Voice Restoration after Laryngectomy based on Magnetic Sensing of Articulator Movement and Statistical Articulation-to-Speech Conversion

Jose A. Gonzalez[1], Lam A. Cheah[2], James M. Gilbert[2], Jie Bai[2], Stephen R. Ell[3], Phil D. Green[1], and Roger K. Moore[1]

[1] Department of Computer Science, University of Sheffield, Sheffield, U.K.
{j.gonzalez,p.green,r.k.moore}@sheffield.ac.uk
[2] School of Engineering, University of Hull, Kingston upon Hull, U.K.
{l.cheah,j.m.gilbert,j.bai}@hull.ac.uk
[3] Hull and East Yorkshire Hospitals Trust, Castle Hill Hospital, Cottingham, U.K.
srell@doctors.org.uk

Abstract. In this work, we present a silent speech system that is able to generate audible speech from captured movement of speech articulators. Our goal is to help laryngectomy patients, i.e. patients who have lost the ability to speak following surgical removal of the larynx most frequently due to cancer, to recover their voice. In our system, we use a magnetic sensing technique known as Permanent Magnet Articulography (PMA) to capture the movement of the lips and tongue by attaching small magnets to the articulators and monitoring the magnetic field changes with sensors close to the mouth. The captured sensor data is then transformed into a sequence of speech parameter vectors from which a time-domain speech signal is finally synthesised. The key component of our system is a parametric transformation which represents the PMA-to-speech mapping. Here, this transformation takes the form of a statistical model (a mixture of factor analysers, more specifically) whose parameters are learned from simultaneous recordings of PMA and speech signals acquired before laryngectomy. To evaluate the performance of our system on voice reconstruction, we recorded two PMA-and-speech databases with different phonetic complexity for several non-impaired subjects. Results show that our system is able to synthesise speech that sounds as the original voice of the subject and also is intelligible. However, more work still need to be done to achieve a consistent synthesis for phonetically-rich vocabularies.

Keywords: Silent speech interfaces, speech rehabilitation, speech synthesis and permanent magnet articulography.

1 Introduction

People whose larynx have been surgically removed following throat cancer, trauma or destructive throat infection normally find themselves struggling with oral

communication after losing their voice. This often has a severe impact on people's lives and can lead to social isolation, feelings of loss of identity and, sometimes, clinical depression [3,2,7]. To make things worse, existing methods for voice restoration are far from ideal [13,18]. The 'gold-standard', the tracheo-oesoephageal valve, requires frequent replacement every 3-4 months due to biofilm growth and, for this reason, is an expensive treatment [22,12,11]. The electrolarynx, on the other hand, despite being relatively easy to use and safe, produces a robotic voice. Finally, oesophageal speech, a technique in which air is injected into the mouth and then it is released in a controlled manner to make the oesophagus vibrate to create speech, sounds gruff and masculine and, also, is difficult to learn. Other methods such as Augmentative and Alternative Communication (AAC) devices, where the user types words and the device synthesises them, are only suitable for short conversations due to their slow manual text input [15].

In an attempt to surpass the limitations of current existing methods, we propose in this work a completely different approach for post-laryngectomy voice rehabilitation. Our proposed method can be seen as a Silent Speech Interface (SSI), which is a system that enables oral communication in the absence of audible speech [9]. Somewhat akin to lip reading, the most common way a SSI works is by first decoding the message encoded in other biosignals associated with speech production using Automatic Speech Recognition (ASR) software and then synthesising the recognised text using a Text-To-Speech (TTS) synthesiser. Many different SSIs have been proposed so far, mainly differing in the type of biosignal they rely on. Thus, we can find SSIs that exploit the electrical signals generated by the neurons in the brain [23] or in the articulator muscles [31,42,49] or the movement of the speech articulators themselves [40,44,9,29,18,14,26,21]. In our work we use a magnetic sensing technique known as Permanent Magnet Articulography (PMA) [13,18] for capturing the movement of the speech articulators. In brief, the principle of PMA is that articulator movement can be captured by attaching a set of permanent magnets to the articulators (typically the lips and tongue) and then sensing the variations of the resultant magnetic field generated while the person articulates words with sensors located close to the mouth. Compared to other techniques for articulator motion capture, such as Electromagnetic Articulography (EMA) or surface Electromiography (sEMG), PMA has the potential advantage of being unobtrusive, since there are no wires coming out of the mouth or electrodes attached to the skin.

The recognise-then-synthesise approach outlined above is the most common way of synthesising speech from articulator movement. This approach, however, is not exempt from problems [21]. Firstly, due to the variable delay introduced by the ASR and TTS systems, speech articulation and the corresponding acoustic feedback produced by the SSI are disconnected (i.e. speech is not generated in real time). An analogy for this would be like having an interpreter: the person 'mouths' words, waits and, after a while, the SSI generates the corresponding acoustic signal. Another limitation is that speech can only be synthesised for the language and vocabulary of the ASR and TTS systems. Finally, as another limitation, the non-linguistic information embedded in the articulatory signal,

such as emotion or speaker identity, is normally lost after speech recognition. To address these shortcomings, we investigate in this work an alternative approach for SSI-based voice restoration known as *direct speech synthesis*.

In the direct synthesis approach, a parametric transformation is applied to the captured articulatory data to obtain a sequence of speech parameter vectors from which a time-domain speech signal is synthesised. Thus, no intermediate speech recognition step is performed. To enable this method, we present a statistical approach in which simultaneous recordings of articulator movement and speech data captured before laryngectomy are used to learn the parameters of the transformation. In particular, in our statistical framework the transformation adopts the form of a joint probability distribution represented as a Mixture of Factor Analysers (MFA) [17]. Once this transformation is learned, it can be used in conversion time to compute the speech-parameter posterior distribution given the articulatory data which, in turn, allows us to estimate the speech parameters associated with the measured articulator gesture. Because the transformation is estimated from recordings of the the patient's voice, our method has the potential to synthesise speech that sounds as the original voice. Moreover, if the conversion can be done in real-time, the voice will sound spontaneous and natural and it will enable the patient to receive real-time acoustic feedback of her/his articulation.

This chapter is organised as follows. First, the details of the PMA technique for capturing articulator movement are provided in Section 2. In Section 3, we describe the proposed technique for synthesising audible speech from PMA data. Then, in Section 3.3, we discuss some practical implementation issues. The details about the experimental evaluation of the proposed technique and the results obtained can be found in Section 5. Finally, we summarise this work and outline some future research in Section 6.

2 Articulator Motion Capture based on Magnetic Sensing

As commented above, in our work we use PMA, a magnetic sensing technique, for capturing the movements of the speech articulators, more typically the lips and tongue [13,18,5]. The principle of PMA is simple as illustrated in Fig. 1: small magnets are attached to the speech articulators whose movement we wish to capture and magnetic sensors located close to the mouth are employed for measuring the magnetic field generated by the magnets. In the current set-up magnets are temporarily attached to the flesh using Histoacryl surgical tissue adhesive (Braun, Melsungen, Germany), but eventually the magnets will be surgically implanted for long term usage. A total of six magnets are used: four are attached to the lips (1 mm diameter $\times$ 5 mm height), one to the tongue tip (2 mm $\times$ 4 mm), and one to the tongue blade (5 mm $\times$ 1 mm). As shown in Fig. 1, an external headset is used to hold the four magnetic sensors employed in the current prototype, though other arrangements such as an intra-oral device similar to a dental retainer has been and are currently being investigated [5,4]. From the four sensors in Fig. 1, only the first three sensors (Sensor1-Sensor3),

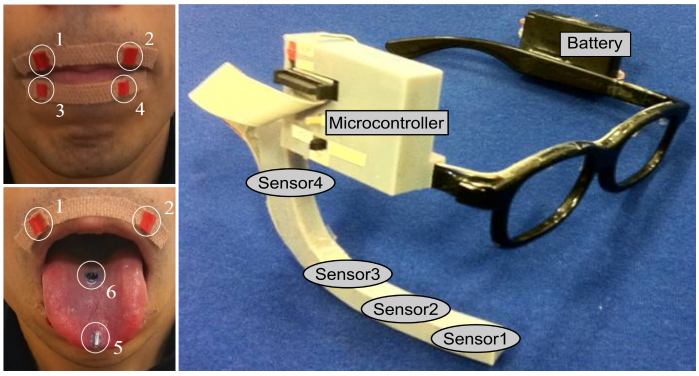

Fig. 1. *Upper-left and lower-left*: placement of magnets used for measuring lips and tongue movements. *Right*: components of the PMA headset: micro-controller, battery and magnetic sensors used to detect the variations of the magnetic field generated by the magnets.

the ones which are closer to the mouth, are actually used for data acquisition, while the last one (Sensor4) is used as a reference sensor for Earth's magnetic field cancellation in the data captured by Sensor1-Sensor3 [5]. Each sensor provides three channels of data for the 3D spatial components of the magnetic field at the sensor location, thus making 9 channels of data in total.

