicing prediction is limited: between 18% to 30% error rate. For F_0 prediction, correlation between the original and predicted F_0 values is significantly better for the woman than for the men. As discussed in Ref. 22, this may be due to the higher correlation between F_0 and the first two speech formants F_1 and F_2 (which are associated with the shape of the mouth during articulation) for female speakers.

Prediction of the speech spectral parameters (represented here as MFCCs) from PMA measurements compares favourably with other methods proposed in the literature for direct speech synthesis.[8,9] LSTM-RNNs seem to be better able to model long-term correlations; thus providing more accurate voice reconstruction. From Fig. 3, MCD results obtained for the men (5.73 ± 0.008 dB, 5.60 ± 0.007 dB and 4.89 ± 0.011 dB, $p < 0.05$, with a delay of 50 ms) are slightly better than for the woman (6.17 ± 0.009 dB, $p < 0.05$, for delay of 50 ms). The results obtained for the third male subject are particularly encouraging. He recorded more training data (35 min compared to around 20 min for the other subjects, and took care to speak slowly (133 words per minute compared to an average of 174 for the other speakers) and clearly. This

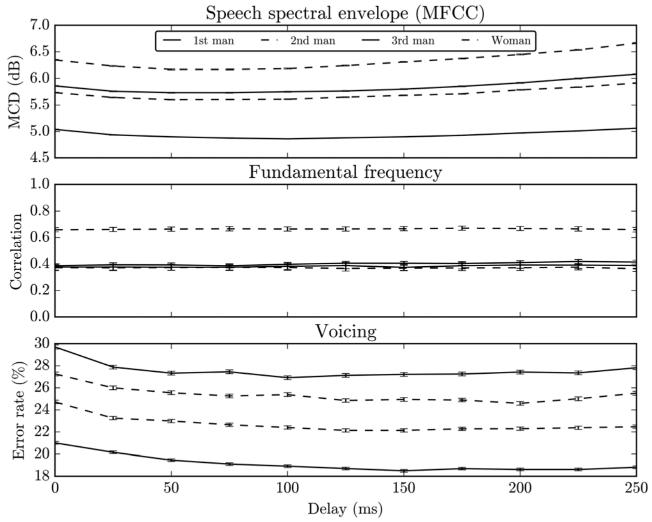

Fig. 3. Objective evaluation of speech reconstruction accuracy for the male and female subjects. Results are presented as a function of the delay corresponding to the future frames supplied to the RNN. Confidence intervals are shown at the 95% level. (a) Average Mel-cepstral distortion (MCD) results in decibels showing the spectral reconstruction accuracy (lower is better). (b) Pearson correlation between the original and predicted F_0 values (logarithmic scale) computed for the voiced speech segments (higher is better). (c) Percentage of errors for the voicing parameter (lower is better).

