Direct Speech Generation for a Silent Speech Interface based on Permanent Magnet Articulography

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.*

[2]*School of Engineering, University of Hull, Kingston upon Hull, U.K.*

[3]*Hull and East Yorkshire Hospitals Trust, Castle Hill Hospital, Cottingham, U.K.*

Keywords: Silent Speech Interfaces, Speech Rehabilitation, Speech Synthesis and Permanent Magnet Articulography.

Abstract: Patients with larynx cancer often lose their voice following total laryngectomy. Current methods for post-laryngectomy voice restoration are all unsatisfactory due to different reasons: requires frequent replacement due to biofilm growth (tracheo-oesoephageal valve), speech sounds gruff and masculine (oesophageal speech) or robotic (electro-larynx) and, in general, are difficult to master (oesophageal speech and electro-larynx). In this work we investigate an alternative approach for voice restoration in which speech articulator movement is converted into audible speech using a speaker-dependent transformation learned from simultaneous recordings of articulatory and audio signals. To capture articulator movement, small magnets are attached to the speech articulators and the magnetic field generated while the user 'mouths' words is captured by a set of sensors. Parallel data comprising articulatory and acoustic signals recorded before laryngectomy are used to learn the mapping between the articulatory and acoustic domains, which is represented in this work as a mixture of factor analysers. After laryngectomy, the learned transformation is used to restore the patient's voice by transforming the captured articulator movement into an audible speech signal. Results reported for normal speakers show that the proposed system is very promising.

SCIENCE AND TECHNOLOGY PUBLICATIONS

1 INTRODUCTION

Every year thousands of people worldwide have their larynx surgically removed because of throat cancer, trauma or destructive throat infection (Fagan et al., 2008; Wang et al., 2012). As speech is seen as a vital part of human communication, post-laryngectomy patients who have lost their voices often find themselves struggling with their daily communication, which can lead to social isolation, feelings of loss of identity and depression (Byrne et al., 1993; Braz et al., 2005; Danker et al., 2010). Unfortunately, the quality of voice generated by conventional post-laryngectomy restoration methods, such as oesophageal speech, tracheo-oesoephageal speech or the electro-larynx, is poor and often these methods are difficult to master (Fagan et al., 2008; Hueber et al., 2010). Augmentative and alternative communication (AAC) devices, on the other hand, are also limited by their slow manual text input.

The use of silent speech interfaces (SSIs) (Denby et al., 2010) provides an alternative solution to the conventional methods by enabling oral communication in the absence of audible speech by exploiting other non-audible signals generated during speech, such as electrical activity in the brain (Herff et al., 2015) or in the articulator muscles (Jou et al., 2006; Schultz and Wand, 2010; Wand et al., 2014) or the movement of the speech articulators themselves (Petajan, 1984; Toda et al., 2008; Denby et al., 2010; Hueber et al., 2010; Gilbert et al., 2010; Freitas et al., 2011; Hofe et al., 2013). Because of this unique feature, SSIs can be suitable for applications other than post-laryngectomy voice rehabilitation, such as communication in noisy environments or in situations where privacy/confidentiality is important.

The present work makes use of permanent magnet articulography (PMA) (Fagan et al., 2008; Gilbert et al., 2010), which is a sensing technique for articulator motion capture. In PMA a set of magnets are attached to the intact articulators and the variations of the resultant magnetic field generated while the user 'mouths' words are captured by a number of sensors located around the mouth. In previous work (Fagan

Gonzalez, J., Cheah, L., Gilbert, J., Bai, J., Ell, S., Green, P. and Moore, R.
Direct Speech Generation for a Silent Speech Interface based on Permanent Magnet Articulography.
DOI: 10.5220/0005754100960105
In *Proceedings of the 9th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2016) - Volume 4: BIOSIGNALS*, pages 96-105
ISBN: 978-989-758-170-0
Copyright © 2016 by SCITEPRESS – Science and Technology Publications, Lda. All rights reserved

et al., 2008; Gilbert et al., 2010; Hofe et al., 2013; Cheah et al., 2015) it has been shown that generation of audible speech from PMA data can be achieved by first decoding the acquired articulatory data using an automatic speech recognition (ASR) system trained on PMA data and then synthesising the recognised text using a text-to-speech (TTS) synthesiser. This approach, however, has limitations that may affect the user's willingness to engage in social interactions: speech articulation and its corresponding auditory feedback are disconnected due to the variable delay introduced by ASR and TTS and the system is constrained to the language and vocabulary of the ASR system being used. Furthermore, the non-linguistic information embedded in the articulatory signal, such as emotion or speaker identity, is normally lost after ASR.