The data recorded by the sensors may then be used to determine the speech associated with the articulatory gestures, either by performing ASR on the PMA data [18,26] or by transforming the data to an acoustic signal, as we do in this work. It should be noted that contrary to other mechanisms for capturing articulator movement data such as EMA, in PMA the exact coordinates of the magnets in the mouth are unknown since the magnetic field sensed by the sensors is a composite of the fields generated by all the magnets. Nevertheless, as each articulatory gesture generates a recognisable pattern in the captured articulatory data, we can resort to statistical approaches for modelling the relationship between the PMA data and the corresponding acoustics. This is developed in the next section.

3 Speech Synthesis from Articulator Movement

In this section we describe the technique for synthesising audible speech from articulator movement data. The goal of the proposed technique is to model the mapping $\boldsymbol{y}_t = \boldsymbol{f}(\boldsymbol{x}_t)$ between source parameter vectors $\boldsymbol{x}_t$ derived from the PMA signal and target parameter vectors $\boldsymbol{y}_t$, which correspond to a low-dimensional, parametric representation of the audio signal. Because the positions of the magnets cannot be easily inferred from PMA data (inference of the Cartesian coordinates of the magnets from the magnetic field captured by the sensors is an inverse problem), we cannot resort here to approaches such as that proposed in

[45,47] where 2-dimensional vocal tract shapes are directly computed from the measured positions of the EMA sensors and, from the vocal tract shapes, speech is synthesised by using an articulatory synthesis method [35]. In contrast, we propose a data-driven approach for modelling the mapping function $\boldsymbol{y}_t = \boldsymbol{f}(\boldsymbol{x}_t)$. The proposed approach assumes the existence of a parallel dataset containing simultaneous recordings of PMA and acoustic data acquired before laryngectomy. From this parallel dataset, the parameters of the mapping function, which is represented here as a Mixture of Factor Analysers (MFA) [17], are estimated during an initial training stage. Later, in conversion time, the learned transformation is used to convert PMA parameter vectors into speech parameter ones. More details about the training and conversion phases are given in the next sections.

3.1 Training Phase

Instead of trying to directly model the mapping function $\boldsymbol{y}_t = \boldsymbol{f}(\boldsymbol{x}_t)$, we assume that $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ are the outputs of a stochastic process whose state $\boldsymbol{v}_t$ is not directly observable. We also assume that the dimensionality of $\boldsymbol{v}_t$ is much less than that of $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$, such that the latent space offers a more parsimonious representation of the observable data. Under these assumptions, we have the following model:

$$\boldsymbol{x}_t = \boldsymbol{f}_x(\boldsymbol{v}_t) + \boldsymbol{\epsilon}_x, \tag{1}$$

$$\boldsymbol{y}_t = \boldsymbol{f}_y(\boldsymbol{v}_t) + \boldsymbol{\epsilon}_y, \tag{2}$$

where $\boldsymbol{\epsilon}_x$ and $\boldsymbol{\epsilon}_y$ are Gaussian-distributed noise processes with zero mean and diagonal covariances $\boldsymbol{\Psi}_x$ and $\boldsymbol{\Psi}_x$, respectively.

In general, $\boldsymbol{f}_x$ and $\boldsymbol{f}_y$ will be non-linear and, hence, difficult to model. To represent them, a piece-wise linear regression approach is adopted in which the functions are approximated by a mixture of K local factor analysis models, each of which has the following form,

$$\boldsymbol{x}_t^{(k)} = \boldsymbol{W}_x^{(k)}\boldsymbol{v}_t + \boldsymbol{\mu}_x^{(k)} + \boldsymbol{\epsilon}_x^{(k)}, \tag{3}$$

$$\boldsymbol{y}_t^{(k)} = \boldsymbol{W}_y^{(k)}\boldsymbol{v}_t + \boldsymbol{\mu}_y^{(k)} + \boldsymbol{\epsilon}_y^{(k)}, \tag{4}$$

where $k = 1, \ldots, K$ is the model index; $\boldsymbol{W}_x^{(k)}$, $\boldsymbol{W}_y^{(k)}$ are the factor loadings matrices; $\boldsymbol{\mu}_x^{(k)}$, $\boldsymbol{\mu}_y^{(k)}$ are bias vectors that allow the data to have a non-zero mean, and $\boldsymbol{x}_t^{(k)}$, $\boldsymbol{y}_t^{(k)}$ are respectively local approximations of $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ around the means $\boldsymbol{\mu}_x^{(k)}$, $\boldsymbol{\mu}_y^{(k)}$. This model can be written more compactly as,

$$\boldsymbol{z}_t^{(k)} = \boldsymbol{W}_z^{(k)}\boldsymbol{v}_t + \boldsymbol{\mu}_z^{(k)} + \boldsymbol{\epsilon}_z^{(k)}, \tag{5}$$

where $\boldsymbol{z}_t = [\boldsymbol{x}_t^{(k)^\top}, \boldsymbol{y}_t^{(k)^\top}]^\top$, $\boldsymbol{W}_z^{(k)} = [\boldsymbol{W}_x^{(k)^\top} \boldsymbol{W}_y^{(k)^\top}]^\top$, $\boldsymbol{\mu}_z^{(k)} = [\boldsymbol{\mu}_x^{(k)^\top}, \boldsymbol{\mu}_y^{(k)^\top}]^\top$, and $\boldsymbol{\epsilon}_z^{(k)} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Psi}_z^{(k)})$, with $\boldsymbol{\Psi}_z^{(k)}$ being the following diagonal covariance matrix,

$$\boldsymbol{\Psi}_z^{(k)} = \begin{bmatrix} \boldsymbol{\Psi}_x^{(k)} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Psi}_y^{(k)} \end{bmatrix}. \tag{6}$$

From (5) we see that the conditional distribution of the observed variables given the latent ones is $p(\boldsymbol{z}|\boldsymbol{v},k)=\mathcal{N}(\boldsymbol{z};\boldsymbol{W}_z^{(k)}\boldsymbol{v}+\boldsymbol{\mu}_z^{(k)},\boldsymbol{\Psi}_z^{(k)})$. By assuming that the latent variables are independent and Gaussian with zero mean and unit variance (i.e. $p(\boldsymbol{v}|k)=\mathcal{N}(\boldsymbol{0},\boldsymbol{I})$), the k-th component marginal distribution of the observed variables, i.e.

$$p(\boldsymbol{z}|k)=\int p(\boldsymbol{z}|\boldsymbol{v},k)p(\boldsymbol{v}|k)d\boldsymbol{v}, \tag{7}$$

also becomes normally distributed as $p(\boldsymbol{z}|k)=\mathcal{N}(\boldsymbol{z};\boldsymbol{\mu}_z^{(k)},\boldsymbol{\Sigma}_z^{(k)})$, where

$$\boldsymbol{\Sigma}_z^{(k)}=\boldsymbol{\Psi}_z^{(k)}+\boldsymbol{W}_z^{(k)}\boldsymbol{W}_z^{(k)^\top} \tag{8}$$

is the reduced-rank covariance matrix.

