# Evaluation of a Silent Speech Interface based on Magnetic Sensing and Deep Learning for a Phonetically Rich Vocabulary

*Jose A. Gonzalez*[1], *Lam A. Cheah*[2], *Phil D. Green*[1], *James M. Gilbert*[2],
*Stephen R. Ell*[3], *Roger K. Moore*[1], *Ed Holdsworth*[4]

[1]Department of Computer Science, University of Sheffield, Sheffield, UK
[2]School of Engineering, University of Hull, Kingston upon Hull, UK
[3]Hull and East Yorkshire Hospitals Trust, Castle Hill Hospital, Cottingham, UK
[4]Practical Control Ltd, Sheffield, UK

j.gonzalez@sheffield.ac.uk

## Abstract

To help people who have lost their voice following total laryngectomy, we present a speech restoration system that produces audible speech from articulator movement. The speech articulators are monitored by sensing changes in magnetic field caused by movements of small magnets attached to the lips and tongue. Then, articulator movement is mapped to a sequence of speech parameter vectors using a transformation learned from simultaneous recordings of speech and articulatory data. In this work, this transformation is performed using a type of recurrent neural network (RNN) with fixed latency, which is suitable for real-time processing. The system is evaluated on a phonetically-rich database with simultaneous recordings of speech and articulatory data made by non-impaired subjects. Experimental results show that our RNN-based mapping obtains more accurate speech reconstructions (evaluated using objective quality metrics and a listening test) than articulatory-to-acoustic mappings using Gaussian mixture models (GMMs) or deep neural networks (DNNs). Moreover, our fixed-latency RNN architecture provides comparable performance to an utterance-level batch mapping using bidirectional RNNs (BiRNNs).

**Index Terms**: speech rehabilitation, articulatory-to-acoustic mapping, recurrent neural network, speech synthesis

## 1. Introduction

In our continuing effort [1–5] to develop an acceptable and discreet speech restoration system for laryngectomees, here we propose a novel technique for transforming data captured from the speech articulators into audible speech. Inspired by recent research, we deploy deep learning techniques [6] to model the articulatory-to-acoustic mapping. In particular, we use recurrent neural networks, which have achieved state-of-the-art performance in various speech tasks in recent years [7–10], to model this mapping. Our speech restoration system is built around permanent magnet articulography (PMA) [1, 2, 11, 12], a technique for capturing the movement of the speech articulators by sensing changes in magnetic field generated by a set of small magnets attached to the articulators. The rest of the paper describes our speech restoration system and its evaluation using recordings made by non-impaired speakers.

## 2. Speech synthesis from articulator movement

To synthesise speech from PMA data, we adopt a data-driven approach in which the articulatory-to-acoustic mapping is learnt from data. In particular, a parallel dataset with simultaneous recordings of speech and PMA data is used[1]. More formally, our goal is to model parametrically the following mapping function between source feature vectors $\boldsymbol{x}_t$ computed from the PMA signal and target speech feature vectors $\boldsymbol{y}_t$ extracted, for instance, using a vocoder:

$$\boldsymbol{y}_t = \boldsymbol{f}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{t+\omega}) + \epsilon, \tag{1}$$

where $t$ is the frame index and $\epsilon$ is a zero-mean Gaussian-distributed approximation error. In addition to the past and current source vectors, we also consider for the mapping $\omega$ future source vectors as this often improves the mapping accuracy at the expense of introducing a small delay in the conversion process [4, 5, 13]. Provided that this delay is less than 50 ms, we will be able to restore the articulatory-auditory feedback without causing disfluencies or mental stress on the speaker [14, 15].

To model the mapping function in (1), we use gated RNNs. A gated RNNs is a type of artificial neural network particularly suited for modelling temporal sequences with long-time dependencies which does not exhibit the problem of vanishing/exploding gradients as other RNN architectures do. Two well-known gated RNN architectures are the long short-term memory (LSTM) [16] and the gated recurrent unit (GRU) [17]. In a set of preliminary experiments, both RNN architectures achieved similar results on our mapping problem, but the GRU-RNNs had lower training times due to having fewer parameters. Hence, in the following, we describe how to apply GRU-RNNs for modelling (1).