To address the shortcomings of the recognise-then-synthesise approach, an alternative approach is investigated in this paper: direct conversion of the PMA data stream into an audible speech signal without an intermediate recognition step. In this approach, which will be referred to as the direct speech synthesis approach, a transformation is applied to the acquired PMA data to obtain a sequence of speech parameter vectors from which an acoustic signal is synthesised. To enable this method, we adopt a statistical approach in which simultaneous recordings of PMA and audio data are used to learn the mapping between the articulatory and acoustic domains. These parallel recordings are used during an initial training phase to estimate the joint distribution of PMA and speech parameter vectors. Then, in the conversion phase, the speech-parameter posterior distribution given the sensor data is computed so that an acoustic signal can be recovered from the captured PMA data. Because the PMA-to-acoustic transformation is learned using the patient's own voice, the proposed method has the potential to synthesise speech that sounds as the original voice. Furthermore, if the conversion can be done in near real-time, the synthesised voice will also sound spontaneous and natural and will allow the patient to receive real-time auditory feedback of her/his articulatory gestures.

Although this is the first work reported which implements a PMA-based SSI using the direct synthesis approach, other authors have also addressed in the past similar problems using different sensing technologies. For example, Toda et al. proposed in (Toda et al., 2008; Toda et al., 2007) a related technique to our proposal for both articulatory-to-acoustic and acoustic-to-articulatory conversion using Gaussian mixture models (GMMs), where the articulatory data was captured using electromagnetic articulography (EMA). More recently, Toda's technique has been also applied to other sensing technologies: non-audible murmur (NAM) (Toda et al., 2012), video and ultrasound (Hueber et al., 2011) and radar (Toth et al., 2010). A different approach for generating speech from articulatory data is that of (Zahner et al., 2014), in which a concatenative, unit-selection approach is employed to generate speech from surface electromiography (sEMG) data. Recently, deep neural networks (DNNs) have also been applied to both acoustic-to-articulatory (Uria et al., 2011) and voice conversion (Chen et al., 2014) problems with very promising results.

The rest of this paper is organised as follows. First, in Section 2, the PMA technique is briefly outlined. Then, in Section 3, the proposed technique for speech generation from PMA data is described. Section 4 discusses some practical implementation issues. In Section 5, direct synthesis is evaluated on parallel databases containing PMA and acoustic data. Finally, we summarise this paper and outline future work in Section 6.

2 PERMANENT MAGNET ARTICULOGRAPHY

PMA is a technique for capturing speech articulator motion by attaching a set of magnets to the articulators (typically the lips and tongue) and measuring the resultant magnetic field changes with sensors close to the articulators (see Fig. 1). The variations of the magnetic field may then be used to determine the speech which the user wishes to produce, either by performing ASR on the PMA data (Gilbert et al., 2010; Hofe et al., 2013) or by transforming the articulatory data to an acoustic signal as we do in this paper. It should be noted that contrary to other mechanisms for articulator motion capture, PMA does not provide the exact position of the individual magnets as the magnetic field detected by each sensor is a composite of the fields generated by all the magnets.

As shown in Fig. 1, six magnets are used in the current PMA device prototype for detecting the movement of the articulators: four on the lips with dimensions 1 mm (diameter) $\times$ 5 mm (height), one on the tongue tip (2 mm $\times$ 4 mm), and one on the middle of the tongue (5 mm $\times$ 1 mm). The magnetic field generated by the magnets when the user 'speaks' is then recorded by four triaxial magnetic sensors mounted on a rigid frame, each one providing three channels of data for the (x,y,z) spatial components of the magnetic field at the sensor location. Only the three sensors that are closest to the mouth are actually used for

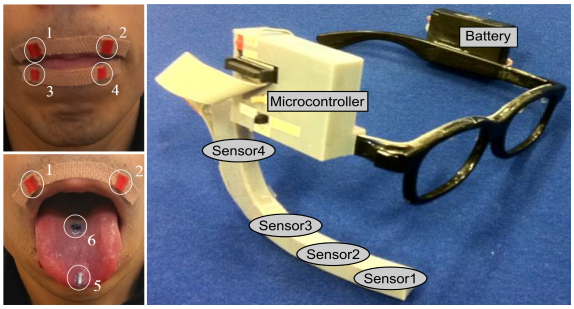

Figure 1: *Upper-left and lower-left*: magnet positioning in the current PMA device. *Right*: components of the PMA headset: microcontroller, battery and magnetic sensors used to detect the variations of the magnetic field generated by the magnets.

capturing articulatory data, while Sensor4 in Fig. 1 is used for cancelling the effects of Earth's magnetic field in the articulatory data.