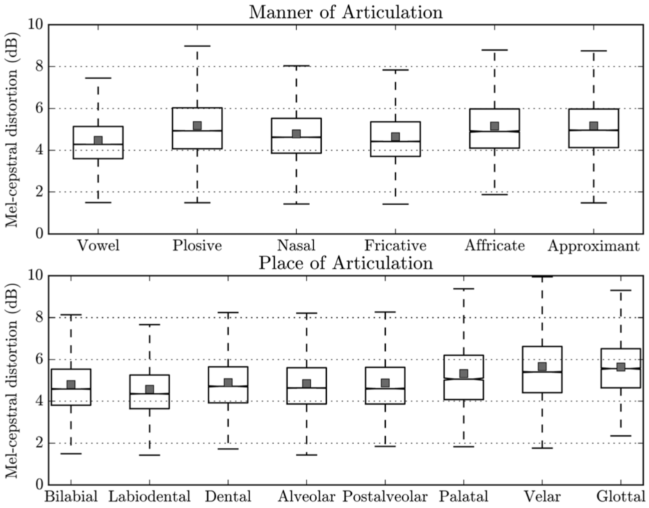

Fig. 4. Detailed spectral reconstruction results for different phone categories. Each box represents the following values of the distribution of MCD errors for each phone category: the main box represents the first quartile (lower edge), the median (red segment inside the box), and third quartile (upper edge). The mean of each distribution is represented as a small red square and the interquartile range (IQR) is IQR = Q3-Q1, where Q1 and Q3 are the first and third quartiles. Whiskers extend vertically from the boxes up to 1.5 × IQR. 95% confidence intervals around the medians are plotted with a notch in the box. (a) Average Mel-cepstral distortion results for all speakers when considering the manner of articulation. (b) Average results when looking at the place of articulation of the phones.

suggests that it may be possible for the users of our system to learn to apply compensatory strategies to improve speech intelligibility.

To determine the performance in speech reconstruction achieved by our system for the speech sounds in English, we performed a second analysis in which the MCD measure was independently computed for each phone. The distributions of MCD results are represented as a boxplot chart, in Fig. 4.

To segment the speech signals into phones, we force-aligned word-level transcriptions using an automatic speech recogniser adapted to the subject's voice. The output of the forced-alignment procedure is a phone-level transcription with timing information. We used this information to segment the original and estimated speech signals and compute the MCD measure for each phone. MCD results for phones with similar articulation are grouped together according to the manner and place of articulation.[23] There are significant differences between the reconstruction quality obtained for different phone categories. When considering the manner of articulation, the vowels and fricatives (MCDs are 4.28 ± 0.006 dB and 4.42 ± 0.009 dB, $p < 0.05$, for the medians) are among the best synthesised sounds while plosive, affricate and approximant consonants (4.93 ± 0.010 dB, 4.90 ± 0.034 Db, and 4.96 ± 0.015 dB, respectively, with $p < 0.05$) are, on average, less well reconstructed due to their more complex articulation and dynamics. When looking at the place of articulation, phones articulated at the middle and back of the mouth (palatal, velar, and glottal consonants, whose MCD values for the medians are, respectively, 5.06 ± 0.050 dB, 5.40 ± 0.023 Db, and 5.56 ± 0.032 dB with $p < 0.05$) are systematically less well reconstructed than the rest of the phones. This can be attributed to the limitations of the current PMA device for modelling those areas.[8,24] In ongoing work we are investigating ways to also capture information about articulator movement in those areas.

Audio examples. The best way to get a feel for what our system can do is to listen to some examples. We provide eight audio files Mm. 1–Mm. 8.

Mm. 1. Original sentence, "He was pressing beyond the limits of his vocabulary," male speaker 1. This is a file of type.wav (91 Kb).

Mm. 2. Original sentence, "My age, in years, is 22," male speaker 2. This is a file of type.wav (91 Kb).

Mm. 3. Original sentence, "Do not you see I hate you," male speaker 3. This is a file of type.wav (83 Kb).

Mm. 4. Original sentence, "In a way, he's my protégé," female speaker. This is a file of type.wav (63 Kb).

Mm. 5. Synthesised sentence, "He was pressing beyond the limits of his vocabulary," male speaker 1. This is a file of type.wav (91 Kb).

Mm. 6. Synthesised sentence, "My age, in years, is 22," male speaker 2. This is a file of type.wav (88 Kb).

Mm. 7. Synthesised sentence, "Do not you see I hate you," male speaker 3. This is a file of type.wav (83Kb).

Mm. 8. Synthesised sentence, "In a way, he's my protégé," female speaker. This is a file of type.wav (63 Kb).

Note that the synthesised "speech" is mostly intelligible and quite natural. The individuality of the speaker is preserved; indeed, the "voice" is recognisable if the speaker is someone you know. When listening to these examples, remember you do not have the visual cues which are of considerable help in following a speaker. These will be available when the device is deployed, because we can reconstruct speech in close to real time.

5. Conclusions

We have shown that state-of-the-art technology has the potential to be used, post laryngectomy, to provide an unobtrusive voice prosthesis which produces a voice which not only sounds natural but is recognizable as that of the patient.