A GRU-RNN consists of a set of recurrently connected blocks, each one representing a GRU block. Similarly to DNNs, to increase the modelling power, several layers of GRU blocks can be stacked to create a deep RNN. The inputs of the block at layer $l$ are the hidden activations of the previous layer $\boldsymbol{h}_t^{l-1}$ (with $\boldsymbol{h}_t^0 = \boldsymbol{x}_t$) and its activations in the previous time instant $\boldsymbol{h}_{t-1}^l$. The outputs $\boldsymbol{h}_t^l$ of the block are computed by iteratively applying the following composite activation function for $t = 1, \ldots, T + \omega$ and $l = 1, \ldots, L$ [17]:

$$\boldsymbol{r}_t^l = \sigma(\boldsymbol{W}_l^r \boldsymbol{h}_t^{l-1} + \boldsymbol{V}_l^r \boldsymbol{h}_{t-1}^l + \boldsymbol{b}_l^r) \tag{2}$$

$$\boldsymbol{u}_t^l = \sigma(\boldsymbol{W}_l^u \boldsymbol{h}_t^{l-1} + \boldsymbol{V}_l^u \boldsymbol{h}_{t-1}^l + \boldsymbol{b}_l^u) \tag{3}$$

$$\tilde{\boldsymbol{h}}_t^l = \tanh(\boldsymbol{W}_l^h \boldsymbol{h}_t^{l-1} + \boldsymbol{V}_l^h (\boldsymbol{r}_t^l \odot \boldsymbol{h}_{t-1}^l) + \boldsymbol{b}_l^h) \tag{4}$$

$$\boldsymbol{h}_t^l = \boldsymbol{u}_t^l \odot \boldsymbol{h}_{t-1}^l + (1 - \boldsymbol{u}_t^l) \odot \tilde{\boldsymbol{h}}_t^l \tag{5}$$

[1] Parallel recordings should be made soon after the patient has been diagnosed with laryngeal cancer and can still speak.

where $\odot$ represents element-wise multiplication between two vectors, $\sigma$ is the logistic sigmoid function, and $\boldsymbol{r}$, $\boldsymbol{u}$ and $\tilde{\boldsymbol{h}}$ are the reset gate, update gate and candidate activation, respectively. These gates regulate the flow of information through the block and the update of the block's hidden state $\boldsymbol{h}_t^l$. The weight matrices $\boldsymbol{W}$, $\boldsymbol{V}$ and bias vectors $\boldsymbol{b}$ are trainable parameters of the RNN.

Finally, the speech features are computed from the hidden activations of the last hidden layer as follows,

$$\boldsymbol{y}_t = \phi(\boldsymbol{W}_y \boldsymbol{h}^L_{t+\omega} + \boldsymbol{b}_y), \tag{6}$$

where $\boldsymbol{W}_y$ and $\boldsymbol{b}_y$ are the trainable weight matrix and bias vector of the output layer and $\phi$ is the output activation function. For regression problems with continuous targets, $\phi$ is the identity function, i.e. $\phi(\boldsymbol{z}) = \boldsymbol{z}$. For binary classification problems, a logistic sigmoid function is used, i.e. $\phi(\boldsymbol{z}) = 1/(1+\exp(\boldsymbol{z}))$. Lastly, for multiclass classification problems, the standard softmax function is used.

To estimate the network's parameters, a stochastic version of the back propagation through time (BPTT) algorithm [18] is employed.

# 3. Experiments

To evaluate the proposed speech restoration system, we recorded parallel data for several non-impaired subjects. At this stage, our goal is to refine the hardware and conversion algorithms to have a consistent speech quality before attempting to evaluate the system on real patients. Hence, the task in this work is to predict the speech recorded by the subjects from the PMA data. The details of the evaluation framework are provided below.