3 SPEECH GENERATION FROM PMA DATA

In this section we present the details of the PMA-to-acoustic conversion technique. Let us denote by $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ the PMA and speech parameter vectors at frame t, respectively. In this work the source vectors $\boldsymbol{x}_t$ are derived from the signal captured by the PMA device, whereas the target vectors $\boldsymbol{y}_t$ correspond to a parametric representation of the audio signal (e.g. Mel-frequency cepstral coefficients). From these definitions, the aim of the proposed technique is to model the mapping $\boldsymbol{y}_t = \boldsymbol{f}(\boldsymbol{x}_t)$. Here, we employ a statistical approach in which the parameters of the mapping are learned from parallel recordings of PMA and acoustic data. The parallel data is used during an initial training phase to learn the joint distribution $p(\boldsymbol{x},\boldsymbol{y})$, which is modelled as a mixture of factor analysers (MFA) (Ghahramani and Hinton, 1996). Then, in the conversion phase, the learned transformation is used to convert PMA parameter vectors into speech parameter ones. As we show below, this involves finding the conditional distribution $p(\boldsymbol{y}|\boldsymbol{x}_t)$. The training and conversion phases are described in more detail below.

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 compact 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 piecewise 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,

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

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

where $k = 1,\ldots,K$ is the model index, $\boldsymbol{W}_x^{(k)}$ and $\boldsymbol{W}_y^{(k)}$ are the factor loadings matrices, and $\boldsymbol{\mu}_x^{(k)}$ and $\boldsymbol{\mu}_y^{(k)}$ are bias vectors that allow the data to have a non-zero mean. This model can be written in a compact form as,

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

where $\boldsymbol{z}_t = [\boldsymbol{x}_t^\top, \boldsymbol{y}_t^\top]^\top$, $\boldsymbol{W}_z^{(k)} = [\boldsymbol{W}_x^{(k)^\top}\boldsymbol{W}_y^{(k)^\top}]^\top$, $\boldsymbol{\mu}_z^{(k)} = [\boldsymbol{\mu}_x^{(k)^\top}, \boldsymbol{\mu}_y^{(k)^\top}]^\top$, 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 $\boldsymbol{\Sigma}_z^{(k)} = \boldsymbol{\Psi}_z^{(k)} + \boldsymbol{W}_z^{(k)}\boldsymbol{W}_z^{(k)^\top}$ is the reduced-rank covariance matrix.

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

$$p(\boldsymbol{z}) = \sum_{k=1}^{K} \pi^{(k)} p(\boldsymbol{z}|k). \tag{8}$$

Finally, the expectation-maximization (EM) algorithm proposed in (Ghahramani and Hinton, 1996) is used in this paper to estimate the parameters of the MFA model $\{\langle \pi^{(k)}, \boldsymbol{\mu}_z^{(k)}, \boldsymbol{W}_z^{(k)}, \boldsymbol{\Psi}_z^{(k)} \rangle, k = 1, \ldots, K\}$ from a training dataset consisting of pairs of source and target vectors $\left\{ \boldsymbol{z}_i = [\boldsymbol{x}_i^\top, \boldsymbol{y}_i^\top]^\top, i = 1, \ldots, N \right\}$.

3.2 Conversion Phase

In the conversion phase the joint distribution $p(\boldsymbol{x}, \boldsymbol{y})$ is used to estimate the sequence of speech parameter vectors associated with the articulatory data captured by the PMA device. Then, the final time-domain acoustic signal is synthesised from the estimated speech parameters by using the corresponding vocoder. Here, a frame-by-frame procedure based on the well-known minimum mean square error (MMSE) estimator is used to represent the mapping:

$$\hat{\boldsymbol{y}}_t = \mathbb{E}[\boldsymbol{y}|\boldsymbol{x}_t] = \int \boldsymbol{y} p(\boldsymbol{y}|\boldsymbol{x}_t) d\boldsymbol{y}. \tag{9}$$

From the expression of $p(\boldsymbol{x}, \boldsymbol{y})$ in (8), it can be deduced that the posterior distribution $p(\boldsymbol{y}|\boldsymbol{x}_t)$ is given by,

$$p(\boldsymbol{y}|\boldsymbol{x}_t) = \sum_{k=1}^{K} P(k|\boldsymbol{x}_t) p(\boldsymbol{y}|\boldsymbol{x}_t, k), \tag{10}$$

where

$$P(k|\boldsymbol{x}_t) = \frac{\pi^{(k)} \mathcal{N}\left(\boldsymbol{x}_t; \boldsymbol{\mu}_x^{(k)}, \boldsymbol{\Sigma}_{xx}^{(k)}\right)}{\sum_{k'=1}^{K} \pi^{(k')} \mathcal{N}\left(\boldsymbol{x}_t; \boldsymbol{\mu}_x^{(k')}, \boldsymbol{\Sigma}_{xx}^{(k')}\right)}, \tag{11}$$

$$p(\boldsymbol{y}|\boldsymbol{x}_t, k) = \mathcal{N}\left(\boldsymbol{y}; \boldsymbol{\mu}_{y|x_t}^{(k)}, \boldsymbol{\Sigma}_{y|x}^{(k)}\right). \tag{12}$$