The generative model is completed by adding mixture weights $\pi^{(k)}$ for each mixture component ($\sum_k \pi^{(k)}=1$). Then, the joint distribution $p(\boldsymbol{z})\equiv p(\boldsymbol{x},\boldsymbol{y})$ finally becomes the following mixture model,

$$p(\boldsymbol{z})=\sum_{k=1}^{K}\pi^{(k)}p(\boldsymbol{z}|k). \tag{9}$$

To learn the parameters $\{\langle \pi^{(k)},\boldsymbol{\mu}_z^{(k)},\boldsymbol{W}_z^{(k)},\boldsymbol{\Psi}_z^{(k)}\rangle, k=1,\ldots,K\}$ of the MFA model in (9) we use a slightly modified version the expectation-maximization (EM) algorithm proposed in [17] in which the noise covariances $\boldsymbol{\Psi}_z^{(k)}$ are cluster dependent. Let $\left\{\boldsymbol{z}_i=[\boldsymbol{x}_i^\top,\boldsymbol{y}_i^\top]^\top, i=1,\ldots,N\right\}$ be the parallel dataset used for training. Then, the E-step and M-step of the EM algorithm are as follows:

1) *E-step*: Using the MFA parameters estimated in the previous iteration, compute the component posterior probabilities $\gamma_i^{(k)}=P(k|\boldsymbol{z}_i)$ and the expectations $\langle\boldsymbol{v}_{ik}\rangle$ and $\langle\boldsymbol{v}_{ik}\boldsymbol{v}_{ik}^\top\rangle$ for the hidden variables:

$$\gamma_i^{(k)}=\frac{p(\boldsymbol{z}_i|k)\pi^{(k)}}{\sum_{k'=1}^{K}p(\boldsymbol{z}_i|k')\pi^{(k')}}, \tag{10}$$

$$\langle\boldsymbol{v}_{ik}\rangle=\boldsymbol{S}_k^{(k)}\left(\boldsymbol{z}_i-\boldsymbol{\mu}_z^{(k)}\right), \tag{11}$$

$$\langle\boldsymbol{v}_{ik}\boldsymbol{v}_{ik}^\top\rangle=\boldsymbol{I}-\boldsymbol{S}^{(k)}\boldsymbol{W}_z^{(k)}+\langle\boldsymbol{v}_{ik}\rangle\langle\boldsymbol{v}_{ik}\rangle^\top, \tag{12}$$

with $\boldsymbol{S}^{(k)}=\boldsymbol{W}_z^{(k)^\top}\boldsymbol{\Sigma}_z^{(k)^{-1}}$ and $\boldsymbol{\Sigma}_z^{(k)}$ as given by (8).

2) *M-step*: To simplify the derivation of the updating equations, we define the following augmented variables,

$$\langle\tilde{\boldsymbol{v}}_{ik}\rangle=\begin{bmatrix}\langle\boldsymbol{v}_{ik}\rangle\\1\end{bmatrix}, \tag{13}$$

$$\langle\tilde{\boldsymbol{v}}_{ik}\tilde{\boldsymbol{v}}_{ik}^\top\rangle=\begin{bmatrix}\langle\boldsymbol{v}_{ik}\boldsymbol{v}_{ik}^\top\rangle & \langle\boldsymbol{v}_{ik}\rangle\\ \langle\boldsymbol{v}_{ik}\rangle^\top & 1\end{bmatrix}, \tag{14}$$

$$\tilde{\boldsymbol{\Sigma}}_{v|z}^{(k)}=\langle\tilde{\boldsymbol{v}}_{ik}\tilde{\boldsymbol{v}}_{ik}^\top\rangle-\langle\tilde{\boldsymbol{v}}_{ik}\rangle\langle\tilde{\boldsymbol{v}}_{ik}\rangle^\top. \tag{15}$$

Then, the updated MFA parameters are obtained as follows,

$$\hat{\pi}_z^{(k)} = \frac{1}{N}\sum_{i=1}^{N}\gamma_i^{(k)}, \tag{16}$$

$$\hat{\boldsymbol{\Lambda}}_z^{(k)} = \left[\sum_{i=1}^{N}\gamma_i^{(k)}\boldsymbol{z}_i\langle\tilde{\boldsymbol{v}}_i\rangle^\top\right]\left[\sum_{i=1}^{N}\gamma_i^{(k)}\langle\tilde{\boldsymbol{v}}_{ik}\tilde{\boldsymbol{v}}_{ik}^\top\rangle\right]^{-1}, \tag{17}$$

$$\hat{\boldsymbol{\Psi}}_z^{(k)} = \mathrm{diag}\left(\hat{\boldsymbol{\Lambda}}_z^{(k)}\tilde{\boldsymbol{\Sigma}}_{v|z}^{(k)}\hat{\boldsymbol{\Lambda}}_z^{(k)^T} + \frac{\sum_{i=1}^{N}\gamma_i^{(k)}\tilde{\boldsymbol{\varepsilon}}_i^{(k)}\tilde{\boldsymbol{\varepsilon}}_i^{(k)^\top}}{\sum_{i=1}^{N}\gamma_i^{(k)}}\right). \tag{18}$$

with $\tilde{\boldsymbol{\varepsilon}}_i^{(k)} = \boldsymbol{z}_i - \hat{\boldsymbol{\Lambda}}_z^{(k)}\langle\tilde{\boldsymbol{v}}_i\rangle$. The updated values for the factor loadings $\hat{\boldsymbol{W}}_z^{(k)}$ and mean vectors $\hat{\boldsymbol{\mu}}_z^{(k)}$ are obtained from the augmented factor loadings matrix $\hat{\boldsymbol{\Lambda}}_z^{(k)} = [\hat{\boldsymbol{W}}_z^{(k)}\hat{\boldsymbol{\mu}}_z^{(k)}]$.

3.2 Conversion Phase

The conversion phase involves two steps. First, the sequence of speech parameter vectors $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T)$ associated with the articulatory gesture captured by the PMA device is estimated under the probabilistic framework presented above. Then, a parametric synthesis algorithm is used to generate the final time-domain signal from the estimated speech parameters. To estimate the speech parameter vectors, we employ a frame-by-frame procedure based on the well-known Minimum Mean Square Error (MMSE) estimator:

$$\hat{\boldsymbol{y}}_t = \mathbb{E}[\boldsymbol{y}|\boldsymbol{x}_t] = \int \boldsymbol{y}p(\boldsymbol{y}|\boldsymbol{x}_t)d\boldsymbol{y}. \tag{19}$$

where $\hat{\boldsymbol{y}}_t$ is the estimate for the speech parameters at time t, $\mathbb{E}[\cdot]$ represents the expected value, and $p(\boldsymbol{y}|\boldsymbol{x}_t)$ is the speech parameter posterior distribution. This distribution is derived from the joint distribution $p(\boldsymbol{x}, \boldsymbol{y})$ in (9) as

$$p(\boldsymbol{y}|\boldsymbol{x}_t) = \sum_{k=1}^{K}P(k|\boldsymbol{x}_t)p(\boldsymbol{y}|\boldsymbol{x}_t, k), \tag{20}$$

where

$$P(k|\boldsymbol{x}_t) = \frac{\pi^{(k)}\mathcal{N}\left(\boldsymbol{x}_t; \boldsymbol{\mu}_x^{(k)}, \boldsymbol{\Sigma}_{xx}^{(k)}\right)}{\sum_{k'=1}^{K}\pi^{(k')}\mathcal{N}\left(\boldsymbol{x}_t; \boldsymbol{\mu}_x^{(k')}, \boldsymbol{\Sigma}_{xx}^{(k')}\right)}, \tag{21}$$

$$p(\boldsymbol{y}|\boldsymbol{x}_t, k) = \mathcal{N}\left(\boldsymbol{y}; \boldsymbol{\mu}_{y|x_t}^{(k)}, \boldsymbol{\Sigma}_{y|x}^{(k)}\right). \tag{22}$$

The parameters of the k-th component conditional distribution $p(\boldsymbol{y}|\boldsymbol{x}_t, k)$ are derived from those of the joint pdf $p(\boldsymbol{x}, \boldsymbol{y}|k)$ in (7). As already mentioned,

the latter distribution is Gaussian with mean $\boldsymbol{\mu}_z^{(k)}$ and covariance matrix $\boldsymbol{\Sigma}_z^{(k)}$. Then, using the standard properties of the joint Gaussian distribution, we can derive the parameters of the conditional distribution as follows,

$$\boldsymbol{\mu}_{y|x_t}^{(k)} = \boldsymbol{\mu}_y^{(k)} + \boldsymbol{\Sigma}_{yx}^{(k)} \boldsymbol{\Sigma}_{xx}^{(k)^{-1}} \left(\boldsymbol{x}_t - \boldsymbol{\mu}_x^{(k)} \right), \tag{23}$$

$$\boldsymbol{\Sigma}_{y|x}^{(k)} = \boldsymbol{\Sigma}_{yy}^{(k)} + \boldsymbol{\Sigma}_{yx}^{(k)} \boldsymbol{\Sigma}_{xx}^{(k)^{-1}} \boldsymbol{\Sigma}_{xy}^{(k)}, \tag{24}$$

where the marginal means $\boldsymbol{\mu}_x^{(k)}$, $\boldsymbol{\mu}_y^{(k)}$ and covariance matrices $\boldsymbol{\Sigma}_{xx}^{(k)}$, $\boldsymbol{\Sigma}_{yy}^{(k)}$, $\boldsymbol{\Sigma}_{xy}^{(k)}$ are obtained by partitioning $\boldsymbol{\mu}_z^{(k)}$ and $\boldsymbol{\Sigma}_z^{(k)}$ into their x and y components.