## 3.1. Parallel database

Six non-impaired British subjects participated in this study: 4 males (M1 to M4) and 2 females (F1 and F2). For each subject, articulatory data was recorded in synchrony with the speech for a random subset of the CMU Arctic corpus of phonetically-rich sentences [19]. The amount of data recorded by each subject is given in Table 1.

Data recording was conducted in an acoustically-isolated room as follows. Lips and tongue movement was captured using the PMA device described in [11]. Six cylindrical neodymium-iron-boron magnets were temporarily glued to the articulators using tissue adhesive to track their movements: two on the upper lip, two on the lower lip, one at the tongue tip and, finally, one on the tongue blade. The magnetic field variations caused by the magnets were then measured by 3 tri-axial magnetoresistive sensors sampled at 100 Hz. Subjects' speech was simultaneously recorded using a AKG C1000S condenser microphone located about 20 cm from the subject at a sampling rate of 48 kHz. Later, the signals were digitally downsampled to 16 kHz. More details about the recording protocol can be found in [4, 11].

## 3.2. Signal processing

Speech and PMA signals are parametrized as sequences of feature vectors computed every 5 ms from 25 ms length analysis windows. The STRAIGHT vocoder [20] is used to parametrize the speech signals as 32-dimensional feature vectors: spectral envelope is encoded as 25 Mel-Frequency Cepstral Coefficients (MFCCs) [21] and the remaining 7 parameters represent the excitation signal as 5-band aperiodicities (BAPs) (0-1, 1-2, 2-4, 4-6, 6-8 kHz), unvoiced/voiced (U/V) decision and continuous $F_0$ in logarithmic scale ($\log F_0$ is linearly interpolated in unvoiced frames). PMA signals, on the other hard, are firstly oversampled from 100 Hz to 200 Hz to match the 5 ms analysis rate. Data frames are then extracted from the oversampled signals at the same frame rate as the speech signals. To improve the performance, the GMM and DNN based mappings described below are trained with segmental features computed by applying the partial least squares (PLS) dimensionality reduction technique [22] over short symmetric windows with $\delta$ consecutive PMA frames. Finally, the PMA and speech features are normalised in mean and variance.

Table 1: *Details of the parallel PMA-and-speech database recorded for the experiments.*

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

## 3.3. Model training

Speaker-dependent RNN models are trained for each speech feature type (i.e. MFCCs, BAPs, $\log F_0$, and U/V decision) using the subject's recordings. To determine the best RNN architecture, we conducted a set of preliminary experiments using a development dataset. We found that RNNs with 4 hidden layers and 150 GRUs in each layer provide the best objective results for our data (using more layers or more units per layer only gives marginal improvements). Regarding the length of look-ahead window used by the RNNs, in our previous work [5] we found that using $\omega = 10$ inputs in the future (i.e. a latency of 50 ms) provides a good trade-off between mapping latency and accuracy, so we use the same value here.

RNN training and inference are implemented using TensorFlow [23]. In training, the RNN weights are initialized randomly (without pretraining) and optimised using the Adam algorithm [24] with minibatches of 50 sentences and a learning rate of 3e-3. As a regularization technique, we add white noise to the inputs ($\sigma_{\text{noise}} = 0.5$). We employ the sum-of-squared errors (SSE) loss when optimizing the RNN parameters for the continuous speech features (MFCCs, BAPs, and $\log F_0$) and the cross-entropy function for the U/V decision. RNNs are trained for 100 epochs or until the error on a validation set does not improve after 20 epochs.