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,

$$\boldsymbol{\mu}_{y|x_t}^{(k)} = \boldsymbol{\mu}_y^{(k)} + \boldsymbol{\Sigma}_{yx}^{(k)} \boldsymbol{\Sigma}_{xx}^{(k)^{-1}} \left(\boldsymbol{x}_t - \boldsymbol{\mu}_x^{(k)}\right), \tag{13}$$

$$\boldsymbol{\Sigma}_{y|x}^{(k)} = \boldsymbol{\Sigma}_{yy}^{(k)} + \boldsymbol{\Sigma}_{yx}^{(k)} \boldsymbol{\Sigma}_{xx}^{(k)^{-1}} \boldsymbol{\Sigma}_{xy}^{(k)}, \tag{14}$$

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 (10) into (9), we reach the following expression for the MMSE estimation of the speech parameter vectors,

$$\begin{aligned} \hat{\boldsymbol{y}}_t &= \sum_{k=1}^{K} P(k|\boldsymbol{x}_t) \int \boldsymbol{y} p(\boldsymbol{y}|\boldsymbol{x}_t, k) d\boldsymbol{y} \\ &= \sum_{k=1}^{K} P(k|\boldsymbol{x}_t) \left(\boldsymbol{A}^{(k)} \boldsymbol{x}_t + \boldsymbol{b}^{(k)}\right), \end{aligned} \tag{15}$$

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 (13).

4 PRACTICAL IMPLEMENTATION

The underlying principle of the proposed technique for direct speech synthesis is that the patient should attend a recording session soon after it has been agreed that a laryngectomy will be performed. During the session, the patient's voice and corresponding PMA data will be recorded using adhesively attached magnets. In addition, the information on the location of the glued magnets during the session will be documented, so that they can then be later surgically implanted accordingly. From the collected data the PMA-to-acoustic mapping is then estimated, so it can be readily available to be used by the patient soon after the laryngectomy.

In certain conditions, however, the above training procedure might fail. For example, it is highly unlikely that the exact magnet positions can be replicated during surgical implantation. Any misplacement of the magnets will inevitably lead to mismatches with respect to the data used for training the MFA model, thus leading to degradation in the quality of synthesised speech. 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 re-calibrate the system in order to compensate for any magnet misplacement with respect to their original positions used for acquiring the training data. In the following we propose a feature-space compensation approach to this end.

We will assume that the positions of the magnets pre- (magnets glued) and post-operation (magnets implanted) only vary slightly, so that the mismatch between the articulatory data captured for the same articulatory gesture pre- and post-operation can be modelled as,

$$\boldsymbol{x}_t = \boldsymbol{h}(\tilde{\boldsymbol{x}}_t), \tag{16}$$

where $\boldsymbol{x}_t$ and $\tilde{\boldsymbol{x}}_t$ denote the PMA data obtained using the pre-operation and post-operation magnet arrangement, respectively, and $\boldsymbol{h}$ is the mismatch function.

To estimate $\boldsymbol{h}$, the following procedure is applied. First, after magnet implantation, the patient has to attend another recording session in which he or 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, the PMA data for both recording sessions are aligned using dynamic time warping (DTW) and used to estimate the mismatch function. Two different alternative methods are proposed in this paper for modelling this function. First, a simple linear mapping is used to model the mismatch, i.e.

$$\boldsymbol{x}_t = \boldsymbol{A}\tilde{\boldsymbol{x}}_t + \boldsymbol{b}, \tag{17}$$

with $\boldsymbol{A}$ and $\boldsymbol{b}$ 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 vectors $\tilde{\boldsymbol{x}}_t$ and it tries to predict the corresponding PMA 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 adaptation procedure to improve the alignment of the PMA data captured in both sessions. Thus, the adaptation data is first compensated using the estimated transformation and then DTW-aligned with the original data. 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.

5 EXPERIMENTAL EVALUATION

The proposed direct synthesis technique is evaluated here only for normal speakers. Despite the fact that our ultimate goal is to use this technique for voice restoration after laryngectomy, we believe that at this initial stage of the development our priority is to assess voice reconstruction accuracy for normal speakers and then, once the system is robust, it can be tested with those whose voice has been altered by disease. More details about the evaluation framework are given in the following.