Finally, by substituting the expression of the conditional distribution $p(\boldsymbol{y}|\boldsymbol{x}_t)$ in (20) into (19), we reach the following expression for the MMSE estimation of the speech parameter vectors,

$$\begin{aligned} \hat{\boldsymbol{y}}_t &= \sum_{k=1}^{K} P(k|\boldsymbol{x}_t) \int \boldsymbol{y} p(\boldsymbol{y}|\boldsymbol{x}_t, k) d\boldsymbol{y} \\ &= \sum_{k=1}^{K} P(k|\boldsymbol{x}_t) \left(\boldsymbol{A}^{(k)} \boldsymbol{x}_t + \boldsymbol{b}^{(k)} \right), \end{aligned} \tag{25}$$

with $\boldsymbol{A}^{(k)} = \boldsymbol{\Sigma}_{yx}^{(k)} \boldsymbol{\Sigma}_{xx}^{(k)^{-1}}$ and $\boldsymbol{b}^{(k)} = \boldsymbol{\mu}_y^{(k)} - \boldsymbol{A}^{(k)} \boldsymbol{\mu}_x^{(k)}$ as can be deduced from (23).

For comparison purposes, we also evaluate a fast, approximate version of the above estimator, which we will refer to as fast MMSE (fMMSE), in which only the most likely Gaussian component k^* is involved:

$$k^* = \underset{1 \le k \le K}{\operatorname{argmax}} P(k|\boldsymbol{y}_t).$$

Then, the fMMSE estimate is defined as:

$$\hat{\boldsymbol{y}}_t \approx \boldsymbol{A}^{(k^*)} \boldsymbol{x}_t + \boldsymbol{b}^{(k^*)}. \tag{26}$$

3.3 Recalibration Procedure

The principle of the direct synthesis technique is that the mapping between articulator movement and the corresponding acoustics can be estimated from a parallel dataset containing simultaneous recordings of PMA and speech signals. Ideally, this dataset should be recorded soon after it has been agreed that a laryngectomy will be performed to the patient. During the recording session, the patient's voice and corresponding PMA data are acquired using adhesively attached magnets. In addition, the information on the location of the magnets is documented, so that they can be later surgically implanted accordingly. From the collected data the PMA-to-acoustic transformation is learned as described in Section 3.1, so it can be readily available to be used by the patient soon after the laryngectomy, as described in Section 3.2.

In certain conditions, however, the above training procedure might fail. For example, it is highly unlikely that the magnet positions can be exactly replicated during surgical implantation. Any magnet misplacement will inevitably lead to discrepancies between the PMA data used for training and the data captured during the use of the system, hence leading to the degradation of the speech quality. Furthermore, variations of the relative positions of the magnets with respect the head-frame used to hold the magnetic sensors (see Figure 1) will also lead to mismatches. Therefore, in most of the practical cases it would be necessary to recalibrate the system to compensate for any magnet misplacement with respect to their original positions used for acquiring the training data. In the following we present a data-driven recalibration procedure to this end.

We will assume that the positions of the magnets before (magnets glued) and after magnet implantation only vary slightly. In this case, the mismatch between the articulatory data captured for the same articulatory gesture pre- and post-magnet implantation can be approximately modelled as,

$$\boldsymbol{x}_t = \boldsymbol{h}(\tilde{\boldsymbol{x}}_t), \tag{27}$$

where $\boldsymbol{x}_t$ and $\tilde{\boldsymbol{x}}_t$ denote the data for the pre- and post- magnet implantation arrangements, respectively, and $\boldsymbol{h}$ is the mismatch function.

We propose the following procedure to estimate the mismatch function $\boldsymbol{h}$. First, after magnet implantation, the patient has to attend another recording session in which he/she is asked to mouth along to some of the utterances recorded during the first recording session. In this case, however, only PMA data is acquired since the patient has already lost their voice. Furthermore, only a small fraction of the data recorded during the first session needs to be recorded during the second session, as the aim of it is not to estimate the full PMA-to-acoustic mapping (as in the first recording session), but to learn the mismatch produced by the magnet misplacement. Next, as the durations of the PMA signals obtained for the same sentence in both sessions may be different, the PMA data for both recording sessions are time-aligned using the Dynamic Time Warping (DTW) technique [41]. From the time-aligned signals the mismatch function $\boldsymbol{h}$ is estimated. Here, we investigate two alternative methods for modelling this function. First, the function is represented as a simple linear mapping:

$$\boldsymbol{x}_t = \boldsymbol{C}\tilde{\boldsymbol{x}}_t + \boldsymbol{d}, \tag{28}$$

with $\boldsymbol{C}$ and $\boldsymbol{d}$ being estimated by least squares regression from the aligned data.

Alternatively, a Multilayer Perceptron (MLP) is used to model $\boldsymbol{x}_t = \boldsymbol{h}(\tilde{\boldsymbol{x}}_t)$. The input to the MLP are the PMA parameter vectors $\tilde{\boldsymbol{x}}_t$ and it tries to predict the corresponding vectors $\boldsymbol{x}_t$ used for training the MFA model. More details about the MLP architecture and its training are given in Section 5.6.

After $\boldsymbol{h}$ is estimated (either as a linear operator or a neural network), it is used in a second round of the recalibration procedure to improve the alignment of the PMA data. Thus, the PMA data recorded after magnet implantation is first compensated using the estimated transformation and then DTW-aligned with the original data (recorded with magnets glued). Next, the alignments obtained

for the compensated data are used to estimate a more accurate transformation between the PMA data captured in both sessions. This procedure is repeated several times until convergence.

4 Related Work

The direct synthesis technique presented in the previous section shares some similarities with other recently proposed methods. In this section we discuss the relationships between our proposal and those methods, pointing out the similarities and differences.

First, regarding our own previous work, we reported in [13,18,26] that not only speech recognition from PMA data is possible, but also that the performance obtained is on par with that obtained using audio, at least on isolated words and connected digits recognition tasks. This study was later successfully extended to multiple subjects in [24]. With respect to the direct synthesis approach, in [25] we carried out a feasibility study on prediction of the first two speech formants (F1 and F2) from the sensor data using a simple linear transformation. Though promising, the results showed that a more powerful transformation is required for modelling the non-linear mapping between sensor data and acoustic parameters. In [19], a more powerful conversion technique was investigated: a statistical mapping based on shared Gaussian process dynamical models [6,50], which are non-parametric models providing a shared low-dimensional embedding of the articulatory and acoustic data as well as a dynamic model in the latent space. Results reported for isolated-digit synthesis showed the superiority of this approach for modelling the PMA-to-acoustic mapping. Finally, in [21], we proposed a conversion system based on mixture of factor analysers, similar to the one described in this work, and showed the viability of voice reconstruction from PMA data for continuous speech.

In addition to ourselves, other authors have also made important contributions to the field of silent speech interfaces. A good introduction to this subject can be found in [9]. In general, approaches for direct speech synthesis from sensor data can be classified into two categories: model-based and data-driven approaches. Model-based approaches are those in which the sensor data provides interpretable information about the position of the speech articulators. From this information, the shape of the vocal tract can be recovered and speech can be synthesised by using an articulatory synthesiser. For example, in [36,37] the articulatory synthesiser is driven by Magnetic Resonance Imaging (MRI) and X-ray images, respectively, while in [45,47] EMA data is used instead. In data-driven approaches, on the other hand, the relationship between the sensor data and the acoustics is learned from parallel datasets, as with our method. Various techniques have been investigated in the past to model this relationship: Gaussian Mixture Models (GMMs) [44,38], Hidden Markov Models (HMMs) [28,27], neural networks [10,48], support vector regression [46], and a concatenative, unit-selection approach [51]. In the next section, we will compare the performance

of our MFA-based conversion technique with the well-known GMM-based technique proposed in [43,44].