For comparison purposes, we also evaluate mappings using GMMs [4, 25, 26] and DNNs [13, 27], which have been successfully applied by ourselves and other authors to model the articulatory-to-acoustic mapping. For a fair comparison, GMMs and DNNs with approximately the same number of parameters as the RNN architecture (~$1/2$ million parameters) are employed. The GMMs have 128 mixtures with full covariance matrices. DNNs with 4 hidden layers and 400 rectified linear units (ReLUs) in each layer are used. Moreover, again for a fair comparison, those models are trained with segmental features computed from symmetric windows with $\delta = 21$ frames (i.e. the length of the look-ahead window is $\omega = 10$ frames as in the RNNs). In both mappings, the maximum likelihood parameter generation (MLPG) algorithm considering dynamic features proposed in [25, 28] is applied to smooth out the predicted speech feature trajectories, as has been repeatedly shown

Table 2: *Objective results for the mapping techniques.*

|  | MCD (dB) | BAP (dB) | $F_0$ RMSE (Hz) | U/V error rate (%) |
|---|---|---|---|---|
| GMM | 5.84 | 4.56 | 26.30 | 18.71 |
| DNN | 5.74 | 4.68 | 26.13 | 15.88 |
| RNN | 5.56 | **4.50** | **23.77** | 13.23 |
| BiRNN | **5.52** | 4.52 | 24.87 | **13.12** |

[25, 26, 29, 30], MLPG outperforms the basic frame-by-frame mappings in terms of objective and perceived speech quality.

Finally, BiRNNs [31,32], which perform an utterance-level batch mapping, are also evaluated in this work. Although BiRNNs are not suitable for real-time processing, it is interesting to compare their performance with that obtained by the proposed fixed-lag RNN architecture above.

### 3.4. Performance evaluation

A 10-fold cross-validation scheme is used to assess the performance of the proposed techniques. Performance was measured using objective and subjective quality metrics. To objectively evaluate mapping performance, the Mel-cepstral distortion (MCD) (dB) [33], root mean squared error (RMSE) of band aperiodicities (dB), RMSE of $F_0$ on a linear scale, and U/V error rate (%) metrics are used. These objective metrics are known to not correlate well with perceived speech quality, but are useful for comparing performance among different techniques and for tuning the systems. We also conducted a listening test to subjectively evaluate the techniques. The details of the listening test are provided below.

## 4. Results

Table 2 summarises the average across all subjects of the objective results for the mapping techniques[2]. The best result for each metric is in bold. Clearly, the mappings using neural networks outperform the more traditional GMM-based mapping in all the objective measures (except for the aperiodicities, where GMM outperforms DNN). Moreover, it is also clear that both types of RNN produce significantly better results than the DNN. For instance, the relative improvements of BiRNN wrt DNN are 3.83%, 3.42%, 4.82% and 17.38%, for metrics MCD to U/V error rate. Although the MLPG algorithm used to post-process the DNN predictions makes use of all past and future contexts (as the BiRNNs), the recurrent networks are better at modelling the sequential nature of speech compared to the simple smoothing carried out by this algorithm. The improvement of RNN and BiRNN over DNN is particularly noticeable for the excitation parameters ($F_0$ and voicing), for which the recurrent networks more accurately capture their long-term correlations. From the comparison between the RNN and BiRNN methods, we see that RNN achieves comparable performance to BiRNN (actually, the RNN mapping obtain better results for the aperiodicities and $F_0$ parameters), but with much lower latency.

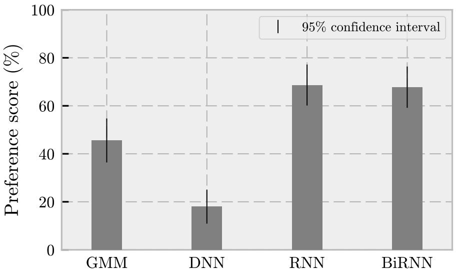

Figure 1: *Results of the ABX test on speech quality.*

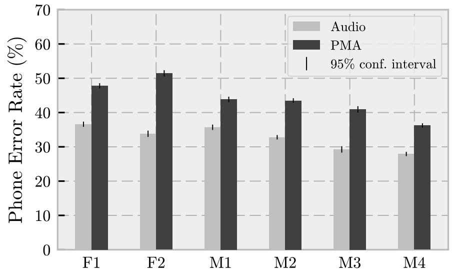

Figure 2: *Frame-wise phoneme classification results for the audio and PMA modalities.*