5.1 Vocabulary and Data Recording

To evaluate the proposed PMA-to-acoustic conversion technique, two parallel databases with different phonetic coverage were recorded. The first one is based on the TIDigits speech database (Leonard, 1984) 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 (Kominek and Black, 2004). 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 (different day) with the aim of evaluating the recalibration procedure proposed in Section 4. The magnet arrangement used during 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, 420 utterances were recorded by speaker M1 making a total of 22 minutes of data.

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 in order to mitigate the effect of the Earth's magnetic field on the articulatory data. 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 Signal Processing

In the proposed technique source $\boldsymbol{x}_t$ and target $\boldsymbol{y}_t$ parameter vectors are computed as follows. For PMA, the PMA signals are firstly segmented into overlapping frames using a 20 ms analysis window with 10 ms overlap. Next, sequences of ω consecutive frames, with a single-frame displacement, are concatenated together in order to better capture contextual phonetic

information. Due to the high dimensionality of the resultant windows of frames, the partial least squares (PLS) technique (De Jong, 1993) is applied to reduce the dimensionality and obtaining the final PMA parameter vectors used by the proposed conversion technique. The audio signals are represented in this work as 25 Mel-frequency cepstral coefficients (MFCCs) (Fukada et al., 1992) 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 (Gonzalez et al., 2014). Rather, the audio signals are re-synthesised 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 Evaluation of the Conversion Accuracy

To objectively evaluate the accuracy of speech reconstruction we use the well-known Mel-cepstral distortion (MCD) measure (Kubichek, 1993). The MCD measure is computed 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:

$$\mathrm{MCD[dB]} = \frac{10}{\ln 10}\sqrt{2\sum_{d=1}^{D}(c_d - \hat{c}_d)^2}. \qquad (18)$$

Subjective evaluation of resynthesised speech via listening tests, although also useful, is not reported here due to the preliminarily nature of this work. This is left for future work.

For estimating the performance of direct synthesis, a 10-fold cross-validation scheme is used. Hence, the available data for each speaker is randomly divided into ten sets 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 TIDigits Results

Fig. 2 shows a contour plot with the average MCD results obtained for all the speakers as a function of the number of mixture components used in the MFA model and the length of the sliding window used to extract the PMA parameter vectors. As expected, the more mixture components are used in the MFA model, the more accurately the non-linear PMA-to-acoustic mapping is represented and, hence, better MCD results are obtained. Moreover, increasing the length of the sliding-window used to extract the PMA feature vectors also helps, as this reduces the uncertainty of the mapping by taking into account more contextual information. In terms of speech intelligibility, informal listening of the resynthesised samples show that speech is intelligible and the speaker's voice is clearly identifiable[1].

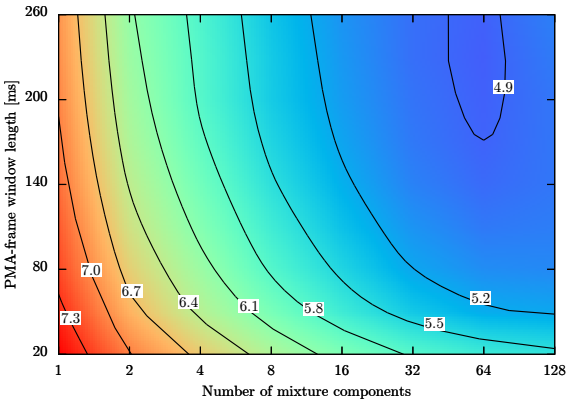

Figure 2: Average MCD results obtained for all the speakers in the TIDigits datasets as a function of the number of mixture components used in the MFA model and the length of the PMA-frame window.

The performance of direct synthesis for each speaker is shown in Fig. 3. A 64-component MFA model, which is the model providing better overall results in Fig. 2, is chosen now. It can be seen that the results for all the speakers follow a similar trend and by increasing the length of the PMA-frame window better results are achieved. For example, the relative improvements of using a window spanning 25 frames (260 ms) in comparison with just a single frame (20 ms) are 10.88%, 17.49%, 10.49%, 12.64% and 14.76% for speakers speakers M1-M4 and F1, respectively. In terms of reconstruction accuracy, the better results are obtained for speakers M1 and M4, whereas M2 and F1 are the worst speakers in this sense. We also see that the absolute difference between the results for M4 and F1 (best and worst speaker, respectively) is approximately 0.6 dB, which more or less correspond to the performance difference between using a window of 20 ms and a window of 260 ms for speaker M4. 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 the first reason, 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

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

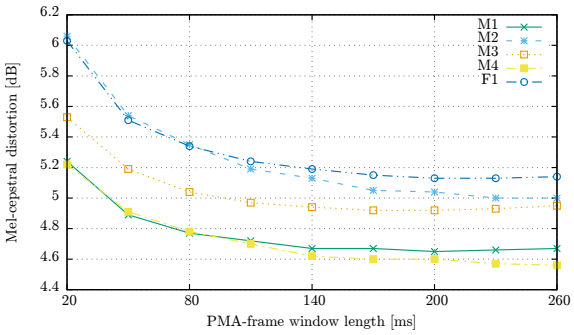

Figure 3: Performance of direct synthesis (in terms of Mel-Cepstral distortion) for different speakers in the TIDigits database.