5 Experimental Evaluation

In this section, we evaluate the performance of our silent speech system on a voice reconstruction task for non-impaired speakers. Although our system is thought to help laryngectomy patients to recover the voice, at this initial stage of the development our priority is to assess performance for normal speakers and then, once the system is robust, it can be tested with real patients. More details about the evaluation framework are given in the following.

5.1 Vocabulary and Data Recording

We recorded two parallel databases with different phonetic coverage to evaluate our system. The first one is based on the TIDigits speech database [34] and consists of sequences of up to seven connected English digits. The vocabulary is made up of eleven words: the digits from 'one' to 'nine' plus 'zero' and 'oh'. The second database consists of utterances selected at random from the CMU Arctic corpus of phonetically balanced sentences [32]. Parallel data was then recorded for the two databases by adult speakers with normal speaking ability. For the TIDigits database, four male speakers (M1 to M4) and a female speaker (F1) recorded 308 sentences (385 sentences for M2) comprising 7.2, 10.5, 8.0, 9.7 and 8.5 minutes of data, respectively. Speaker M1 also recorded a second dataset with 308 sentences (7.4 minutes of data) in a different recording session with the aim of evaluating the recalibration procedure described in Section 3.3. The magnet arrangement in the first recording session was documented and replicated in the second session. Despite this, as will be discussed below, small variations in the magnet positions and/or orientations unintentionally occurred. For the Arctic database, parallel data was recorded for two male speakers: M1 (same as in the TIDigit database) and M5. M1 recorded 420 utterances comprising 22 minutes of data and M5 recorded 509 sentences comprising 26 minutes.

The audio and 9-channel PMA signals were recorded simultaneously at sampling frequencies of 16 kHz and 100 Hz, respectively, using an AKG C1000S condenser microphone and the PMA device shown in Figure 1. Background cancellation was later applied to the PMA signals to mitigate the effect of the Earth's magnetic field on the articulatory data [5]. Finally, all data were endpointed in the audio domain using an energy-based algorithm to prevent modelling silence parts, as the speech articulators may adopt any position during the silence parts.

5.2 Feature Extraction

The source $\boldsymbol{x}_t$ and target $\boldsymbol{y}_t$ parameter vectors are computed as follows in the proposed system. The PMA signals are first segmented into overlapping frames

using a 20 ms analysis window with 10 ms overlap. Next, to better model contextual phonetic information, sequences of ω consecutive frames, with a single-frame displacement, are concatenated and the Partial Least Squares (PLS) technique [8] is applied to reduce the dimensionality of the resultant frames. In PLS, the number of principal components retained are those explainng the 95% of the variance in the target speech features. The audio signals are represented in this work as 25 Mel-Frequency Cepstral Coefficients (MFCCs) [16] obtained at the same frame rate as that for PMA. Neither F_0 nor voicing information are extracted from the audio signals because of the limited ability of PMA to model this aspect of speech articulation [20]. Rather, the audio signals are re-synthesised without voicing as whispered speech. Finally, the PMA and speech parameter vectors are converted to z-scores with zero mean and unit variance to improve statistical training.

5.3 Objective Evaluation of Voice Reconstruction Accuracy

In this work we use objective quality measures to evaluate the performance of the techniques under different conditions. As we have access to the original speech signals recorded by the subjects, we can compare the speech signals predicted from articulator movement with the original ones to evaluate reconstruction accuracy. In particular, we use the well-known Mel-Cepstral Distortion (MCD) measure [33] between the MFCCs extracted from the original audio signals, $\boldsymbol{c}$, and the ones predicted from PMA data, $\hat{\boldsymbol{c}}$, with smaller values indicating better results:

$$\text{MCD[dB]} = \frac{10}{\ln 10}\sqrt{2\sum_{d=1}^{D}(c_d - \hat{c}_d)^2}. \tag{29}$$

Results reported in this work are obtained using a 10-fold cross-validation scheme. Hence, the available data for each subject is randomly divided into ten sets of same length and, in each round, 9 sets are used for training and the remaining one for testing. The MCD results reported in the following sections correspond to the average MCD result for the 10 rounds.

5.4 Results on the TIDigits Database

The left and middle panels in Fig. 2 show contour plots with the average MCD results across the five subjects for the two conversion algorithms introduced in Section 3.2: MMSE and fast MMSE (fMMSE). The results are presented as a function of the number of mixture components in the MFA model (i.e. K in (20)) and the length of the sliding window used to extract the PMA parameter vectors. As can be seen, results greatly improve when more mixture components and longer windows are used for modelling the PMA-to-acoustic mapping. Using more mixtures means that this mapping, which is known to be highly non-linear [1,39], is more finely represented. For example, the mapping is approximated by a linear transformation when using 1-mixture models, while a piece-wise

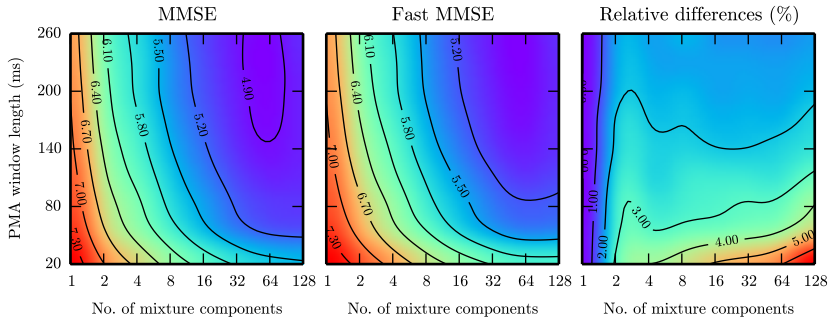

Fig. 2. Mel-cepstral distortion results on the TIDigits database. *Left and middle*: Average MCD results across all the speakers for the MMSE and fMMSE conversion systems as a function of the number of mixture components used in the MFA model and the length of the PMA-frame window. *Right*: Relative degradation of the fMMSE system with respect to the MMSE-based one.

linear approximation is employed for $K > 1$. For a PMA frame window of $\omega = 200$ ms, the relative MCD reduction when using $K = 64$ components, which is the optimum number of mixtures as can be seen in the figure, with respect to just using a single mixture is 30.33% for the MMSE system and 29.08% for the fMMSE-based one. Increasing the length of the PMA frame window is also beneficial for the mapping because it reduces its uncertainty by taking into account more contextual information about the temporal evolution of the PMA signal. In particular, for $K = 64$ mixtures, the relative reduction in MCD achieved when using a window of $\omega = 200$ ms instead of $\omega = 20$ ms is 13.52% and 17.28% for the MMSE and fMMSE systems, respectively.

The right panel in Fig. 2 shows the relative differences (expressed in percent) between the MMSE method and its fast version fMMSE. As it can be seen, both methods perform almost equally except when $K > 1$ or when short windows are used. In those cases, the performance of the fMMSE method degrades because the mapping uncertainty is higher when using short windows and, hence, it is more difficult for the fMMSE method to choose the 'correct' mixture component for performing the mapping. For example, for $K = 128$ and $\omega = 20$ ms, the fMMSE algorithm is 7.21% worse than MMSE. Conversely, the differences are almost insignificant ($\leq 2\%$ of degradation) for $K = 1$ or when long windows are used (e.g. $\omega = 200$ ms). In terms of speech intelligibility, though not formally evaluated in this work, informal listening show that speech generated by both methods is intelligible and that the speaker's voice is clearly identifiable [4].

The detailed MCD results obtained by the MMSE conversion system for each of the five subjects in the TIDigits database are shown in Fig. 3. A 64-component MFA model, which is the best model in Fig. 2, is chosen. As can be seen in Fig.

[4] Several speech samples are available in the Demos section of `http://www.hull.ac.uk/speech/disarm`

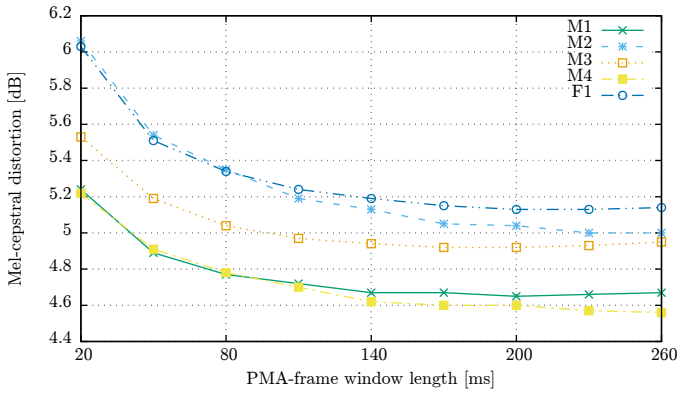

Fig. 3. Mel-ceptral distortion results obtained by the MMSE conversion system for each speaker in the TIDigits database.

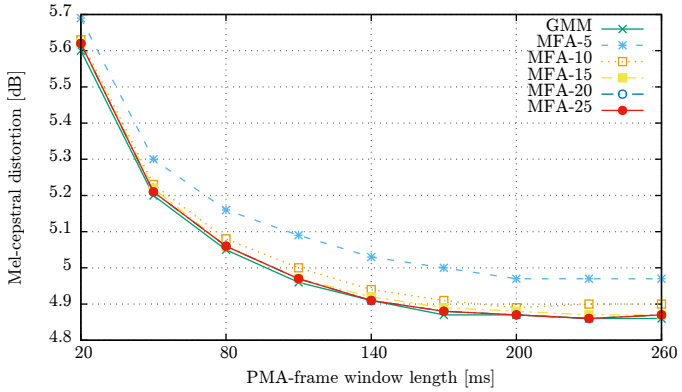

Fig. 4. Comparison between the proposed approach for articulatory-to-acoustic conversion using MFAs and Toda's et al. approach using GMMs [44]. For our proposal, the conversion accuracy using different latent space dimensions (i.e. 5, 10, 15, 20, and 25) for $\boldsymbol{v}_t$ in (5) is evaluated.

3, the best results are obtained for subjects M1 and M4 and the worst results for M2 and F1. The differences in performance among speakers can be mainly attributed to two factors: the user's experience in using the PMA device and how well the device fits her/his anatomy. In regard of user's experience, it must be pointed out that M1, M3 and M4 were proficient in the use of the PMA device, while for M2 and F1 the data recording session was also the first time they used the PMA device. With respect to the second reason, the current PMA device prototype was specifically designed for M1, so it is reasonable to think that articulatory data is more accurately captured for him than for the other subjects.

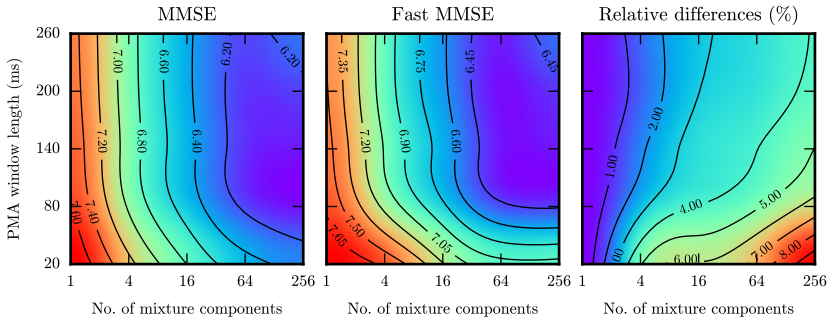

Fig. 5. Mel-cepstral distortion results on the Arctic database. *Left and middle*: Average results across all subjects for the MMSE and fMMSE systems as a function of the number of mixture components used in the MFA model and the length of the PMA-frame window. *Right*: Relative degradation of the fMMSE system with respect to the MMSE-based one.

Next, we compare our proposal with the well-known GMM-based conversion technique proposed by Toda et al. in [44,43]. For a fairer comparison, both methods are evaluated using the MMSE-based mapping algorithm. Also, we evaluate our proposal using different dimensions for the latent space variable $\boldsymbol{v}_t$ in (5). The dimensions are 5, 10, 15, 20, and 25, the latter being the dimensionality of the speech feature vectors. Results are shown in Fig. 4 for both systems. It can be seen that both methods perform almost equally except when the dimensionality of the latent space in the our system is very small (i.e. 5 or 10). In this case, the quality of synthetic speech is slightly degraded due to the inability of properly capturing the correlations between the acoustic and PMA spaces in such low-dimensional latent spaces. For dimensions greater than 15, both approaches report more or less the same results, with the benefit that our proposed approach is more computationally efficient because of the savings of carrying out the computations in the reduced-dimension space.

5.5 Results on the Arctic Database

Figure 5 shows the MCD results obtained by the MMSE and fMMSE systems on the Arctic database (left and middle contour plots), as well as the relative degradation of the fast MMSE algorithm w.r.t. the MMSE algorithm (right contour plot). It can be seen that the results for the Arctic database are not as good as the results for the TIDigit database in Figure 2. This is due to greater phonetic variability of the Arctic sentences. In fact, the Arctic corpus was designed for phonetic balance. This greater complexity results in a more complex PMA-to-acoustic mapping in the case of the Arctic sentences. Apart from that, another reason the MCD results are worse on the Arctic database is the limitations of the current PMA device for modelling some areas of the vocal

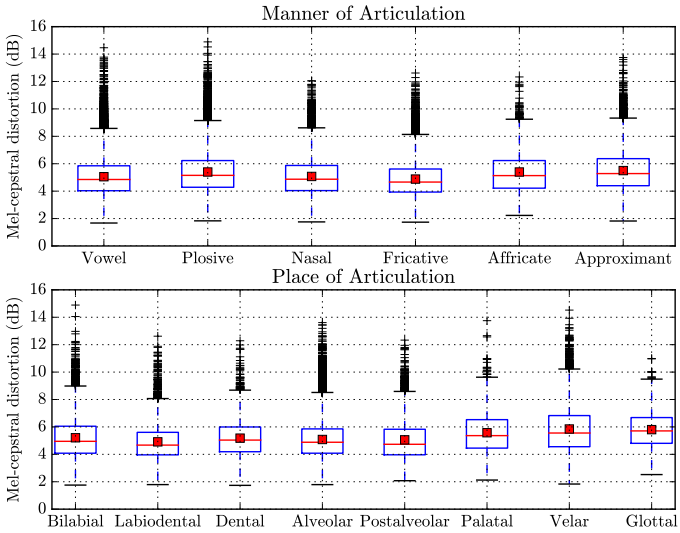

Fig. 6. Box plot of the distributions of the MCD results for different phone categories on the Arctic database. Bottom and top edges of the boxes are the first (Q1) and third (Q3) quartiles of the data. Red bands inside the boxes represents the median, while the means are represented with small, red boxes. Whiskers extend up to 1.5 times the interquartile range (i.e. Q3-Q1) and the outliers are plotted with black crosses. Phone categories are those in the IPA chart [30]. *Top*: Results when considering the manner of articulation of the phones. *Bottom*: Results for the place of articulation.

tract (e.g. sounds articulated at the back of the mouth) [20,21], as discussed below. Since the Arctic sentences contain more phones articulated in those areas than the digits vocabulary, this harms the overall reconstruction performance achieved on the Arctic database. Regarding the two conversion algorithms, the MMSE-based system outperforms its fast version (fMMSE) again. Nevertheless, the differences between both systems are small, particularly when both high number of mixtures and long windows are used for the mapping. The best overall results are obtained using 256 mixtures in the MFA model and an analysis window spaning 80 ms. This is configuration used in the rest of this section.