Acknowledgments

This is a summary of independent research funded by the National Institute for Health Research (NIHR)'s Invention for Innovation Programme (Grant No. 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 and links

[1] Cancer Research UK, Laryngeal Cancer Statistics, http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/laryngeal-cancer#heading-Two (Last viewed 10 March 2017).

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

[3] Department of Veterans Affairs, Laryngeal Cancer Study Group, "Induction chemotherapy plus radiation compared with surgery plus radiation in patients with advanced laryngeal cancer," New Engl. J. Med. **324**, 1685–1690 (1991).

[4] A. J. Timmermans, B. A. Dijk, L. I. Overbeek, M. L. F. Velthuysen, H. Tinteren, F. J. Hilgers, and M. W. Brekel, "Trends in treatment and survival for advanced laryngeal cancer: A 20-year population-based study in The Netherlands," Head Neck **38**(S1), E1247–E1255 (2016).

[5] P. Sunil, S. P. Verma, and H. Mahboubi, "The changing landscape of total laryngectomy surgery," Otolaryng. Head Neck Surg. **150**(3), 413–418 (2014).

[6] B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, and J. S. Brumberg, "Silent speech interfaces," Speech Commun. **52**, 270–287 (2010).

[7] 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. **10**, 1189–1197 (2010).

[8] 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 speech synthesis," Comput. Speech Lang. **39**, 67–87 (2016).

[9] T. Hueber and G. Bailly, "Statistical conversion of silent articulation into audible speech using full-covariance HMM," Comput. Speech Lang. **36**, 274–293 (2015).

[10] F. Bocquelet, T. Hueber, L. Girin, C. Savariaux, and B. Yvert, "Real-time control of an articulatory-based speech synthesizer for brain computer interfaces," PLOS Comput. Biol. **12**(11), e1005119 (2016).

[11] S. Aryal and R. Gutierrez-Osuna, "Data driven articulatory synthesis with deep neural networks," Comput. Speech Lang. **36**, 260–273 (2016).

[12] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Comput. **9**(8), 1–32 (1997).

[13] F. A. Gers, N. N. Schraudolph, and J. Schmidhuber, "Learning precise timing with LSTM recurrent retworks," J. Mach. Learn. Res. **3**(1), 115–143 (2002).

[14] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proceedings of ICASSP 2013* (2013), pp. 6645–6649.

[15] Y. Fan, Y. Qian, F. Xie, and F. K. Soong, "TTS synthesis with bidirectional LSTM based recurrent neural networks," in *Proceedings of Interspeech 2014* (2014), pp. 1964–1968.

[16] A. J. Yates, "Delayed auditory feedback," Psychol. Bull. **60**, 213–232 (1963).

[17] J. Kominek and A. W. Black, "CMU ARCTIC databases for speech synthesis," technical report, Language Technologies Institute, Carnegie Mellon University (2003).

[18] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigné, "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 (1999).

[19] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for Mel-cepstral analysis of speech," in *Proceedings of ICASSP 1992* (1992), pp. 137–140.

[20] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proceedings IEEE Pacific Rim Conference on Communications, Computers and Signal Processing* (1993), pp. 125–128.

[21] J. Kominek, T. Schultz, and A. W. Black, “Synthesizer voice quality of new languages calibrated with mean Mel cepstral distortion,” in *Proceedings Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU)* (2008).

[22] K. Chládková, P. Boersma, and V. J. Podlipský, “On-line formant shifting as a function of F0,” in *Proceedings of Interspeech 2009* (2009), pp. 464–467.

[23] P. Roach, *Phonetics* (Oxford University Press, Oxford, 2001).

[24] J. A. Gonzalez, L. J. Cheah, J. Bai, S. R. Ell, J. M. Gilbert, R. K. Moore, and P. D. Green, “Analysis of phonetic similarity in a silent speech interface based on permanent magnetic articulography,” *Proceedings of Interspeech 2014* (2014), pp. 1018–1022.