We also conducted an ABX listening test to subjectively evaluate speech quality. In the test, listeners heard a reference sample (one of the signals recorded by the subjects) and two versions of it produced by any of the 4 mapping techniques. Listeners were asked to judge which of the resynthesised samples was more similar to the reference. Each of the 18 listeners who participated in the test evaluated 10 sample pairs for each of the 6 possible mapping combinations (i.e. 60 pair evaluated in total). Fig. 1 shows the results of the listening test. As expected, speech generated by the RNN-based systems is preferred to that synthesised by the other mappings. Encouragingly, the RNN and BiRNN mappings are considered equally good, despite BiRNNs having the potential advantage of exploiting all the future context while RNNs only looks a little way into the future. Therefore, in the rest of this paper, we will only focus in the RNN-based mapping. Interestingly, GMMs obtained higher subjective scores than DNNs despite the latter achieving better objective results in Table 2. By listening to the speech samples generated by both mappings, the problem seems to be that the DNN samples sounds 'buzzy', which might be related to the fact that this method obtained the worse results for the BAPs in Table 2.

Next, we conducted a speech recognition experiment to determine the limitations of PMA and the fixed-lag RNN mapping for phonetic modelling. In this experiment, we trained RNNs to perform framewise phoneme classification from either audio or PMA data. Phone-level transcriptions for training the RNNs were obtained by force-aligning the audio signals using a speaker-dependent, triphone-based speech recogniser. Fig. 2 shows the recognition results for all subjects in the database. Firstly, we see that significantly better recognition results are obtained when using audio: the average phone error rate (PER) across all subjects is 32.68% for the audio and 43.95% for PMA. Also, the recognition results using PMA are significantly better for the male than the female subjects. As discussed in [4], this might be due to the fact that the PMA prototype was designed to fit the head anatomy of subject M1. Interestingly, the best recognition results are obtained for M4, who recorded more data and also and took care to speak clearly and slowly (133 words per minute (wpm) compared to an average of 174 wpm for the other subjects).

Fig. 3 shows the phone confusions averaged across all subjects for PMA. To sum up, it can be seen that the major confusions are for the following phones:

- Vowels: /ah/ (30% accuracy), /ax/ (40%), /ea/ (38%), /ia/

[2] Speech samples generated by the mapping techniques can be found at `http://www.jandresgonzalez.com/is2017`

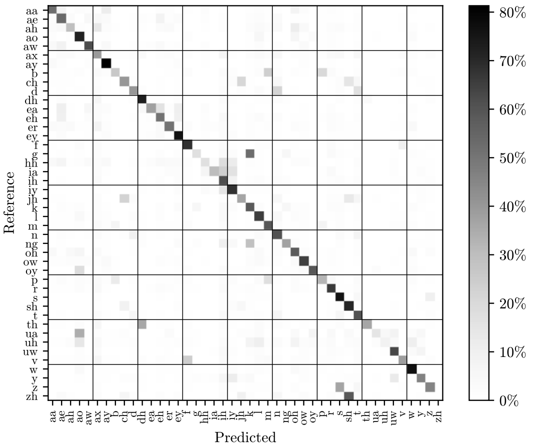

Figure 3: *Normalized confusion matrix for phoneme classification from PMA data.*

Table 3: *Objective results of the RNN-based mapping for different types of input features.*

| Input features | MCD (dB) | BAP (dB) | $F_0$ RMSE (Hz) | U/V error rate (%) |
|---|---|---|---|---|
| PMA | 5.56 | 4.50 | 23.77 | 13.23 |
| Phones | 4.95 | **3.72** | **22.70** | **8.84** |
| Senones | 5.15 | 3.85 | 22.97 | 9.28 |
| PMA+Phones | **4.75** | 3.79 | 22.81 | 9.18 |
| PMA+Senones | 4.90 | 3.88 | 22.88 | 9.65 |

(30%), /ua/ (15%) and /uh/ (9%), which are mainly confused with other vowels with similar articulation.