To investigate in detail the performance of our system, we conduct a second analysis at the phone level in which the distortions are computed for each phone and the resultant distributions are represented as a box plot. For the sake of clarity, the MCD results of the phones sharing similar articulation properties are grouped together rather than presenting the results for each individual phone. Here, phones are grouped according to their manner and place of articulation. For segmenting the speech signals into phones, we force-aligned their word-level transcriptions using a cross-word, triphone-based speech recogniser adapted to each subject. The phone-level transcriptions with timing information provided by the forced-alignment procedure are then used to segment the original and

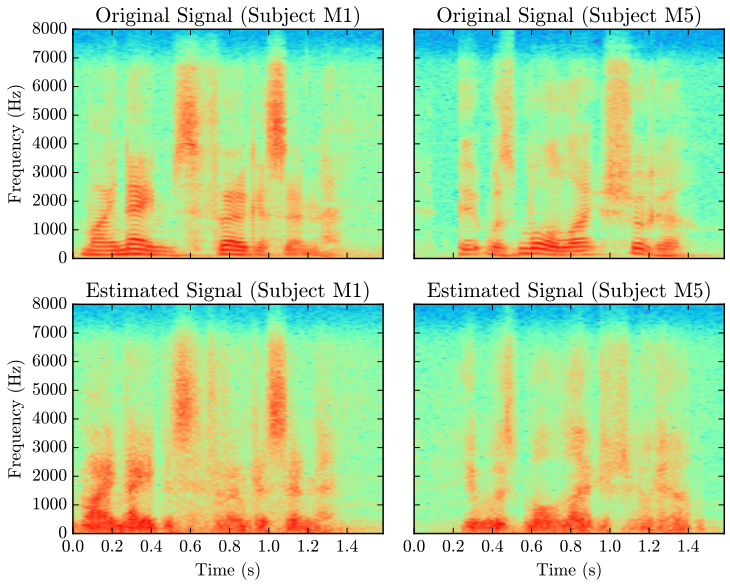

Fig. 7. *Top-left and bottom-left*: Spectrograms of natural speech (top), and speech estimated from PMA data (bottom) corresponding to the sentence "My name is Ferguson" spoken by the subject M1. *Top-right and right-left*: Spectrograms of the sentence "It was more like sugar" spoken by M5.

estimated speech signals. The results of this second analysis are shown in Figure 6. When considering the manner of articulation, we can see that the plosive, affricate and approximant consonants tend to be synthesised less accurately than other sound classes due to their more complex articulation and dynamics. When considering the place of articulation, it can seen that sounds articulated at the middle and back of the mouth (palatal, velar, and glottal consonants) are, on average, less well reconstructed than other phones. This is due to the limitations of the current PMA prototype for modelling those areas of the vocal tract [20,21].

Finally, in Figure 7, a comparison between the spectrograms computed from natural speech and from the speech synthesised by our system is presented for the two subjects in the Arctic database. As can be seen, our system is able to predict with high level of accuracy the speech formants and, in general, the spectral envelope of the signals. Spectral detail, however, is lost due to over-smoothing when training the MFA models [52] and the limited information provided by PMA about the articulation process. It also can be seen that the estimated signals are synthesised with no voicing. This is also due to the limited ability of the current PMA prototype for capturing the movement of the vocal folds.

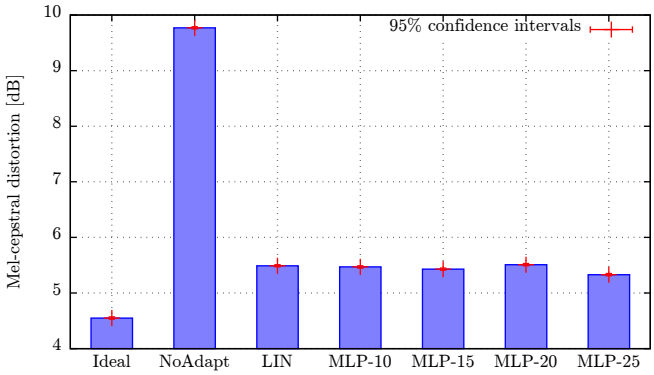

Fig. 8. Cross-session synthesis results: MCD results obtained when synthesising PMA data from Session 2 in the TIDigits database using a MFA model trained on the Session 1 dataset.

5.6 Cross-session synthesis results

So far it has been assumed that there is no mismatch between the data used for training and that used for testing. However, as already discussed in Section 3.3, this is not always true. Variations in the positions of the magnets pre- and post- implantation as well as variations in the relative position of magnets with respect to the head-frame used to hold the magnetic sensors (see Figure 1), will inevitably lead to mismatches that will degrade the quality of speech synthesised from sensor data. In this section, we evaluate the performance of the direct synthesis technique in one scenario which introduces such mismatch: speech is synthesised from PMA data recorded by the speaker M1 in his second recording session (Session 2) using a MFA model trained on parallel data from his first session (Session 1).

Figure 8 shows the MCD results obtained for the above experiment when a 64-component MFA model and a PMA-frame window of 200 ms are used. In the figure, Ideal refers to the ideal case in which there is no mismatch between training and testing (i.e. parallel data from Session 2 is used for training and testing within the cross-validation scheme), the NoAdapt system directly convert the sensor data from Session 2 using the model trained on data from Session 1 with no compensation, and the remaining results are for the compensation technique proposed in Section 3.3: LIN models the mismatch function as a linear transformation, while MLP uses a multilayer perceptron with 10, 15, 20 and 25 sigmoid units in the hidden layer.

In the figure, the best results are obtained in the Ideal case where there is no mismatch between training and testing. Even though magnet placement was documented to avoid misplacement between sessions, we see from the NoAdapt results that even small changes between sessions are catastrophic in terms of the synthesised speech quality. This is greatly alleviated, however, by the proposed compensation technique. In this case, the results are only slightly worse than

the result obtained in the ideal case. Regarding the different approaches for mismatch compensation, it can be seen that the best results are obtained using a MLP with 25 hidden units due to the greater modelling flexibility allowed by this model. Nevertheless, a simple linear transformation (LIN) also achieves very similar results to MLP-25 with the benefit of LIN being more computationally efficient.

6 Conclusions

In this chapter we have described a system for synthesising speech from motion data captured from the lips and tongue using magnetic sensing. Preliminary evaluation of the system via objective performance metrics show that the proposed system is able to generate speech of sufficient quality for some vocabularies. However, problems still remain to scale up the system to work consistently for phonetically rich vocabularies. It has also been reported that one of the current limitations of our sensing technique, that is, the differences between the articulatory data captured in different sessions, can be greatly reduced by applying a pre-processing technique to the sensor data before the conversion. This brings us closer to being able to apply our voice reconstruction system in a realistic treatment scenario. These results encourage us in pursuing our goal of developing a system that will ultimately allow laryngectomy patients to recover their voice. In order to reach this point, a number of questions will need to addressed in future research such as improving the performance for phonetically rich vocabularies, ways of predicting the prosodic information from PMA data, and extending the technique to impaired subjects.

7 Acknowledgements

This is a summary of independent research funded by the National Institute for Health Research (NIHR)'s Invention for Innovation Programme. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

References

1. Atal, B.S., Chang, J.J., Mathews, M.V., Tukey, J.W.: Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique. J. Ac. Soc. Am. 63(5), 1535–1555 (1978)
2. Braz, D.S.A., Ribas, M.M., Dedivitis, R.A., Nishimoto, I.N., Barros, A.P.B.: Quality of life and depression in patients undergoing total and partial laryngectomy. Clinics 60(2), 135–142 (2005)
3. Byrne, A., Walsh, M., Farrelly, M., O'Driscoll, K.: Depression following laryngectomy. A pilot study. Brit J Psychiat. 163(2), 173–176 (1993)
4. Cheah, L.A., Bai, J., Gonzalez, J.A., Gilbert, J.M., Ell, S.R., Green, P.D., Moore, R.K.: Preliminary evaluation of a silent speech interface based on intra-oral magnetic sensing. In: Proc. BioDevices. pp. 108–116 (2016)
5. Cheah, L.A., Bai, J., Gonzalez, J.A., Ell, S.R., Gilbert, J.M., Moore, R.K., Green, P.D.: A user-centric design of permanent magnetic articulography based assistive speech technology. In: Proc. BioSignals. pp. 109–116 (2015)
6. Chen, J., Kim, M., Wang, Y., Ji, Q.: Switching Gaussian process dynamic models for simultaneous composite motion tracking and recognition. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition. pp. 2655–2662 (2009)
7. Danker, H., Wollbrück, D., Singer, S., Fuchs, M., Brähler, E., Meyer, A.: Social withdrawal after laryngectomy. Eur Arch Oto-Rhino-L 267(4), 593–600 (2010)
8. De Jong, S.: SIMPLS: an alternative approach to partial least squares regression. Chemometrics Intell. Lab. Syst. 18(3), 251–263 (1993)
9. Denby, B., Schultz, T., Honda, K., Hueber, T., Gilbert, J., Brumberg, J.: Silent speech interfaces. Speech Commun. 52(4), 270–287 (Apr 2010)
10. Desai, S., Raghavendra, E.V., Yegnanarayana, B., Black, A.W., Prahallad, K.: Voice conversion using artificial neural networks. In: Proc. ICASSP. pp. 3893–3896 (2009)
11. Ell, S.R.: Candida: the cancer of silastic. J Laryngol Otol 110(03), 240–242 (1996)
12. Ell, S.R., Mitchell, A.J., Parker, A.J.: Microbial colonization of the groningen speaking valve and its relationship to valve failure. Clin Otolaryngol Allied Sci. 20(6), 555–556 (1995)
13. Fagan, M.J., Ell, S.R., Gilbert, J.M., Sarrazin, E., Chapman, P.M.: Development of a (silent) speech recognition system for patients following laryngectomy. Med Eng Phys. 30(4), 419–425 (2008)
14. Freitas, J., Teixeira, A., Bastos, C., Dias, M.: Speech Technologies, vol. 10, chap. Towards a multimodal silent speech interface for European Portuguese, pp. 125–150. InTech (2011)
15. Fried-Oken, M., Fox, L., Rau, M.T., Tullman, J., Baker, G., Hindal, M., Wile, N., Lou, J.S.: Purposes of AAC device use for persons with ALS as reported by caregivers. Augment Altern Commun. 22(3), 209–221 (2006)
16. Fukada, T., Tokuda, K., Kobayashi, T., Imai, S.: An adaptive algorithm for Mel-cepstral analysis of speech. In: Proc. ICASSP. pp. 137–140 (1992)
17. Ghahramani, Z., Hinton, G.E.: The EM algorithm for mixtures of factor analyzers. Tech. Rep. CRG-TR-96-1, University of Toronto (1996)
18. Gilbert, J.M., Rybchenko, S.I., Hofe, R., Ell, S.R., Fagan, M.J., Moore, R.K., Green, P.: Isolated word recognition of silent speech using magnetic implants and sensors. Med Eng Phys. 32(10), 1189–1197 (2010)
19. Gonzalez, J.A., Green, P.D., Moore, R.K., Cheah, L.A., Gilbert, J.M.: A non-parametric articulatory-to-acoustic conversion system for silent speech using shared gaussian process dynamical models. In: UK Speech. p. 11 (2015)