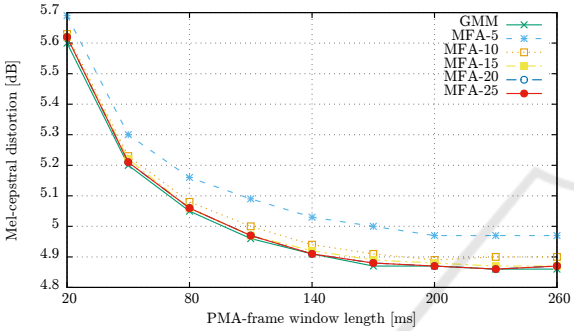

Figure 4: Comparison between the proposed approach for articulatory-to-acoustic conversion using MFAs and Toda's et al. approach using GMMs (Toda et al., 2008). For our proposal, the conversion accuracy using different latent space dimensions (i.e. 5, 10, 15, 20, and 25) for $\boldsymbol{v}_t$ in (5) is evaluated.

device. With respect to the latter reason, the current PMA device prototype was specifically designed for M1's anatomy, so it is reasonable to think that articulatory data is more accurately captured for him than for the other speakers.

We now compare the quality of the resynthesised voices obtained by our technique with the voices obtained by the well-known GMM-based conversion technique proposed by Toda et al. in (Toda et al., 2008; Toda et al., 2007). For a fairer comparison, both methods are evaluated using the MMSE-based mapping algorithm. Furthermore, 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 MFCC parameter vectors. Results are shown in Fig. 4. As can be seen, both methods perform almost equally except when the dimensionality of the latent space in the MFA-based conversion 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 latent spaces.

Table 1: MCD results (in dB) obtained for the Arctic database. Two conversion systems are compared: oracle and non oracle (see the text for details).

	PMA frame window length (ms)								
	20	50	80	110	140	170	200	230	260
Oracle	4.75	4.81	4.88	4.90	4.92	4.94	4.95	4.98	5.00
Non oracle	6.39	6.28	6.17	6.14	6.11	6.10	6.08	6.08	6.07

For dimensions greater than 15, we see that both approaches (GMM and MFA) 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 Arctic Results

Table 1 presents the MCD results obtained for the Arctic database when a 64-component MFA model is employed (the dimensionality of the latent space is 25). Two systems are compared. The non-oracle system corresponds to the PMA-to-acoustic conversion procedure described in Section 3. In the oracle system, on the other hand, the posterior probabilities $P(k|\boldsymbol{x}_t)$ used by the MMSE estimator in (15) are computed using both $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$, where $\boldsymbol{y}_t$ are the speech parameters extracted from the original audio signals (i.e. we are cheating). The rationale behind the oracle system is to evaluate the conversion performance when less uncertainty is involved in the mapping by simulating that we know in advance the dominant mixture component in each frame. As can be seen, there is a big drop between the results for the TIDigits database in Figure 2 and those presented here for the non-oracle system. The reason is the greater phonetic variability of the Arctic sentences. Thus, the PMA-to-acoustic mapping is more difficult to model for the Arctic sentences due to the greater uncertainty of the one-to-many conversion. Nevertheless, we see from the results obtained by the oracle system that better results can be obtained by reducing the uncertainty associated with the mapping. An example of the speech spectrograms obtained by the oracle and non-oracle conversion systems is shown in Figure 5. Perception of re-synthesised Arctic sentences is highly variable: some phrases are captured well, while others are incoherent.

Figure 6 shows a box plot with the detailed MCD results obtained by the non-oracle conversion system for different phone categories. When considering the manner of articulation, we can see that the nasals and plosives tend to be synthesised less accurately than other sound classes. For nasals, this is due to the current PMA prototype not modelling the velum area, whereas for the plosives the problem is the difficulty

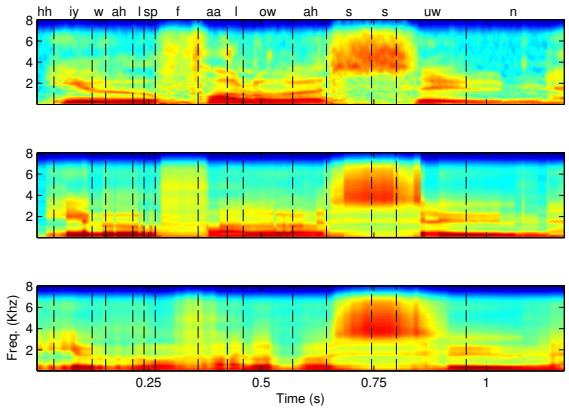

Figure 5: Examples of spectrograms of natural speech (top), oracle PMA-to-acoustic conversion (middle), and non-oracle conversion (bottom) for the utterance "He will follow us soon".