- Plosive consonants: /p/ (33%), /b/ (26%), /d/ (40%). For these phones, most errors are voicing confusions (e.g. /b/→/p/ (20% confusions), /d/→/t/ (17%), /p/→/b/ (18%)) and manner confusions (e.g. /b/→/m/ (24%), /d/→/n/ (22%), /p/→/m/ (18%)). As discussed in [4,34], those aspects of speech articulation (i.e. voicing and manner) are not well captured by PMA.
- Consonants articulated at the back of the mouth: velars /g/ (17% accuracy), /k/ (58%) and /ng/ (37%), and glottal /hh/ (17%). PMA fails to model those areas because no magnet were attached to them in our experiments.
- Fricatives: /th/ (35%), /zh/ (0%) and /v/ (38%). Most errors correspond to voicing confusions: /th/→/dh/ (36% confusions), /zh/→/sh/ (59%) and /v/→/f/ (25%).
- Affricates: /ch/ (40%) and /jh/ (37%).

From results in Figs. 2 and 3, we can conclude that PMA does not capture the movements of some vocal tract areas well. While these limitations could in principle be addressed in future work, we ask ourselves if it also possible to address them from a machine learning perspective: by exploiting linguistic information as a prior knowledge in the conversion process. To shed some light into this question, we designed an experiment to compare the performance of the RNN-based mapping trained with different input features: PMA features, linguistic features[3] (either phone or senone labels), and both PMA and linguistic features. In this oracle experiment, linguistic features are obtained from the force-aligned phonetic transcriptions of the audio signals, but in a real-system they could be obtained by running a speech recogniser in parallel with our mapping technique.

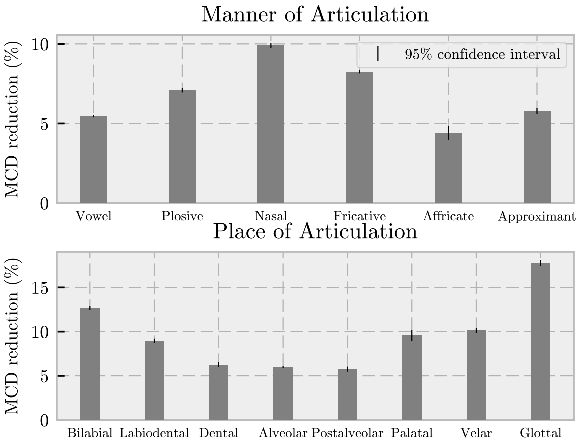

Figure 4: *Relative improvement in MCD per phone category when using PMA+Phones features wrt just using PMA features.*

Table 3 shows the objective results for each type of feature. Clearly, the linguistic features provide more information than the PMA data alone, but when both types of feature are combined together, the resulting system obtains better results than the separate systems (except for the excitation parameters, for which the best results are obtained using only phone labels[4]). Thus, it seems that the two types of feature provide complementary information. While it might be difficult to obtain these improvements in a real system due to automatic speech recognition (ASR) errors, the results seem to indicate that, indeed, the exploitation of linguistic knowledge might be useful for obtaining better speech quality in PMA-to-acoustic mapping. Finally, Fig. 4 shows the relative improvement in the MCD metric of the PMA+Phones system wrt the baseline system using only PMA features. Not surprisingly, the biggest improvements are for the sounds that PMA has most problems with: phones articulated at the back of the mouth (palatal, velar and glottal consonants), plosives and nasals.