20. Gonzalez, J.A., Cheah, L.A., Bai, J., Ell, S.R., Gilbert, J.M., 1, R.K.M., Green, P.D.: Analysis of phonetic similarity in a silent speech interface based on permanent magnetic articulography. In: Proc. Interspeech. pp. 1018–1022 (2014)
21. Gonzalez, J.A., Cheah, L.A., Gilbert, J.M., Bai, J., Ell, S.R., Green, P.D., Moore, R.K.: A silent speech system based on permanent magnet articulography and direct synthesis. Comput Speech Lang. 39, 67–87 (2016)
22. Heaton, J.M., Parker, A.J.: Indwelling tracheo-oesophageal voice prostheses post-laryngectomy in sheffield, uk: a 6-year review. Acta Otolaryngol. 114(6), 675–678 (1994)
23. Herff, C., Heger, D., de Pesters, A., Telaar, D., Brunner, P., Schalk, G., Schultz, T.: Brain-to-text: decoding spoken phrases from phone representations in the brain. Frontiers in Neuroscience 9(217) (Jun 2015)
24. Hofe, R., Bai, J., Cheah, L.A., Ell, S.R., Gilbert, J.M., Moore, R.K., Green, P.D.: Performance of the MVOCA silent speech interface across multiple speakers. In: Proc. Interspeech. pp. 1140–1143 (2013)
25. Hofe, R., Ell, S.R., Fagan, M.J., Gilbert, J.M., Green, P.D., Moore, R.K., Rybchenko, S.I.: Speech synthesis parameter generation for the assistive silent speech interface MVOCA. In: Proc. Interspeech. pp. 3009–3012 (2011)
26. Hofe, R., Ell, S.R., Fagan, M.J., Gilbert, J.M., Green, P.D., Moore, R.K., Rybchenko, S.I.: Small-vocabulary speech recognition using a silent speech interface based on magnetic sensing. Speech Commun. 55(1), 22–32 (2013)
27. Hueber, T., Bailly, G.: Statistical conversion of silent articulation into audible speech using full-covariance HMM. Med Eng Phys. 36, 274–293 (2016)
28. Hueber, T., Bailly, G., Denby, B.: Continuous articulatory-to-acoustic mapping using phone-baseda trajectory HMM for a silent speech interface. In: Proc. Interspeech. pp. 723–726 (2012)
29. Hueber, T., Benaroya, E.L., Chollet, G., Denby, B., Dreyfus, G., Stone, M.: Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips. Speech Commun. 52(4), 288–300 (2010)
30. International Phonetic Association: The international phonetic alphabet (2005)
31. Jou, S.C., Schultz, T., Walliczek, M., Kraft, F., Waibel, A.: Towards continuous speech recognition using surface electromyography. In: Proc. Interspeech. pp. 573–576 (2006)
32. Kominek, J., Black, A.W.: The CMU Arctic speech databases. In: Fifth ISCA Workshop on Speech Synthesis. pp. 223–224 (2004)
33. Kubichek, R.: Mel-cepstral distance measure for objective speech quality assessment. In: Proc. IEEE Pacific Rim Conference on Communications, Computers and Signal Processing. pp. 125–128 (1993)
34. Leonard, R.: A database for speaker-independent digit recognition. In: Proc. ICASSP. pp. 328–331 (1984)
35. Maeda, S.: A digital simulation method of the vocal-tract system. Speech Commun. 1(3), 199–229 (1982)
36. Mullen, J., Howard, D.M., Murphy, D.T.: Waveguide physical modeling of vocal tract acoustics: flexible formant bandwidth control from increased model dimensionality. IEEE Trans. Audio Speech Lang. Process. 14(3), 964–971 (2006)
37. Murphy, D.T., Jani, M., Ternström, S.: Articulatory vocal tract syntheis in supercollider. In: Proc. Int. Conference on Digital Audio Effects. pp. 1–7 (2015)
38. Nakamura, K., Toda, T., Saruwatari, H., Shikano, K.: Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech. Speech Commun. 54(1), 134–146 (Jan 2012)

39. Neiberg, D., Ananthakrishnan, G., Engwall, O.: The acoustic to articulation mapping: non-linear or non-unique? In: Proc. Interspeech. pp. 1485–1488 (2008)
40. Petajan, E.D.: Automatic lipreading to enhance speech recognition (speech reading). Ph.D. thesis, University of Illinois at Urbana-Champaign (1984)
41. Sakoe, H., Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans. Acoust., Speech, Signal Processing 26(1), 43–49 (1978)
42. Schultz, T., Wand, M.: Modeling coarticulation in EMG-based continuous speech recognition. Speech Commun. 52(4), 341–353 (Apr 2010)
43. Toda, T., Black, A.W., Tokuda, K.: Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. IEEE Trans. Audio Speech Lang. Process. 15(8), 2222–2235 (Nov 2007)
44. Toda, T., Black, A.W., Tokuda, K.: Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model. Speech Commun. 50(3), 215–227 (Mar 2008)
45. Toutios, A., Maeda, S.: Articulatory VCV synthesis from EMA data. In: Proc. Interspeech (2012)
46. Toutios, A., Margaritis, K.G.: A support vector approach to the acoustic-to-articulatory mapping. In: Proc. Interspeech. pp. 3221–3224 (2005)
47. Toutios, A., Narayanan, S.: Articulatory synthesis of french connected speech from EMA data. In: Proc. Interspeech. pp. 2738–2742 (2013)
48. Uria, B., Renals, S., Richmond, K.: A deep neural network for acoustic-articulatory speech inversion. In: Proc. NIPS 2011 Workshop on Deep Learning and Unsupervised Feature Learning (2011)
49. Wand, M., Janke, M., Schultz, T.: Tackling speaking mode varieties in EMG-based speech recognition. IEEE Trans. Bio-Med. Eng. 61(10), 2515–2526 (Oct 2014)
50. Wang, J.M., Fleet, D.J., Hertzmann, A.: Gaussian process dynamical models for human motion. IEEE Trans. Pattern Anal. Mach. Intell. 30(2), 283–298 (2008)
51. Zahner, M., Janke, M., Wand, M., Schultz, T.: Conversion from facial myoelectric signals to speech: a unit selection approach. In: Proc. Interspeech. pp. 1184–1188 (2014)
52. Zen, H., Tokuda, K., Black, A.W.: Statistical parametric speech synthesis. Speech Commun. 51(11), 1039–1064 (Nov 2009)