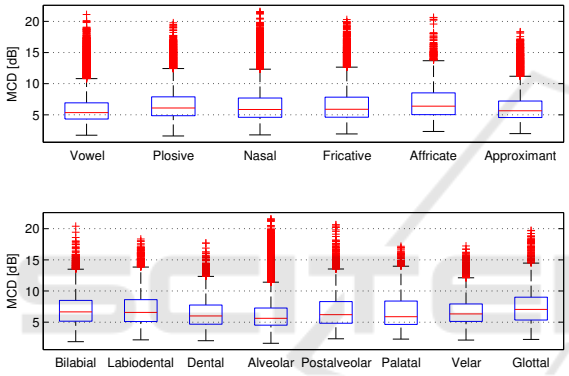

Figure 6: Detailed MCD results on the Arctic database for different speech sounds considering their manner (top) and place of articulation (bottom).

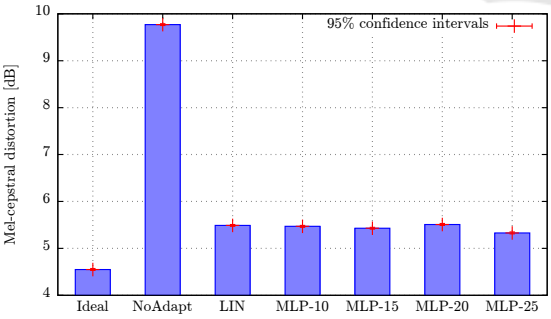

Figure 7: Cross-session synthesis results: MCD results obtained when synthesising PMA data from Session 2 in the TIDigits database using a MFA model trained on the Session 1 dataset.

in modelling their dynamics. When considering the place of articulation, it is difficult to extract any meaningful conclusions but we can see that many errors are made for the sounds articulated in the back of the mouth, which the current PMA prototype is not capturing well.

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 4, 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 7 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 4: 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.

As can be seen, 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 paper we have introduced a system for synthesising audible speech from speech articulator move-

ment captured from the lips and tongue using permanent magnet articulography. Preliminary evaluation of the system via objective 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 tasks. It has also been reported that one of the current limitations of PMA, 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 result brings us closer to being able to apply the direct synthesis method in a realistic treatment scenario. These results encourage us in pursuing our goal of developing a SSI that will ultimately allow laryngectomised patients to recover their voice. In order to reach this point, a number of questions will need to addressed in future research such as making better use of temporal context, improving the conversion accuracy for a large vocabulary, ways of recovering the prosodic information (i.e. voicing information and stress), and extending the technique to speech impaired speakers.

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

Braz, D. S. A., Ribas, M. M., Dedivitis, R. A., Nishimoto, I. N., and Barros, A. P. B. (2005). Quality of life and depression in patients undergoing total and partial laryngectomy. *Clinics*, 60(2):135–142.

Byrne, A., Walsh, M., Farrelly, M., and O'Driscoll, K. (1993). Depression following laryngectomy. A pilot study. *The British Journal of Psychiatry*, 163(2):173–176.

Cheah, L. A., Bai, J., Gonzalez, J. A., Ell, S. R., Gilbert, J. M., Moore, R. K., and Green, P. D. (2015). A user-centric design of permanent magnetic articulography based assistive speech technology. In *Proc. BioSignals*, pages 109–116.

Chen, L.-H., Ling, Z.-H., Liu, L.-J., and Dai, L.-R. (2014). Voice conversion using deep neural networks with layer-wise generative training. *IEEE/ACM Trans. Audio Speech Lang. Process.*, 22(12):1859–1872.

Danker, H., Wollbrück, D., Singer, S., Fuchs, M., Brähler, E., and Meyer, A. (2010). Social withdrawal after laryngectomy. *European Archives of Oto-Rhino-Laryngology*, 267(4):593–600.

De Jong, S. (1993). SIMPLS: an alternative approach to partial least squares regression. *Chemometrics Intell. Lab. Syst.*, 18(3):251–263.

Denby, B., Schultz, T., Honda, K., Hueber, T., Gilbert, J., and Brumberg, J. (2010). Silent speech interfaces. *Speech Commun.*, 52(4):270–287.