## 5. Conclusions

We have described a technique for synthesising speech from articulatory data acquired through PMA, which could potentially restore speech after laryngectomy. Through an extensive evaluation, we have shown that our method produces reasonable speech quality and with room for improvement either by improving the sensing technique or by introducing linguistic constraints in the conversion process. We are about to enter a clinical trial where our system will be evaluated on real patients.

## 6. 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.

[3] The resulting mapping could be seen as a very basic TTS system.

[4] Again, this is due to voicing parameters not being well captured by PMA.

## 7. References

[1] 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.

[2] 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.

[3] R. Hofe, S. R. Ell, M. J. Fagan, J. M. Gilbert, P. D. Green, R. K. Moore, and S. I. Rybchenko, "Small-vocabulary speech recognition using a silent speech interface based on magnetic sensing," *Speech Commun.*, vol. 55, no. 1, pp. 22–32, 2013.

[4] 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.

[5] J. M. Gilbert, J. A. Gonzalez, L. A. Cheah, S. R. Ell, P. Green, R. K. Moore, and E. Holdsworth, "Restoring speech following total removal of the larynx by a learned transformation from sensor data to acoustics," *J. Acoust. Soc. Am.*, vol. 141, no. 3, pp. EL307–EL313, 2017.

[6] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[7] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. ICASSP*, 2013, pp. 6645–6649.

[8] Y. Fan, Y. Qian, F.-L. Xie, and F. K. Soong, "TTS synthesis with bidirectional LSTM based recurrent neural networks," in *Proc. Interspeech*, 2014, pp. 1964–1968.

[9] H. Zen and H. Sak, "Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis," in *Proc. ICASSP*, 2015, pp. 4470–4474.

[10] L. Sun, S. Kang, K. Li, and H. Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," in *Proc. ICASSP*, 2015, pp. 4869–4873.

[11] 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.

[12] 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.

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

[14] A. J. Yates, "Delayed auditory feedback," *Psychological bulletin*, vol. 60, no. 3, p. 213, 1963.

[15] A. Stuart, J. Kalinowski, M. P. Rastatter, and K. Lynch, "Effect of delayed auditory feedback on normal speakers at two speech rates," *J. Acoust. Soc. Am.*, vol. 111, no. 5, pp. 2237–2241, 2002.

[16] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[17] K. Cho, B. Van Merriënboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proc. EMNLP*, 2014, pp. 1724–1734.

[18] P. J. Werbos, "Backpropagation through time: what it does and how to do it," *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1550–1560, 1990.

[19] J. Kominek and A. W. Black, "The CMU Arctic speech databases," in *Fifth ISCA Workshop on Speech Synthesis*, 2004, pp. 223–224.

[20] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne, "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 communication*, vol. 27, no. 3, pp. 187–207, Apr. 1999.

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

[22] S. De Jong, "SIMPLS: an alternative approach to partial least squares regression," *Chemometrics Intell. Lab. Syst.*, vol. 18, no. 3, pp. 251–263, 1993.

[23] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mane, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viegas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. (2015) Tensorflow: Large-scale machine learning on heterogeneous distributed systems. [Online]. Available: www.tensorflow.org

[24] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, 2015.

[25] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 8, pp. 2222–2235, Nov. 2007.

[26] ——, "Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model," *Speech Commun.*, vol. 50, no. 3, pp. 215–227, Mar. 2008.

[27] 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 Computational Biology*, vol. 12, no. 11, p. e1005119, 2016.

[28] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. ICASSP*, 2000, pp. 1315–1318.

[29] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. ICASSP*. IEEE, 2013, pp. 7962–7966.

[30] K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "The effect of neural networks in statistical parametric speech synthesis," in *Proc. ICASSP*, 2015, pp. 4455–4459.

[31] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, 1997.

[32] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Networks*, vol. 18, no. 5, pp. 602–610, 2005.

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

[34] J. A. Gonzalez, L. A. 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," in *Proc. Interspeech*, 2014, pp. 1018–1022.