Fagan, M. J., Ell, S. R., Gilbert, J. M., Sarrazin, E., and Chapman, P. M. (2008). Development of a (silent) speech recognition system for patients following laryngectomy. *Medical engineering & physics*, 30(4):419–425.

Freitas, J., Teixeira, A., Bastos, C., and Dias, M. (2011). *Speech Technologies*, volume 10, chapter Towards a multimodal silent speech interface for European Portuguese, pages 125–150. InTech.

Fukada, T., Tokuda, K., Kobayashi, T., and Imai, S. (1992). An adaptive algorithm for Mel-cepstral analysis of speech. In *Proc. ICASSP*, pages 137–140.

Ghahramani, Z. and Hinton, G. E. (1996). The EM algorithm for mixtures of factor analyzers. Technical Report CRG-TR-96-1, University of Toronto.

Gilbert, J. M., Rybchenko, S. I., Hofe, R., Ell, S. R., Fagan, M. J., Moore, R. K., and Green, P. (2010). Isolated word recognition of silent speech using magnetic implants and sensors. *Medical engineering & physics*, 32(10):1189–1197.

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

Herff, C., Heger, D., de Pesters, A., Telaar, D., Brunner, P., Schalk, G., and Schultz, T. (2015). Brain-to-text: decoding spoken phrases from phone representations in the brain. *Frontiers in Neuroscience*, 9(217).

Hofe, R., Ell, S. R., Fagan, M. J., Gilbert, J. M., Green, P. D., Moore, R. K., and Rybchenko, S. I. (2013). Small-vocabulary speech recognition using a silent speech interface based on magnetic sensing. *Speech Commun.*, 55(1):22–32.

Hueber, T., Benaroya, E.-L., Chollet, G., Denby, B., Dreyfus, G., and Stone, M. (2010). Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips. *Speech Commun.*, 52(4):288–300.

Hueber, T., Benaroya, E.-L., Denby, B., and Chollet, G. (2011). Statistical mapping between articulatory and acoustic data for an ultrasound-based silent speech interface. In *Proc. Interspeech*, pages 593–596.

Jou, S.-C., Schultz, T., Walliczek, M., Kraft, F., and Waibel, A. (2006). Towards continuous speech recognition using surface electromyography. In *Proc. Interspeech*, pages 573–576.

Kominek, J. and Black, A. W. (2004). The CMU Arctic speech databases. In *Fifth ISCA Workshop on Speech Synthesis*, pages 223–224.

Kubichek, R. (1993). Mel-cepstral distance measure for objective speech quality assessment. In *Proc. IEEE Pa-*

cific Rim Conference on Communications, Computers and Signal Processing, pages 125–128.

Leonard, R. (1984). A database for speaker-independent digit recognition. In *Proc. ICASSP*, pages 328–331.

Petajan, E. D. (1984). *Automatic lipreading to enhance speech recognition (speech reading)*. PhD thesis, University of Illinois at Urbana-Champaign.

Schultz, T. and Wand, M. (2010). Modeling coarticulation in EMG-based continuous speech recognition. *Speech Commun.*, 52(4):341–353.

Toda, T., Black, A. W., and Tokuda, K. (2007). Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *IEEE Trans. Audio Speech Lang. Process.*, 15(8):2222–2235.

Toda, T., Black, A. W., and Tokuda, K. (2008). Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model. *Speech Commun.*, 50(3):215–227.

Toda, T., Nakagiri, M., and Shikano, K. (2012). Statistical voice conversion techniques for body-conducted unvoiced speech enhancement. *IEEE Trans. Audio Speech Lang. Process.*, 20(9):2505–2517.

Toth, A. R., Kalgaonkar, K., Raj, B., and Ezzat, T. (2010). Synthesizing speech from Doppler signals. In *Proc. ICASSP*, pages 4638–4641.

Uria, B., Renals, S., and Richmond, K. (2011). A deep neural network for acoustic-articulatory speech inversion. In *Proc. NIPS 2011 Workshop on Deep Learning and Unsupervised Feature Learning*.

Wand, M., Janke, M., and Schultz, T. (2014). Tackling speaking mode varieties in EMG-based speech recognition. *IEEE Trans. Bio-Med. Eng.*, 61(10):2515–2526.

Wang, J., Samal, A., Green, J. R., and Rudzicz, F. (2012). Sentence recognition from articulatory movements for silent speech interfaces. In *Proc. ICASSP*, pages 4985–4988.

Zahner, M., Janke, M., Wand, M., and Schultz, T. (2014). Conversion from facial myoelectric signals to speech: a unit selection approach. In *Proc. Interspeech*, pages 1184–1188.

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS