
Article:
Gonzalez, J.A., Cheah, L.A., Gilbert, J.M. et al. (4 more authors) (2016) A silent speech system based on permanent magnet articulography and direct synthesis. Computer Speech & Language, 39. C. pp. 67-87. ISSN 0885-2308

https://doi.org/10.1016/j.csl.2016.02.002

A Silent Speech System based on Permanent Magnet Articulography and Direct Synthesis

Jose A. Gonzalez[a,*], Lam A. Cheah[b], James M. Gilbert[b], Jie Bai[b], Stephen R. Ell[c], Phil D. Green[a], Roger K. Moore[a]

[a] *Department of Computer Science, The University of Sheffield, Sheffield, UK*
[b] *School of Engineering, University of Hull, Kingston upon Hull, UK*
[c] *Hull and East Yorkshire Hospitals Trust, Castle Hill Hospital, Cottingham, UK*

Abstract

In this paper we present a silent speech interface (SSI) system aimed at restoring speech communication for individuals who have lost their voice due to laryngectomy or diseases affecting the vocal folds. In the proposed system, articulatory data captured from the lips and tongue using permanent magnet articulography (PMA) are converted into audible speech using a speaker-dependent transformation learned from simultaneous recordings of PMA and audio signals acquired before laryngectomy. The transformation is represented using a mixture of factor analysers, which is a generative model that allows us to efficiently model non-linear behaviour and perform dimensionality reduction at the same time. The learned transformation is then deployed during normal usage of the SSI to restore the acoustic speech signal associated with the captured PMA data. The proposed system is evaluated using objective quality measures and listening tests on two databases containing PMA and audio recordings for normal speakers. Results show that it is possible to reconstruct speech from articulator movements captured by an unobtrusive technique without an intermediate recognition step. The SSI is capable of producing speech of sufficient intelligibility and naturalness that the speaker is clearly identifiable, but problems remain in scaling up the process to function consistently for phonetically-rich vocabularies.

Keywords: Silent speech interfaces, speech rehabilitation, speech synthesis, permanent magnet articulography, augmentative and alternative communication
2010 MSC: 68T05, 69T10, 92C55

*Corresponding author
Email address: j.gonzalez@sheffield.ac.uk (Jose A. Gonzalez)

1. Introduction

Despite speech being our preferred and most natural form of communication, normal speech communication can be impossible or undesirable in some situations. Adverse noise conditions might make speech unintelligible and there are diseases that lead to a person losing their voice or having their ability to speak severely impaired. These include trauma, cancer of the larynx, and some neurological disorders. Sometimes audible speech may not be desirable, e.g. private conversations in public areas. The main obstacle to communication in these situations derives from the acoustic speech signal: its quality is severely affected or non-existent in some situations, whereas it is desirable to avoid generating it in other situations. In all these situations silent speech interfaces (SSIs) can help.

A SSI is a system that enables speech communication in the absence of audible speech by exploiting other biosignals associated with speech production (Denby et al., 2010). Several types of SSIs have been proposed using different sensing technologies to capture speech-related biosignals. Some work has been done, with limited success, using brain-computer interfaces (BCIs) such as intracranial electrocorticography (ECoG) (Brumberg et al., 2010, 2011; Herff et al., 2015) or electroencephalography (EEG) (Wester, 2006; Brigham and Vijaya Kumar, 2010) to decode the brain activity associated with particular thoughts or intentions of a subject. Other SSIs use the electrical activity of the articulator muscles. The most widespread technology for capturing this information is surface electromyography (sEMG) (Jou et al., 2006; Schultz and Wand, 2010; Janke et al., 2012; Wand et al., 2014; Zahner et al., 2014; Deng et al., 2014). Alternatively, SSIs can also be based on the movement of the speech articulators. Different technologies have been used to capture articulator motion including video (Petajan, 1984; Petajan et al., 1988; Matthews et al., 2002), ultrasound (Cai et al., 2013), both video and ultrasound (Hueber et al., 2010, 2011), electromagnetic articulography (EMA) (Toda et al., 2008; Toutios and Narayanan, 2013), magnetic resonance imaging (MRI) (Badin et al., 2002; Birkholz and Jackel, 2003) and radar (Toth et al., 2010). In this paper we employ an alternative approach for capturing articulator movement: permanent magnet articulography (PMA) (Fagan et al., 2008; Gilbert et al., 2010; Hofe et al., 2013b,a; Cheah et al., 2015). In PMA a set of magnets are attached to the articulators (typically the lips and tongue) and the magnetic field generated while the user 'speaks' is captured by a number of sensors located around the mouth. Compared with other techniques for capturing articulator movement such as EMA or sEMG, PMA has the potential to be unobtrusive as there are no wires coming out of the mouth or electrodes attached to the skin.

The speech-related biosignals generated during speech production can then be used to determine the acoustic signal associated with those signals. The most common way of doing this would be to decode the message encoded in the biosignals using automatic speech recognition (ASR), and then use a text-to-speech (TTS) synthesiser to generate the final acoustic signal from the recognised text. Although this approach for speech reconstruction has several advantages, such

as rapid development by using readily available ASR and TTS tools and the possibility of obtaining a better speech signal reconstruction (especially of the voicing[1]) by exploiting the textual representation in the TTS synthesiser, it also has drawbacks (Hofe et al., 2011). First, the approach is constrained to the language and vocabulary of the recogniser. Second, speech articulation and its associated auditory feedback are disconnected due to the variable delay introduced by the ASR and TTS steps. Third, non-linguistic information encoded in the articulatory data (e.g. subject's gender, age or mood) is normally lost after the ASR step. These drawbacks, particularly the second one, may affect the willingness of a SSI user to engage in social interactions. This means that, at best, a recognise-then-synthesise system would be like having an interpreter. To address these problems, we can resort to an alternative approach for speech restoration: *direct speech synthesis from the biosignals without an intermediate recognition step*.

The direct synthesis (DS) approach attempts to model the relationship between the speech-related biosignals and their acoustics. In comparison with the recognise-then-synthesise approach, DS has the advantage that is not limited to a specific vocabulary and is language-independent. Moreover, it can allow real-time speech synthesis. There is also the possibility that real-time auditory feedback might enable the user to learn to produce better speech: like learning to play an instrument. At best, DS could restore the user's voice, lost by excision of the larynx. Assuming that the biosignals represent articulatory data, as with PMA, and that a parametric representation of speech (i.e. a vocoder) is adopted, the modelling of the articulatory-to-acoustic mapping presents some challenging problems. First, this mapping is known to be non-linear (Atal et al., 1978; Qin and Carreira-Perpiñán, 2007; Neiberg et al., 2008; Ananthakrishnan et al., 2012). Furthermore, in some cases the mapping is non-unique, that is, the same articulatory data might correspond to different acoustic realizations. The reason for this non-uniqueness is that typically the sensing technology used by the SSI only provides an incomplete picture of the speech production process and some of the information about this process is missing or not well captured.

Several techniques have been proposed in the literature for representing the articulatory-to-acoustic mapping. In general, these techniques can be classified into two categories: model-based and stereo-based. Model-based techniques such as those proposed in Schroeter and Sondhi (1994); Birkholz et al. (2008); Toutios et al. (2011); Toutios and Narayanan (2013) use articulatory data to drive an articulatory synthesiser, which implements a physical model of speech production that can be controlled using a small set of control parameters (Rubin et al., 1981; Maeda, 1982). Thus, these techniques attempt to find a mapping between the articulatory data and the control parameters of the synthesiser. Stereo-based techniques, in contrast, learn the direct correspondence between the articulatory and acoustic domains using parallel data, i.e.

[1]The speech biosignals generally provide little information about the voicing of speech, particularly when the SSI is used by laryngectomees (Gonzalez et al., 2014).

simultaneous recordings of articulatory and speech data. To learn the transformation between these domains from the parallel data, several approaches have been proposed including statistical approaches based on Gaussian mixture models (GMMs) (Toda et al., 2008, 2012b; Nakamura et al., 2012), hidden Markov models (HMMs) (Hueber et al., 2012), shared Gaussian process dynamical models (Gonzalez et al., 2015), neural networks (Desai et al., 2009), support vector regression (Toutios and Margaritis, 2005), and a concatenative, unit-selection approach (Zahner et al., 2014). Most of these approaches were originally developed for voice conversion (VC) (Stylianou et al., 1998; Toda et al., 2007) and, in general terms, the techniques developed for VC can also be applied to stereo-based articulatory-to-acoustic tasks.

In this paper we present a silent speech system that is able to convert articulator motion data captured using PMA into audible speech. From the two DS approaches outlined above, we opt for a stereo-based approach for two reasons. First, the availability of parallel datasets enables the direct modelling of the PMA-to-acoustic mapping using machine learning techniques. The second reason is that current models of speech production (i.e. articulatory synthesisers) are still not mature enough compared to other approaches such as statistical parametric speech synthesis. In our proposed technique, 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 the training phase to estimate the joint probability distribution of PMA and speech parameter vectors. To represent the distribution, a generative approach based on mixture of factor analysers (MFA) is proposed in this work. Then, during normal usage of the SSI, the speech-parameter posterior distribution given the PMA data is evaluated in order to convert the captured articulatory data into an acoustic signal. Two alternative conversion algorithms are investigated in this work for transforming PMA parameter vectors to speech parameter ones. The first one is based on the well-known minimum mean square error (MMSE) estimator. A limitation of this algorithm is that it works on a frame-by-frame basis; thus imposing no temporal constraints on the reconstructed speech signal. To encourage smooth trajectories on the reconstructed speech parameters, we also investigate the application of the maximum likelihood estimation (MLE) algorithm proposed by Tokuda et al. (2000); Toda et al. (2007), which takes into account the statistics of both the static and dynamic speech parameters, to our specific problem. The proposed techniques are evaluated using objective and subjective quality measures on parallel datasets with PMA and audio material recorded for several speakers.

This work forms part of our continuing effort to develop an acceptable and discrete PMA-based SSI for laryngectomy patients. Key milestones in our previous work that build up to this paper are as follows:

- First, in Gilbert et al. (2010); Hofe et al. (2013b), speech recognition from PMA data was reported to achieve similar accuracy results to using audio on isolated words and connected digits recognition tasks.
- Then, in Hofe et al. (2013a), the study of PMA-based speech recognition

was successfully extended to multiple speakers.

- More recently, extensive investigation into the effectiveness of PMA data for discriminating the voicing, place and manner of articulation of English phones was presented in Gonzalez et al. (2014).
- With respect to the direct synthesis approach, a feasibility study was presented in Hofe et al. (2011) in which speech formants were estimated from PMA data.

As previously stated, our long term plan is to build an SSI that is able to generate high quality speech from PMA data in real time. For laryngectomy patients, this will involve simultaneously recording both PMA data and the patient's voice before laryngectomy. Then, after laryngectomy has been performed, the direct synthesis models trained on the patient's voice will be used to generate speech. In cases where it is impractical to record parallel data before the operation we can, for instance, record acoustics only and then, after laryngectomy, ask the patient to mime along to their own pre-recorded voice to provide the sensor data stream. In cases where the voice has been destroyed prior to the laryngectomy, patients could be asked to mime along to a 'donor voice'.

The rest of this paper is organised as follows. First, in Section 2, the functional principles of the PMA technique are outlined. Section 3 presents the mathematical details of the direct synthesis technique we use in this paper for generating speech from PMA data. Then, in Section 4, the technique is evaluated on parallel databases containing PMA and acoustic data. The results obtained are discussed in Section 5. Finally, we summarise this paper and outline future work in Section 6.

2. Permanent Magnet Articulography

The principle of PMA is that the motion of the articulators may be determined by attaching a set of magnets to the articulators and measuring the resultant magnetic field variations using a set of magnetic sensors located around the mouth. These field variations may then be used to determine the speech which the user wishes to produce. It should be noted that the magnetic field detected at each sensor is a composite of the field from each magnet and that the contribution from each magnet is a non-linear function of its position and orientation. Due to the complexity of the interaction between magnets and the sensed field, it is not currently the intention that the sensor information be used to determine the Cartesian positions/orientations of the magnets, but rather that the composite field be mapped to speech features.

A number of implementations of PMA have been investigated in recent years (Fagan et al., 2008; Gilbert et al., 2010; Hofe et al., 2013b,a). Earlier prototypes provided acceptable recognition performance but were not particularly satisfactory in terms of their appearance, comfort and ergonomic factors for the users. To address these limitations, the latest PMA device was developed based on a

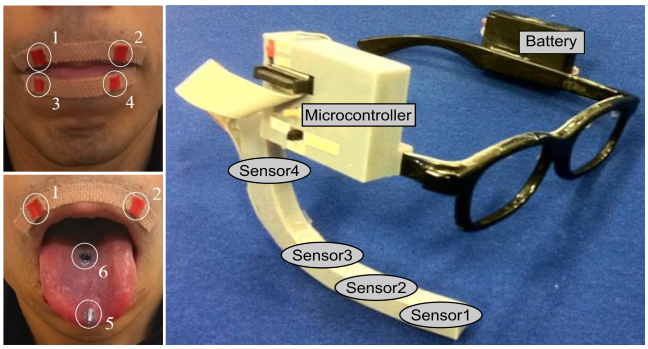

Figure 1: Overview of the PMA technique for capturing articulator motion data. *Upper-left and lower-left panels*: placement of the magnets used to detect the movement of the lips and tongue. The size of the magnets is 1mm (diameter) × 5mm (length) for magnets 1-4, 2mm × 4mm for magnet 5 and 5mm × 1mm for magnet 6. *Right panel*: headset used to house the four magnetic sensors that detect the variations of the magnetic field generated by the magnets.

user-centric approach (Cheah et al., 2015). The prototype was re-designed based on feedback from user questionnaires and through discussion with stakeholders including clinicians, potential users and their families.

The new PMA device has much improved appearance, portability and miniaturised hardware. Nevertheless, the prototype showed a comparable performance to its predecessor (Cheah et al., 2015; Gonzalez et al., 2014). Key components of the device include a set of six Neodynium Iron Boron (NdFeB) permanent magnets attached to the lips and tongue as illustrated in Fig. 1. These magnets are currently attached using Histoacryl surgical tissue adhesive (Braun, Melsungen, Germany), but would be surgically implanted for long term use. The remainder of the PMA system is composed of a set of four tri-axial Anisotropic Magnetoresistive (AMR) magnetic sensors mounted on a bespoke wearable headset, a control unit, a rechargeable battery and a processing unit (e.g. computer/tablet PC). Detailed information on these hardware modules and their operation is presented in Cheah et al. (2015).

During data acquisition the outputs of the first three magnetic sensors in Fig. 1 (i.e. 9 channels) are used to capture the magnetic field changes arising from the movements of the magnets attached to the articulators. The output of the fourth sensor (which is placed further away from the mouth) is used as a reference for background cancellation to compensate for the effect of the earths magnetic field on the other three sensors. The acquired data is sampled at 100Hz and transmitted, either through a USB connection or via Bluetooth, to a computer/tablet PC for further processing. The PMA data is first low-pass filtered to remove 50Hz electrical noise, and then normalised prior to further processing (Hofe et al., 2013b; Cheah et al., 2015).

3. Direct speech synthesis from PMA data

In this section we present the details of the proposed technique for speech parameter generation from PMA data. Formally, the aim of this technique is to find a mapping function $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x})$ for transforming source feature vectors $\boldsymbol{x}$ into target feature vectors $\boldsymbol{y}$. In our case, the source vectors are derived from the PMA data captured by the SSI, while the target vectors correspond to a parametric representation of speech, typically Mel-frequency cepstral coefficient (MFCC) parameters (Fukada et al., 1992). To model the PMA-to-acoustic mapping, we resort to a statistical approach in which the parameters of the mapping function are learned from training data containing parallel recordings of PMA and acoustic data. The proposed approach consists of two phases. Firstly, in the training phase, the parallel data is used to learn the parameters of the joint distribution of source and target vectors $p(\boldsymbol{x}, \boldsymbol{y})$. The details of the training phase are given in Section 3.1. Then, in the conversion phase, the learned parameters are used to derive the conditional distribution $p(\boldsymbol{y}|\boldsymbol{x})$ which, in turn, allows us to find the target acoustic vector associated with a particular observation (i.e. a PMA feature vector). This is discussed in Section 3.2.

3.1. Training phase

Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be the PMA and acoustic parameter vectors with dimensions D_x and D_y, respectively. Instead of directly modelling the relationship between both variables as $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x})$, we assume that $\boldsymbol{x}$ and $\boldsymbol{y}$ are outputs of an underlying stochastic process whose state $\boldsymbol{v}$ is not directly observable. Furthermore, we will assume that $D_v \ll D_x, D_y$, such that the latent variable offers a more compact representation of the data. Then, the relationship between the latent space and the observed variables can be expressed as,

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

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

where $\boldsymbol{\epsilon}_x$ and $\boldsymbol{\epsilon}_y$ are noise processes. To make inference tractable, a common assumption in latent variable models is to consider $\boldsymbol{\epsilon}_x$ and $\boldsymbol{\epsilon}_y$ as Gaussian with zero mean and diagonal covariances $\boldsymbol{\Psi}_x$ and $\boldsymbol{\Psi}_y$, respectively.

Assuming that $\boldsymbol{v}$ encodes the vocal tract shape at a given time instant, the mapping functions $\boldsymbol{f}_x$ and $\boldsymbol{f}_y$ will be non-linear as indicated above. Although $\boldsymbol{v}$ might not have any physical interpretation, the non-linearity of $\boldsymbol{f}_x$ and $\boldsymbol{f}_y$ will still hold in general terms. Hence, in order to accurately model $\boldsymbol{f}_x$ and $\boldsymbol{f}_y$, we have to deploy non-linear regression techniques. In this work we adopt a mixture of factor analysers (MFAs) Ghahramani and Hinton (1996) approach in which the mapping functions are approximated in a piecewise linear fashion. The functions are approximated by a mixture of K factor analysis (FA) models Anderson (2003), each of which has the following form,

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

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

where $k = 1, \ldots, K$ is the FA model index; $\boldsymbol{W}_x^{(k)}$, $\boldsymbol{W}_y^{(k)}$ are linear transformation matrices (a.k.a. *factor loadings*); $\boldsymbol{\mu}_x^{(k)}$, $\boldsymbol{\mu}_y^{(k)}$ are vectors that allow the data to have a non-zero mean; and $\boldsymbol{x}^{(k)}$, $\boldsymbol{y}^{(k)}$ denote local approximations of $\boldsymbol{x}$ and $\boldsymbol{y}$ around the means $\boldsymbol{\mu}_x^{(k)}$ and $\boldsymbol{\mu}_y^{(k)}$, respectively. The number of local models K can be optimised by cross-validation or using a validation set. The above equation can be written more compactly as,

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

where $\boldsymbol{z} = [\boldsymbol{x}^\top, \boldsymbol{y}^\top]^\top$, $\boldsymbol{W}_z^{(k)} = [{\boldsymbol{W}_x^{(k)}}^\top {\boldsymbol{W}_y^{(k)}}^\top]^\top$, $\boldsymbol{\mu}_z^{(k)} = [{\boldsymbol{\mu}_x^{(k)}}^\top, {\boldsymbol{\mu}_y^{(k)}}^\top]^\top$, 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,

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

Using the generative model in (5), we can now write the joint pdf of source and target vectors as the following mixture distribution,

$$p(\boldsymbol{z}) = \sum_{k=1}^{K} \pi^{(k)} p(\boldsymbol{z}|k), \tag{7}$$

where $\pi^{(k)}$ are the mixture weights and the likelihood $p(\boldsymbol{z}|k)$ is given by

$$p(\boldsymbol{z}|k) = \int p(\boldsymbol{z}|\boldsymbol{v}, k) p(\boldsymbol{v}|k) d\boldsymbol{v}, \tag{8}$$

with $p(\boldsymbol{z}|\boldsymbol{v}, k) = \mathcal{N}(\boldsymbol{W}_z^{(k)}\boldsymbol{v} + \boldsymbol{\mu}_z^{(k)}, \boldsymbol{\Psi}_z^{(k)})$ (deduced from (5)). As in standard factor analysis, the factors $\boldsymbol{v}$ are assumed to be distributed according to $p(\boldsymbol{v}|k) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. Under this assumption, it can be shown (see e.g. Appendix B of Bishop (2006)) that $p(\boldsymbol{z}|k)$ above simplifies to a Gaussian distribution with mean $\boldsymbol{\mu}_z^{(k)}$ and reduced-rank covariance matrix given by $\boldsymbol{\Sigma}_z^{(k)} = \boldsymbol{\Psi}_z^{(k)} + \boldsymbol{W}_z^{(k)} {\boldsymbol{W}_z^{(k)}}^\top$. This matrix can also be expressed in terms of the correlations between the source and target vectors by defining the following partitions,

$$\begin{aligned} \boldsymbol{\Sigma}_z^{(k)} &= \begin{bmatrix} \boldsymbol{\Sigma}_{xx}^{(k)} & \boldsymbol{\Sigma}_{xy}^{(k)} \\ \boldsymbol{\Sigma}_{yx}^{(k)} & \boldsymbol{\Sigma}_{yy}^{(k)} \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{\Psi}_x^{(k)} + \boldsymbol{W}_x^{(k)} {\boldsymbol{W}_x^{(k)}}^\top & \boldsymbol{W}_x^{(k)} {\boldsymbol{W}_y^{(k)}}^\top \\ \boldsymbol{W}_y^{(k)} {\boldsymbol{W}_x^{(k)}}^\top & \boldsymbol{\Psi}_y^{(k)} + \boldsymbol{W}_y^{(k)} {\boldsymbol{W}_y^{(k)}}^\top \end{bmatrix}. \end{aligned} \tag{9}$$

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

3.2. Conversion phase

The conversion phase of the proposed approach for direct speech synthesis addresses the problem of transforming articulatory data into audible speech. In this section we will only address the problem of transforming PMA feature vectors into speech parameter vectors, relying on the corresponding vocoder for obtaining the final time-domain acoustic signal from the estimated speech parameters. Formally, the PMA-to-acoustic conversion problem involves finding the target speech vector $\boldsymbol{y}$ that corresponds to an observed PMA feature vector $\boldsymbol{x}$. In our statistical-based conversion system, the information about this mapping is available in the form of the conditional distribution $p(\boldsymbol{y}|\boldsymbol{x})$, which is derived from the joint distribution $p(\boldsymbol{x}, \boldsymbol{y})$ in (7) as,

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

where

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

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

The mean and covariance of the k-th component conditional distribution $p(\boldsymbol{y}|\boldsymbol{x}, k)$ are obtained using the properties of the joint Gaussian distribution:

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

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

where $\boldsymbol{\Sigma}_{xx}^{(k)}$, $\boldsymbol{\Sigma}_{yy}^{(k)}$, $\boldsymbol{\Sigma}_{xy}^{(k)}$, and $\boldsymbol{\Sigma}_{yx}^{(k)}$ are given by (9).

From (10) we see that $p(\boldsymbol{y}|\boldsymbol{x})$ adopts the form of a mixture distribution with possibly more than one mode. Hence, different estimated values for the speech parameters may be obtained depending on the specific estimator employed in the conversion process. In the following, we introduce two different conversion techniques based on two well-known statistical estimators: MMSE and MLE considering the dynamic speech features.

3.2.1. MMSE conversion

The MMSE estimator is defined as the conditional expectation of $\boldsymbol{y}$ given the observation $\boldsymbol{x}$:

$$\hat{\boldsymbol{y}} = \mathbb{E}[\boldsymbol{y}|\boldsymbol{x}] = \int \boldsymbol{y} p(\boldsymbol{y}|\boldsymbol{x}) d\boldsymbol{y}. \tag{15}$$

By substituting the expression of the conditional distribution $p(\boldsymbol{y}|\boldsymbol{x})$ in (10) into (15), the estimator finally becomes,

$$\hat{\boldsymbol{y}} = \sum_{k=1}^{K} P(k|\boldsymbol{x}) \left(\boldsymbol{A}^{(k)} \boldsymbol{x} + \boldsymbol{b}^{(k)}\right), \tag{16}$$

where $\boldsymbol{A}^{(k)}$ and $\boldsymbol{b}^{(k)}$ are derived from the k-th component conditional mean in (13) as,

$$\boldsymbol{A}^{(k)} = \boldsymbol{\Sigma}_{yx}^{(k)} \boldsymbol{\Sigma}_{xx}^{(k)^{-1}}, \tag{17}$$

$$\boldsymbol{b}^{(k)} = \boldsymbol{\mu}_{y}^{(k)} - \boldsymbol{A}^{(k)} \boldsymbol{\mu}_{x}^{(k)}. \tag{18}$$

We can see from (16) that no continuity constraints are imposed when reconstructing the speech parameter trajectories in the MMSE estimator, which may lead to reduced speech quality. In order to address this issue, an MLE-based conversion technique is introduced in the next section.

3.2.2. MLE conversion

Let $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ be the source and target parameter vectors at frame t, respectively. From the sequence of acoustic speech vectors $\boldsymbol{Y} = \left[\boldsymbol{y}_1^\top, \ldots, \boldsymbol{y}_T^\top\right]^\top$ we define the first-order difference parameters (i.e. dynamic features) at time t as,

$$\Delta \boldsymbol{y}_t = \boldsymbol{y}_t - \boldsymbol{y}_{t-1}. \tag{19}$$

The augmented target vector containing both the static and dynamic parameters is then denoted by $\overline{\boldsymbol{y}}_t = [\boldsymbol{y}_t^\top, \Delta \boldsymbol{y}_t^\top]^\top$ and similarly the sequence of augmented target vectors by $\overline{\boldsymbol{Y}} = [\overline{\boldsymbol{y}}_1^\top, \ldots, \overline{\boldsymbol{y}}_T^\top]^\top$. For the purpose of MLE, it is convenient to express the relationship between $\boldsymbol{Y}$ and $\overline{\boldsymbol{Y}}$ as the following linear transformation,

$$\overline{\boldsymbol{Y}} = \boldsymbol{R}\boldsymbol{Y}, \tag{20}$$

where $\boldsymbol{R}$ is the following $(2 \cdot D_y \cdot T) \times (D_y \cdot T)$ block matrix,

$$\boldsymbol{R} = \begin{bmatrix} \mathbf{1} & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} & \mathbf{0} \\ \mathbf{1} & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \mathbf{0} & \ldots & \mathbf{0} & \mathbf{0} \\ -\mathbf{1} & \mathbf{1} & \mathbf{0} & \ldots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} & \mathbf{1} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \ldots & -\mathbf{1} & \mathbf{1} \end{bmatrix}, \tag{21}$$

with $\mathbf{0}$, $\mathbf{1}$ and $-\mathbf{1}$ denoting the $D_y \times D_y$ zero, identity and negative identity matrices, respectively. In the above matrix we have assumed that $\Delta \boldsymbol{y}_0 = \boldsymbol{y}_0$.

From the sequence of articulatory data $\boldsymbol{X}$ the MLE conversion algorithm tries to recover the sequence of acoustic speech parameters $\boldsymbol{Y}$ that simultaneously maximises the likelihood of the static and dynamic parameters. Mathematically, this can be expressed as follows,

$$\hat{\boldsymbol{Y}} = \arg\max_{\boldsymbol{Y}} p(\overline{\boldsymbol{Y}}|\boldsymbol{X}) = \arg\max_{\boldsymbol{Y}} p(\boldsymbol{R}\boldsymbol{Y}|\boldsymbol{X}), \tag{22}$$

where $\hat{\boldsymbol{Y}} = [\hat{\boldsymbol{y}}_1^\top, \ldots, \hat{\boldsymbol{y}}_T^\top]^\top$ is the estimated sequence of acoustic speech parameters and the likelihood $p(\overline{\boldsymbol{Y}}|\boldsymbol{X})$ is obtained by assuming independence among frames (see Toda et al. (2007, 2008)),

$$p(\overline{\boldsymbol{Y}}|\boldsymbol{X}) = \prod_{t=1}^{T} p(\overline{\boldsymbol{y}}_t|\boldsymbol{x}_t). \tag{23}$$

and the conditional distribution $p(\overline{\boldsymbol{y}}_t|\boldsymbol{x}_t)$ is again a mixture distribution as in (10). It must be pointed out, however, that this distribution is now derived from a joint distribution $p(\boldsymbol{x}, \overline{\boldsymbol{y}})$ representing PMA feature vectors and augmented speech parameter vectors (i.e. with static and dynamic features).

Direct maximisation of (22) is not possible because of the hidden mixture component sequence $\boldsymbol{k} = (k_1, k_2, \ldots, k_T)$ that appears in (23) as a consequence of $p(\overline{\boldsymbol{y}}_t|\boldsymbol{x}_t)$ being a mixture distribution. Hence, we adopt the iterative EM algorithm proposed in Tokuda et al. (2000); Toda et al. (2007). Let $\overline{\boldsymbol{Y}}$ be the sequence of augmented acoustic speech parameters to be optimised by the EM algorithm. Similarly, $\overline{\boldsymbol{Y}}^{\text{old}}$ is the current estimate of the augmented sequence. Then, the EM algorithm proceeds by iteratively optimising the following auxiliary Q-function with respect to $\overline{\boldsymbol{Y}}$,

$$\begin{aligned}\mathcal{Q}(\overline{\boldsymbol{Y}}, \overline{\boldsymbol{Y}}^{\text{old}}) &= \sum_{t=1}^{T}\sum_{k=1}^{K} P(k|\boldsymbol{x}_t, \overline{\boldsymbol{y}}_t^{\text{old}}) \log p(\overline{\boldsymbol{y}}_t, k|\boldsymbol{x}_t) \\ &\propto -\frac{1}{2}(\boldsymbol{R}\boldsymbol{Y})^\top \boldsymbol{\Phi}(\boldsymbol{R}\boldsymbol{Y}) + (\boldsymbol{R}\boldsymbol{Y})^\top \boldsymbol{\lambda},\end{aligned} \tag{24}$$

where $\boldsymbol{\Phi}$ is the following $(2 \cdot D_y \cdot T) \times (2 \cdot D_y \cdot T)$ block matrix,

$$\boldsymbol{\Phi} = \begin{bmatrix} \boldsymbol{C}_1 & \boldsymbol{0} & \boldsymbol{0} & \ldots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{C}_2 & \boldsymbol{0} & \ldots & \boldsymbol{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \ldots & \boldsymbol{C}_T \end{bmatrix}, \tag{25}$$

and $\boldsymbol{\lambda}$ is the following $(2 \cdot D_y \cdot T)$-dimensional vector,

$$\boldsymbol{\lambda} = \left[\boldsymbol{m}_1^\top, \boldsymbol{m}_2^\top, \ldots, \boldsymbol{m}_T^\top\right]^\top. \tag{26}$$

The $(2 \cdot D_y) \times (2 \cdot D_y)$ matrices $\boldsymbol{C}_t$ and $(2 \cdot D_y)$-dimensional vectors $\boldsymbol{m}_t$ $(t = 1, \ldots, T)$ that appear in (25) and (26), respectively, are defined as the following expected values,

$$\boldsymbol{C}_t = \sum_{k=1}^{K} P(k|\boldsymbol{x}_t, \overline{\boldsymbol{y}}_t^{\text{old}}) \boldsymbol{\Sigma}_{\overline{\boldsymbol{y}}|x,t}^{(k)^{-1}}, \tag{27}$$

$$\boldsymbol{m}_t = \sum_{k=1}^{K} P(k|\boldsymbol{x}_t, \overline{\boldsymbol{y}}_t^{\text{old}}) \boldsymbol{\Sigma}_{\overline{\boldsymbol{y}}|x,t}^{(k)^{-1}} \boldsymbol{\mu}_{\overline{\boldsymbol{y}}|x,t}^{(k)}, \tag{28}$$

where $\boldsymbol{\mu}_{\overline{y}|x,t}^{(k)}$ and $\boldsymbol{\Sigma}_{\overline{y}|x,t}^{(k)}$ are the mean vector and covariance matrix of the conditional distribution $p(\overline{\boldsymbol{y}}|\boldsymbol{x}_t, k)$ and are given by (13) and (14), respectively.

Finally, by setting the derivatives of (24) w.r.t. $\boldsymbol{Y}$ to zero and solving, we obtain the following expression for updating the estimated sequence of acoustic speech parameters,

$$\boldsymbol{Y}^{\text{new}} = (\boldsymbol{R}^\top \boldsymbol{\Phi} \boldsymbol{R})^{-1} \boldsymbol{R}^\top \boldsymbol{\lambda}. \tag{29}$$

The above equation is iteratively applied until a certain stopping criterion is met (e.g. a number of iterations is reached). The EM algorithm guarantees that each iteration of the iterative procedure increases the log-likelihood $\log p(\boldsymbol{RY}|\boldsymbol{X})$ and, consequently, a better speech reconstruction is expected after each iteration. As initial estimate of the speech parameters we use the MMSE estimate in (16).

As opposed to the MMSE conversion technique, the MLE technique in (29) performs a sequence-by-sequence mapping, rather than a frame-by-frame conversion. Thus, it is expected that more accurate speech parameter reconstructions can be obtained by the MLE technique. The drawback of this technique, in comparison with the MMSE technique, is that it is difficult to implement it in real time due to the sequence-by-sequence conversion process. Nevertheless, recent work (Toda et al., 2012a; Moriguchi et al., 2013) has extended the MLE technique to enable real time voice conversion.

4. Experimental evaluation

In this section we evaluate the reconstruction performance of the proposed technique for PMA-to-acoustic conversion on parallel datasets containing both PMA and acoustic data. Since our goal in this work is to evaluate the feasibility of direct speech synthesis from PMA data, results are only reported for non-impaired subjects. Evaluation of the proposed technique for laryngectomy patients, where we may not be able to directly record parallel data, is left for future work.

4.1. Vocabulary choice and data acquisition

Two parallel PMA-and-acoustic databases with different phonetic coverage were recorded. The first one follows 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 number of phones in this vocabulary is 21: 11 vowels and 10 consonants. With this database we aim to establish that our method creates intelligible output. In order to perform an in-depth analysis of the reconstruction accuracy at the phone level, a second corpus was designed in a more systematic manner. We know from previous work (Gonzalez et al., 2014) that the ability of PMA for detecting the manner of articulation and voicing of speech sounds is limited and, therefore, we need to determine to what extent this limitation affects direct synthesis. The vocabulary in this case consists of 48 isolated consonant-vowel (CV) syllables obtained by combining 12 consonants

	p	m	f	θ	ʃ	t	n	r	l	s	k	h
i	pea	me	fee	thee	she	tea	nee	ree	lee	see	key	he
u	poo	moo	foo	thoo	shoe	too	noo	roo	loo	sue	coo	who
ɔ	por	mor	for	thor	shaw	tor	nor	raw	law	saw	core	hoar
a	pah	mah	fah	thah	shah	tah	nah	rah	lah	sah	kah	hah

Table 1: Consonants and vowels used for building up the vocabulary in the CV database.

with 4 vowels, as shown in Table (1). The construction of this vocabulary was as follows. From the set of English vowels, we choose those four most distinctive from the articulation point of view. Thus, [a i u] were chosen because they are at three of the corners in the International Phonetic Association (IPA) vowel chart. The fourth corner would be [ɑ ʊ], however [ɔ], which is also close to the fourth corner, was selected because we thought it was easy to pronounce for British English speakers. Unvoiced consonants were preferred over voiced ones due to the limited accuracy of PMA for detecting voicing. Apart from that, the consonants were chosen to have a high coverage of the IPA consonant chart, maximising the number of CV minimal pairs differing in the manner of articulation of the consonants.

Parallel data was recorded for the two vocabularies described above by adult subjects with normal speaking ability in a sound proof booth. To prevent fatigue, the recording sessions were carried out on different days and short breaks were allowed during each recording session. For the TIDigits database, two male speakers (M1 and M2) and one female speaker (F1) were involved. The total amount of data for each speaker was 5.54 minutes (231 sentences) for speaker M1, 10.50 minutes (385 sentences) for speaker M2 and 8.46 minutes (308 sentences) for speaker F1. For the CV database, 958 individual consonant-vowel syllables comprising 15.76 minutes of data were recorded only for speaker M1. In each recording session, 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 in-house PMA device shown in Fig. 1, which was specifically designed to fit speaker M1's anatomy. Next, background cancellation was applied to compensate for the effect of the Earth's magnetic field on the captured articulatory data. Finally, all data was endpointed in the audio domain using an energy-based algorithm to prevent modelling the silence parts, during which the speech articulators may adopt any position.

4.2. Feature extraction

In the case of PMA, the background-cancelled, 9-channel signals are first segmented into overlapping frames using a 25 ms analysis windows with 10 ms overlap. Next, in order to combat the loss of information produced when using PMA for acquiring articulatory data and to better capture contextual phonetic information, a sliding-window approach is employed in which consecutive frames are concatenated together to form super-frames. From the sequence of PMA frames, the super-frames are formed by concatenating the ω neighbouring frames

around each particular PMA frame. Because of the high dimensionality of the resulting super-frames, the partial least squares (PLS) technique (De Jong, 1993) is applied to reduce the dimensionality of the super-frames and obtain the final PMA parameter vectors used by direct synthesis. The number of principal components retained after PLS are those explaining the 95% of the total variance.

The STRAIGHT vocoder (Kawahara et al., 1999) is used in this work for parametrising the acoustic signals. The speech parameters, which include the spectral envelope, aperiodicity spectrum and $F0$ value, are extracted from the audio signals at the same frame rate as that for the PMA signals. Then, the spectral envelopes are represented as 25-order Mel-frequency cepstral coefficients (MFCCs) (Fukada et al., 1992). As PMA does not have direct access to voicing information, the F_0 value and aperiodicity are discarded and, consequently, the reconstructed speech signals are synthesised unvoiced (i.e. as 'whispered' speech).

Finally, we apply mean and variance normalisation to the PMA and speech parameter vectors using the statistics computed for the training dataset in order to facilitate statistical training.

4.3. Evaluation of PMA-to-acoustic mapping

A 10-fold cross-validation scheme is used to evaluate the proposed techniques. Thus, the data available for each speaker is randomly divided into ten sets with equal number of utterances: nine of the sets are used for training and the remaining one for testing. This process is then repeated 10 times and results obtained for the 10 rounds are averaged.

For evaluating the accuracy of the PMA-to-acoustic mapping, both objective and subjective quality measures are employed in this paper. In the objective evaluation the Mel-cepstral distortion (MCD) measure (Kubichek, 1993) between the MFCCs extracted from the original audio signals and those estimated from PMA data is compute as follows:

$$\mathrm{MCD[dB]} = \frac{10}{\ln 10}\sqrt{2\sum_{d=1}^{D}(c_d - \hat{c}_d)^2}, \tag{30}$$

where c_d and $\hat{c}_d$ are the d-th MFCC of the original and reconstructed signal, respectively, and $D = 24$ in our case. The zero-order MFCC is not included in the above distortion since it describes the energy of the frame and in this paper we only focus on spectral-envelope reconstruction. As a distortion measure, smaller MCD results indicate better reconstruction accuracy.

For evaluating the techniques subjectively, an anonymous listening test was conducted by 25 subjects. The only requisite for participating in the test was to be adult and native English speaker. In the test, participants were asked to listen carefully to several resynthesised speech samples through a web-based interface and rate them in terms of quality, intelligibility and naturalness (see below for more details). Participants were asked to conduct the test in a quite

place while wearing head-phones, setting the volume to a comfortable hearing level. Subjects were allowed to replay the speech stimuli as many times as they wanted. The samples presented to each listener were randomly chosen from the set of available synthesised utterances.

4.4. Evaluation with the TIDigits database

4.4.1. Objective results

Fig. 2 shows contour plots for the average MCD results achieved by the MMSE and MLE conversion algorithms described in Sections 3.2.1 and 3.2.2, respectively, on the TIDigits database as a function of the number of mixture components used in the MFA model and the length of the sliding window ω used to extract the PMA parameter vectors. The results in the figure correspond to the average distortion computed for the three speakers (M1, M2 and F1) and the 10 rounds in the cross-validation scheme. As expected, the MCD results greatly improve when more mixture components are used in the MFA model, because the non-linear PMA-to-acoustic mapping is represented more accurately. In this regard, we see that for this task the optimum number of mixtures is 64. Moreover, we also see that there is a significant improvement in the conversion accuracy when longer windows are used to extract the PMA feature vectors, as this helps to reduce the uncertainty during the conversion process by taking into account more contextual information. The price for increasing the number of mixtures and the length of the sliding window is in the time it takes to convert sensor data to audio and the delay for the 'speech' to begin. By comparing the two conversion algorithms we see that the MLE-based algorithm performs slightly better than the MMSE-based algorithm on average, but the differences between both algorithms almost disappear when long context windows are used (i.e. 200 ms to 260 ms). This seems to indicate that little is gained by performing the utterance-level conversion achieved by the MLE algorithm when long context windows are used. Conversely, the short-term temporal correlations captured by the contextual windows seem to be more important for the mapping.

Fig. 3 shows example reconstructed spectrograms obtained by the MMSE and MLE methods for the utterance *six one five eight two* when a 64 mixture-component MFA and a context window of $\omega = 200$ ms are used. As can be seen, speech formants are quite accurately estimated by both methods, but the spectral details are lost as a consequence of statistical averaging carried out when estimating the MFA model, leading to the well-known problem of over-smoothing (Toda et al., 2005; Zen et al., 2009). We tried the global variance (GV) conversion algorithm proposed by Toda et al. (2005, 2007) to alleviate this problem, but the results we obtained were no better than those obtained by the MLE algorithm alone. In general, we see that the vowels and fricative consonants are well estimated. However, the stop sounds (e.g. [k] in *six* and [t] in *two* at times 0.28 s and 1.20 s, respectively) are blurred. This is due to the complex dynamics of these sounds and the limited ability of PMA to detect information about the airflow during articulation.

The MCD results obtained by the MLE conversion system with a 64-mixtures MFA model for each of the three speakers in the TIDigits database are shown in

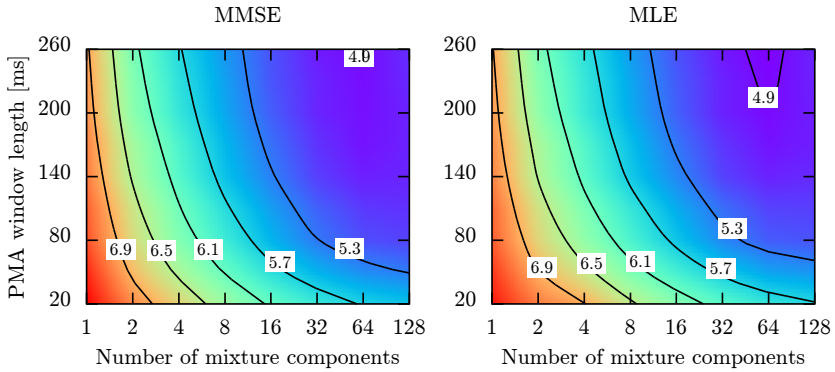

Figure 2: Contour plots for the average MCD results achieved by the MMSE-based and MLE-based conversion systems in the TIDigits database as a function of the number of mixture components in the MFA model and the length of the PMA frame window.

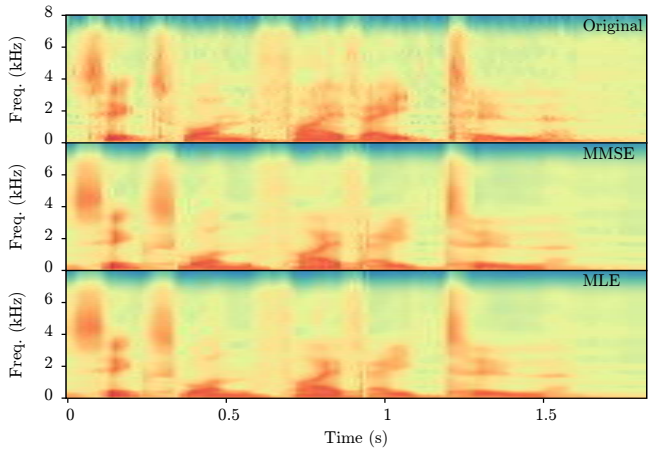

Figure 3: Examples of spectrograms of natural speech (top), MMSE-converted speech (middle), and MLE-converted speech (bottom) for the utterance "six one five eight two".

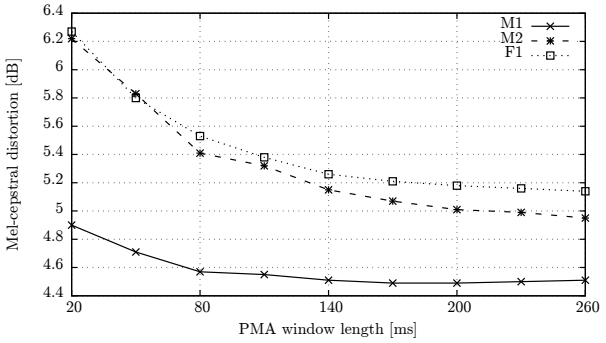

Figure 4: Performance of the MLE-based conversion system as a function of the PMA-frame window length for the three speakers in the TIDigits database.

Fig. 4. As can be seen, there is a noticeable difference between the conversion accuracy achieved for speaker M1 and that obtained for speakers M2 and F1. The reason of this behaviour, as already discussed in our previous work (Hofe et al., 2013a), is that the PMA prototype used for data recording was specifically designed for M1's anatomy.

Finally, a comparison between the GMM-based articulatory-to-acoustic conversion technique proposed by Toda et al. (2007, 2008) and our MFA-based mapping is shown Fig. 5. For a fairer comparison both approaches are evaluated using 64-mixture models and the MLE-based conversion algorithm. We evaluate our proposal using different dimensions for the latent space variable $\boldsymbol{v}$ in (5). The dimensions are 5, 10, 15, 20, and 25, the latter being the dimensionality of the speech parameter vectors. 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 difficulty of capturing the correlations between the acoustic and PMA spaces in such latent spaces. 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 proposal is more computationally efficient because of the savings of carrying out the computations in the reduced-dimension space.

4.4.2. Subjective results

We conducted a listening test to evaluate speech quality and naturalness. Speech intelligibility was not assessed for this database because informal listening revealed that the converted samples were completely intelligible[2]. Intelligi-

[2]The direct synthesis technique can produce speech of surprisingly high quality: the reader may listen to examples on the demos section of `http://www.hull.ac.uk/speech/disarm`. The identity of the speaker comes over strongly, because the mapping is trained to an individual voice.

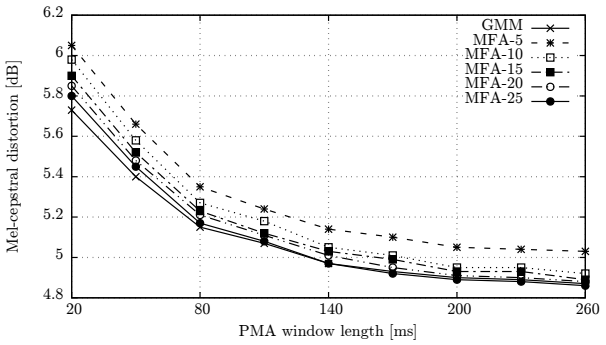

Figure 5: Comparison between the GMM-based articulatory-to-acoustic conversion of Toda et al. (2008) and our proposed method (MFA). For our proposal, the conversion accuracy is evaluated for different latent space dimensions: 5, 10, 15, 20, and 25. Results are averaged for all speakers.

bility experiments for the CV database are presented in the next section. For the TIDigits database, an XAB test was first carried out to evaluate speech quality. In the test, a speech sample was resynthesised without voicing (i.e. as whispered speech) using the STRAIGHT vocoder and presented to the listener as the reference X. Also, two different versions of the same sample converted from PMA data by our proposed method and Toda's technique were also presented to the subject in random order as A and B. Then, the listener was asked to choose which of A or B was more similar to the reference X, offering also the possibility of no preference (N/P) if both A and B sounded equally close to X. In order to evaluate the effect on perceived speech quality of the latent space dimensionality in the MFA model, different versions of the same speech sample were resynthesised from PMA data using dimensions of 5, 15 and 25 for $\boldsymbol{v}$. Each listener evaluated 8 pairs of randomly selected A-B sentences for each condition (i.e. latent dimensionality), thus making a total of 24 sample pairs evaluated per listener. For obtaining the resynthesised samples, mixtures models with 64 components and a context window of 200 ms length were employed. The conversion method chosen in both cases was the MLE algorithm.

Fig. 6 shows the results of the XAB test. It can be seen that, even when a low dimensionality is chosen for the hidden variable in the MFA model, no significant differences between speech synthesised by the MFA and GMM approaches were perceived. As the dimensionality of the latent space increases, so do the number of times listeners judge that there are no differences between both approaches. Thus, we can conclude that our approach can be seen as an efficient approximation to Toda's conversion method.

Next, we conducted a mean opinion score (MOS) test on speech naturalness. Subjects were asked to judge the naturalness of individual speech samples using a five-point scale: from 1 (completely unnatural) to 5 (completely natural). Five systems were evaluated:

- **Original**: the original speech samples with no modification.

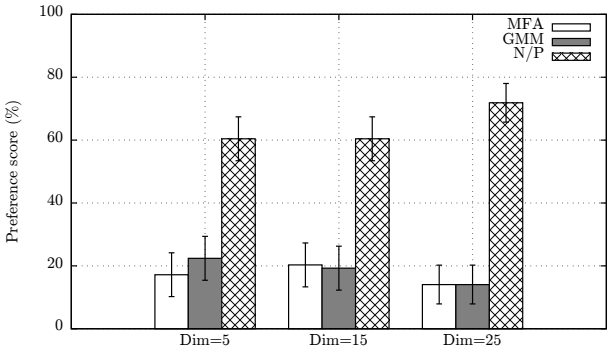

Figure 6: Results of the XAB test on speech quality. N/P indicates no preference. Error bars are plotted at the 95% confidence level.

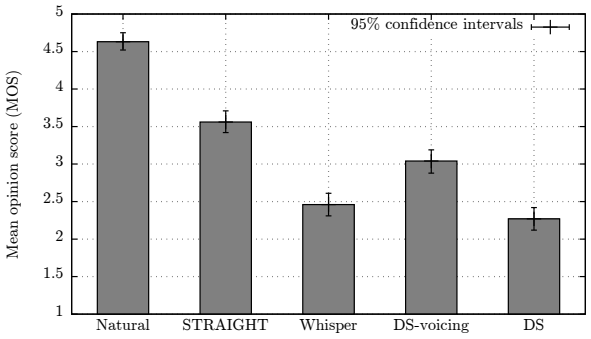

Figure 7: Results of the MOS test on speech naturalness. See text for a description of the systems evaluated.

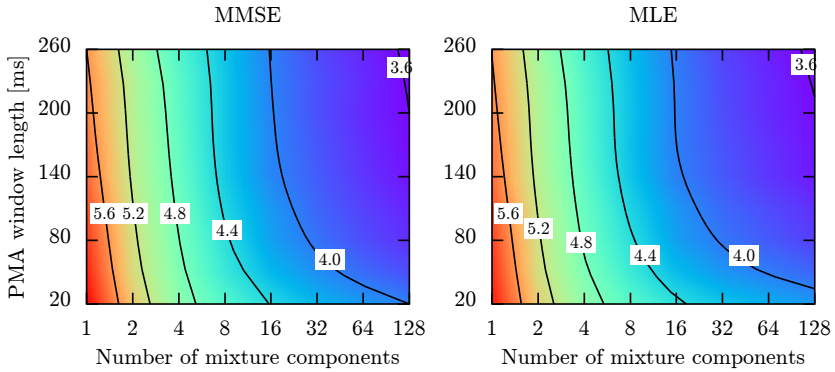

Figure 8: Average MCD results of the MMSE-based and MLE-based conversion systems in the CV database.

- **STRAIGHT**: vocoded speech using STRAIGHT.
- **Whisper**: vocoded speech using STRAIGHT but synthesised as whispered speech (i.e. without voicing).
- **DS**: direct synthesis, that is, speech converted from PMA data by our proposed MLE-based method using a 64-component MFA model.
- **DS-voicing**: same as before, but now speech is synthesised with voicing taken from the original files. In other words, the spectral envelope is estimated from PMA data by our method and the voicing information (i.e. aperiodicity, F0 and voicing decision) is taken from the original files.

Each listener evaluated 8 randomly-chosen samples for each system, thus making a total of 40 samples evaluated per subject. The order of the samples were randomised to control for order effect bias. The results of the MOS test are shown in Fig. 7. It can be seen that speech naturalness is degraded after the analysis-synthesis process carried out by STRAIGHT due to artefacts introduced by this vocoder. This degradation is further amplified when speech stimuli are resynthesised with no voicing and, hence, listeners considered whispered speech quite unnatural. Surprisingly, listeners judge that the naturalness of DS speech is not significantly lower than that of whispered speech. This is a exciting result if we consider that the whispered samples are directly obtained from the original, natural speech samples, while the DS samples are obtained through an error-prone conversion process such as direct synthesis. We also see in the figure that when the DS samples are synthesised with voicing, their naturalness is significantly enhanced, outperforming even the whisper process.

4.5. Evaluation with the CV database

4.5.1. Objective results

The objective MCD results obtained for the CV database are shown in Fig. 8. Again, we see that the MMSE and MLE conversion algorithms perform

equivalently. Compared to the results obtained for the TIDigits database in Fig. 2, we see that the MCD figures for the CV database are slightly better. On the one hand, a possible explanation for this is that the utterances in the CV corpus consist of single-syllable words, while those in the TIDigits database have multiple words. On the other hand, the CV database contains only data for speaker M1, who is the user for whom the PMA prototype was designed. For this database we see that the best results are obtained when using a 128-mixture MFA and a PMA frame window spanning 240 ms. This is the set-up we will use in the rest of this section.

Next, to study in-depth the errors made by the conversion algorithms, we performed phone-level comparisons between the reconstructed MFCCs obtained from PMA data and the original MFCCs extracted from the ground-truth acoustic signals. We did that by first force-aligning the word-level transcriptions of the ground-truth signals using a context-dependent speech recogniser adapted to the speaker's voice. Then, the timing information included in the resultant phonetic transcriptions was used to segment the acoustic signals into phones and the MCD measure was computed for each individual phone. In doing so we assumed that the ground-truth and reconstructed signals were synchronous. Furthermore, only the stable part of the phones was used for computing the MCD distortion in order to avoid considering coarticulation effects. We assumed that the stable part corresponds to the 50% central segment of the phone.

The detailed MCD results obtained for each phone and conversion method are shown in Fig. 9. The results are presented as box plots, each box showing the first three quartiles (i.e. 25%, 50% and 75%) of the error, while the whiskers extending up to 1.5 times the interquartile range (i.e. Q3-Q1). We see that, again, the results obtained by both conversion algorithms are very similar. From the point of view of the different phones, we can made the following observations. Firstly, it can be observed that the vowels are quite accurately reconstructed in both cases. The consonants, however, are not always consistently well reconstructed. In general, we see that direct synthesis perfoms poorly in reconstructing the sounds articulated in the back of the mouth (i.e. [k h]), which the current PMA prototype is not capturing well as no magnet is placed in this area. Other consonant sounds for which the reconstruction error is higher than the mean are the plosives [p t k]. In this case, as commented above, the problem lies in the difficulty in modelling the dynamics of these sounds (i.e. a hold phase where the vocal tract is closed followed by a short burst in which the air is suddenly released), together with the limitations of PMA for accessing airflow information. Apart from these problems, the ability of the direct synthesis technique to synthesise accurately phones sharing the same place of articulation but a different manner (e.g. [n l r]) is remarkable. It might be that contextual information such as coarticulation is well captured by PMA, helping to reduce the uncertainty associated with the articulatory-to-acoustic mapping.

4.5.2. Subjective results

We evaluated the intelligibility of the resynthesised speech samples by conducting a listening test involving 25 human subjects. In the test, the subjects

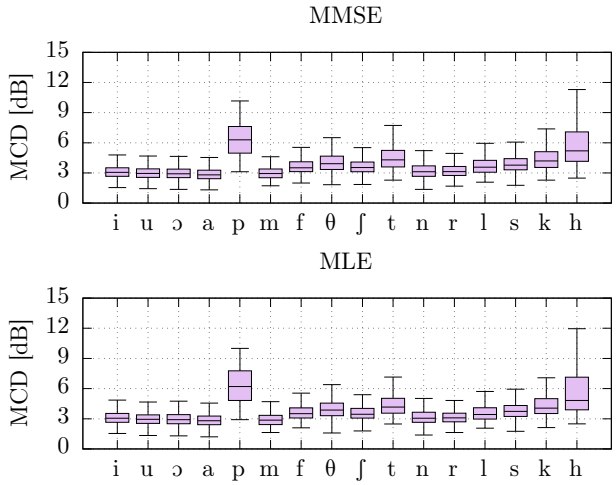

Figure 9: Detailed MCD results for the different phones in the CV database.

were asked to type the syllable they heard when presented with a speech stimulus. No further instructions were given to the subjects apart from that the stimuli were comprised of consonant-vowel syllables and that they might be nonsense words. In particular, the consonants and vowels used to construct the syllables were not revealed to the subjects. Because of this freedom, it was later found during the analysis of the results that some subjects were biased in choosing correct words in English with similar pronunciation (e.g. food for [fu] or zoo for [su]) when uncertain about the transcription of a particular stimulus. Despite this, we preferred this form for evaluating speech intelligibility over other forms (e.g. asking the subjects to choose the transcriptions from a list) because we considered it provides us with a better measure of intelligibility.

In the test, each subject was presented with 24 stimuli chosen at random, consisting of 24 different syllables formed by combining the 12 consonants with two random, but different, vowels. Later, when analysing the responses of the subjects, the original syllables and the subject responses were transcribed phonetically in order to account for possible homophones of the same syllable (e.g. to, too or two). Then, subject responses were compared with the transcriptions of the original stimuli to see whether they match.

Table 2 the accuracy results obtained for each CV syllable, that is, the percentage of each individual syllable which was correctly transcribed by the subjects in the listening test. In addition, the overall average (AVG) results for each consonant and vowel are also shown. With an overall accuracy of 68%, it can be seen that the results are very diverse, ranging from syllables that are always well transcribed (e.g. [fa], [tu], [lɔ]) to those such as [ku] that were not correctly transcribed on any occasion. Speaking roughly, we see that the worst results are obtained for the syllables containing plosive consonants (i.e. [p t k]) and those in which the consonant sound is articulated in the back of the mouth

	p	m	f	θ	ʃ	t	n	r	l	s	k	h	**AVG**
i	71.43	87.50	81.25	61.54	90.91	44.44	50.00	61.54	81.82	100.00	7.14	7.69	61.07
u	23.08	56.25	85.71	30.00	100.00	100.00	92.86	0.00	100.00	78.57	0.00	63.64	64.86
ɔ	66.67	92.31	84.62	58.33	85.71	56.25	100.00	57.14	100.00	90.91	50.00	50.00	72.48
a	64.29	84.62	100.00	33.33	91.67	100.00	80.00	60.00	75.00	91.67	66.67	58.33	73.38
AVG	56.00	78.00	86.00	46.00	92.00	76.00	78.00	50.00	88.00	90.00	32.00	44.00	68.00

Table 2: Results of the intelligibility test for the CV database. For each syllable, the transcription accuracy in percent of the speech stimuli by human subjects is shown. AVG correspond to the average accuracy for each phone.

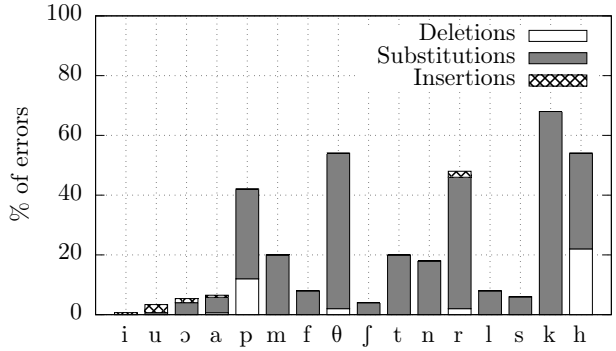

Figure 10: Analysis of the type of errors made by the human subjects in the intelligibility test for each phone.

(i.e. [k h]), as in the objective results shown in Fig. 9. Other syllables for which poor results are obtained are those starting with the consonants [θ] and [r]. For [θ], as will be further discussed later, a large percentage of the errors (70.37% of them) correspond to confusions with [f]. For [r], more than half of the total errors (58.33%) are due to [r] being confused with [l].

Next, we conducted an analysis of the type of errors made by listeners when transcribing the speech stimuli in terms of deletions, insertions and substitutions. To perform this analysis, the phonetic transcriptions of the listener responses were manually aligned with the reference consonant-vowel transcriptions of the stimuli and the number of deletion, insertion and substitutions errors were counted for each phone. For example, [pa] would count as an substitution error for the vowel [ɔ] in the analysis if the original stimuli was the syllable [pɔ]. Similarly, [u] would count as a deletion error of the consonant if the original syllable was [hu]. Finally, we count [blu] and [taɔ] as insertions errors for the consonant and vowel, respectively, if the original syllables were [bu] and [ta].

The results of the error analysis are shown in Fig. 10. Again, as in the objective results shown in Fig. 9, it can be seen that far fewer errors are made for the vowels than for the consonants. In the case of the consonants, big differences exist between different groups of consonants: [p θ r k h] are erroneously transcribed more than 40% of times, other consonants as [f ʃ l s] are quite accurately transcribed (less than 10% of errors), while the renaming consonants

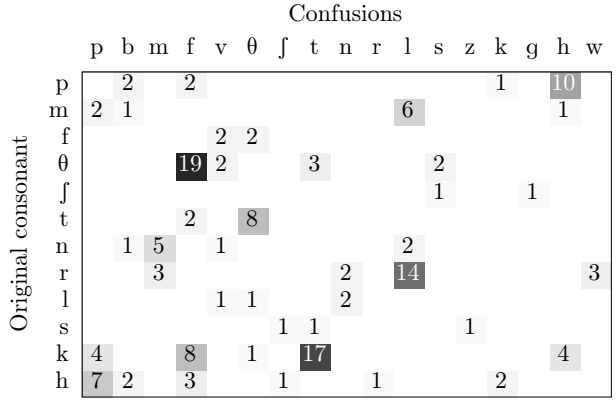

Figure 11: Confusion matrix with the substitutions errors for each consonant. Each cell provides the total number of confusions for each pair of consonants. Note that the axis labels are different because the subjects' responses sometimes include phones that are not in the phone repertory in Table 1.

[m t n] are in the middle of the table, with transcription errors approximately equal to 20%. Regarding the type of errors, it can be seen that most of them correspond to substitutions. These will be discussed below in more detail. For the vowels, a large percentage of the errors are also due to insertions. These errors correspond to phones, usually vowels, identified by the subjects at the end of the syllables that are not present in the original stimuli. The origin of these errors may be due to the endpointing algorithm, which leaves a small fraction of the initial and final silences in the utterances. During the silences the speech articulators may adopt any position and, hence, it is possible that the short silences left by the endpointing algorithm are synthesised as audible speech. For the consonants, the second most frequent errors are the deletions. These errors might be due to the phone being omitted because its duration is very short, as in the stop [p], or PMA not capturing enough information to distinguish the phone, as in [h].

For the purpose of better understanding the intelligibility results shown in Fig. 10, it is illustrative to compare those results with the objective results achieved by the MLE-based system in Fig. 9. We see that both figures are visually similar, thus indicating that, as can be expected, the phones with the highest MCD values are more likely to be mis-recognised. This visual analysis is confirmed when we compute the Pearson correlation coefficient between the results in both figures: the correlation between the average MCD for each phone in Fig. 9 and the sum of errors in Fig. 10 is $\rho = 0.60$. This correlation is further increased up to $\rho = 0.75$ when the phones [θ r] are omitted from the analysis: these are the ones for which the intelligibility and MCD results are less similar. Below, we suggest why the intelligibility and MCD results differ for these phones.

The last statistical analysis we performed on the data from the intelligibility test was an analysis of the consonant confusions due to the substitution errors

shown in Fig. 10. The results of this analysis are presented in Fig. 11 as a confusion matrix. The interpretation of the matrix is as follows. Rows correspond to the the actual consonant in the original stimuli. Columns correspond to substitutions errors, that is, the consonant that the subjects reported hearing. Finally, the cells contain the number of confusions for each pair of consonants (e.g. [k] is confused with [h] 4 times).

In general, we can see that the confusions are more uniformly distributed for the consonants that have fewer errors in Fig. 10, such as [m f l s]. On the other hand, the consonants [p θ r k h], which are the ones that are more frequently confused, posses a less uniform distribution in the errors, being those consonants often confused with another particular consonant. For example, it can be seen that, among the 15 substitutions errors for [p], 67.7% of them (10) correspond to confusions with [h]. By listening to the erroneously transcribed speech stimuli for [p] we realise that the problem is that the onset on [p] is not accurately estimated and is oversmoothed, so that this consonant is easily confused with [h] when accompanied with a vowel. Another interesting confusion is that of [θ] with [f]. Although articulated differently, both consonant are fricatives and are acoustically similar, which might explain why they are so often confused. The confusion of [r] with [l] may be due to the two phones being acoustically similar because they are articulated in roughly the same position of the mouth and both are approximant consonants. Regarding the confusions for [k] and [h], it is hard to extract any meaningful conclusions about the confusions of these consonants, as no magnet is currently attached in the velar and glottal areas in the PMA prototype. We see that, for example, [k] is often confused with [t] and [h] with [p]. For the confusions of [k] with [t], a possible explanation is that both are plosive consonants with similar acoustics. In the case of [h], it might be that PMA is picking some information from the tongue or lips that is similar to that captured when [p] is articulated. Finally, it is worth mentioning that the pattern of confusions in Fig. 11 is somewhat similar to that obtained by means of speech recognition experiments in other SSIs Hueber et al. (2008); Wand and Schultz (2011).

5. Discussion

The results of the last section have clearly demonstrated the feasibility of synthesising audible speech from articulator movement data without the need of an intermediate recognition step. As opposed to our previous work (Fagan et al., 2008; Gilbert et al., 2010; Hofe et al., 2013b,a; Cheah et al., 2015), in which the PMA data were first decoded using an ASR system trained on articulatory data and then audible speech was optionally generated using a TTS system, in the proposed direct synthesis technique a learned transformation is directly applied to the PMA data to obtain the final acoustic signal. As discussed in the introduction, this has significant potential advantages compared to the recognise-then-synthesise approach in terms of the ease with which the technique could be extended to, accents and speakers, real-time implementation, and the amount of data needed to train the system.

For evaluating the proposed system, two parallel PMA-and-speech databases have been employed. For the TIDigits database, informal listening revealed that direct synthesis was able to produce completely intelligible speech. Furthermore, results from the listening test summarised in Fig. 7 demonstrate that the naturalness of direct-synthesis speech is on a par with that of 'whispered' natural speech (i.e. natural speech synthesised without voicing). However, as also shown in the same figure, the naturalness of direct synthesis speech is still far from that of natural speech. From the analysis of the results of the MOS test, it seems that the reason why human listeners consider direct synthesis speech unnatural is because it lacks prosodic features related to the intonation such as the pitch and voicing, although it does incorporate other prosodic features related to the tempo and the stress.

The results obtained for the CV database have allowed us to shed some light on the intelligibility of reconstructing phonetically-rich speech. Despite the objective results obtained for the TIDigits and CV databases on Figs. 2 and 8, respectively, being similar, we have seen that the intelligibility of converted speech in both cases is very different. For the TIDigits database, as already mentioned, informal listening concluded that converted speech was highly intelligible, while the results in Table 2 demonstrates that this is not necessarily true for other vocabularies. In particular, for the CV database, 32% of the speech stimuli were not correctly recognised by the subjects of the listening test. We can think of two reasons that might explain these differences. First of all, intelligibility and understanding of reconstructed speech can greatly benefit from the knowledge of any *a priori* information about the topic being spoken. In the case of the TIDigits database, it is the knowledge of the vocabulary which helps human listeners to disambiguate between words when uncertain about their identity. This information, however, was not made available on purpose to subjects in the CV database. Thus, subjects sometimes struggle to recognise the correct syllable when direct synthesis fails to synthesise it with enough accuracy. In a natural conversation, however, it is expected that we will be somehow in the middle between these two extremes: some words could be predicted with high accuracy from the context, while for others the user will have to rely on the acoustic information solely. A second reason which might explain the differences between both databases is the phonetic complexity. The CV vocabulary was specifically designed to evaluate direct synthesis under a phonetically rich vocabulary with a high number of minimal pairs, while the TIDigits vocabulary was not. This, together with the limitations of the current PMA prototype for detecting certain aspect of speech articulation, make the CV database a harder material for direct synthesis.

The results have also provided us with clues about the performance of direct synthesis for different speakers. As shown in Fig. 4 and also reported in our previous work (Hofe et al., 2013a), direct synthesis performs significantly better for speaker M1, who is the one the PMA device was designed for and also is the more experienced user. Nevertheless, these differences among speakers could be lessened in future by carefully controlling the above two factors, i.e. the experience of the user in using PMA and a user-adapted design of PMA device

suited to her/his anatomy.

Finally, it is also worth commenting about the performance achieved by the two conversion algorithms introduced in Section 3.2: MMSE and MLE. As discussed in that section, the main motivation behind the more computationally-complex algorithm MLE is to achieve better accuracy in speech reconstruction by taking into account the temporal dynamics of the speech parameters in the PMA-to-acoustic mapping. However, in the light of the results in Figs. 2 and 8, this extra complexity does not seem to be justified when compared with the simpler MMSE algorithm. Thus, results show that MMSE performs on a par with that of MLE when long-contextual windows are employed for extracting the articulatory parameters. As previously discussed, these windows might provide the conversion algorithm with contextual information about the phone being spoken which, in turn, helps to reduce the uncertainty in the PMA-to-acoustic mapping. Conversely, long-term correlations such as those exploited by the MLE method seems to be of little help during the conversion procedure.

6. Conclusions

In this paper we have presented a system for producing audible speech from speech-articulator movement captured using a technique known as permanent magnet articulography. We have successfully demonstrated that the proposed technique is able to generate speech of sufficient intelligibility and quality for some vocabularies. This is a big step in our long-term goal of developing a discrete and reliable SSI that will ultimately allow laryngectomees to recover their voice. However, before the proposed technique can be applied in an realistic treatment scenario, a number of questions need to be addressed. A first question is related to the capabilities of PMA for modelling the vocal tract. As demonstrated by the results presented in this paper, the current prototype has some limitations for detecting certain aspects of speech articulation (e.g. the manner of articulation, voicing and the phones articulated at the back of the mouth). A second question relates to the quality of reconstructed speech. This includes improving its naturalness by also recovering the prosodic information (i.e. voicing information and stress) and also improving the conversion accuracy for a large vocabulary. Finally, another important question concerns the practical implementation of the proposed speech restoration system to patients who have already lost their voice and for whom it is impossible to record the parallel data used to train the system. Solutions to all these questions are currently under development.

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

Ananthakrishnan, G., Engwall, O., Neiberg, D., Dec. 2012. Exploring the predictability of non-unique acoustic-to-articulatory mappings. IEEE Trans. Audio Speech Lang. Process. 20 (10), 2672–2682.

Anderson, T. W., 2003. An introduction to multivariate statistical analysis, 3rd Edition. Wiley-Interscience.

Atal, B. S., Chang, J. J., Mathews, M. V., Tukey, J. W., 1978. Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique. J. Ac. Soc. Am. 63 (5), 1535–1555.

Badin, P., Bailly, G., Revret, L., Baciu, M., Segebarth, C., Savariaux, C., Jul. 2002. Three-dimensional linear articulatory modeling of tongue, lips and face, based on MRI and video images. Journal of Phonetics 30 (3), 533–553.

Birkholz, P., Jackel, D., 2003. A three-dimensional model of the vocal tract for speech synthesis. In: Proc. 15th International congress of phonetic sciences. pp. 2597–2600.

Birkholz, P., Steiner, I., Breuer, S., 2008. Control concepts for articulatory speech synthesis. In: Proc. 6th ISCA Workshop on Speech Synthesis. pp. 5–10.

Bishop, C. M., 2006. Pattern recognition and machine learning. Springer.

Brigham, K., Vijaya Kumar, B. V. K., Jun. 2010. Imagined speech classification with EEG signals for silent communication: a preliminary investigation into synthetic telepathy. In: Proc. 4th International Conference on Bioinformatics and Biomedical Engineering. pp. 1–4.

Brumberg, J. S., Nieto-Castanon, A., Kennedy, P. R., Guenther, F. H., Apr. 2010. Brain-computer interfaces for speech communication. Speech Commun. 52 (4), 367–379.

Brumberg, J. S., Wright, E. J., Andreasen, D. S., Guenther, F. H., Kennedy, P. R., May 2011. Classification of intended phoneme production from chronic intracortical microelectrode recordings in speech-motor cortex. Frontiers in neuroscience 5, 1–12.

Cai, J., Hueber, T., Manitsaris, S., Roussel, P., Crevier-Buchman, L., Stone, M., Pillot-Loiseau, C., Chollet, G., Dreyfus, G., Denby, B., 2013. Vocal tract imaging system for post-laryngectomy voice replacement. In: Proc. IEEE International Instrumentation and Measurement Technology Conference (I2MTC). pp. 676–680.

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

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., Brumberg, J., Apr. 2010. Silent speech interfaces. Speech Commun. 52 (4), 270–287.

Deng, Y., Heaton, J. T., Meltzner, G. S., 2014. Towards a practical silent speech recognition system. In: Proc. Interspeech. pp. 1164–1168.

Desai, S., Raghavendra, E. V., Yegnanarayana, B., Black, A. W., Prahallad, K., 2009. Voice conversion using artificial neural networks. In: Proc. ICASSP. pp. 3893–3896.

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

Fukada, T., Tokuda, K., Kobayashi, T., Imai, S., 1992. An adaptive algorithm for Mel-cepstral analysis of speech. In: Proc. ICASSP. pp. 137–140.

Ghahramani, Z., Hinton, G. E., 1996. The EM algorithm for mixtures of factor analyzers. Tech. Rep. CRG-TR-96-1, University of Toronto.

Gilbert, J. M., Rybchenko, S. I., Hofe, R., Ell, S. R., Fagan, M. J., Moore, R. K., 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., Green, P. D., 2014. Analysis of phonetic similarity in a silent speech interface based on permanent magnetic articulography. In: Proc. Interspeech. pp. 1018–1022.

Gonzalez, J. A., Green, P. D., Moore, R. K., Cheah, L. A., Gilbert, J. M., 2015. A non-parametric articulatory-to-acoustic conversion system for silent speech using shared gaussian process dynamical models. In: Proc. UK Speech. p. 11.

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

Hofe, R., Bai, J., Cheah, L. A., Ell, S. R., Gilbert, J. M., Moore, R. K., Green, P. D., 2013a. Performance of the MVOCA silent speech interface across multiple speakers. In: Proc. Interspeech. pp. 1140–1143.

Hofe, R., Ell, S. R., Fagan, M. J., Gilbert, J. M., Green, P. D., Moore, R. K., Rybchenko, S. I., 2011. Speech synthesis parameter generation for the assistive silent speech interface MVOCA. In: Proc. Interspeech. pp. 3009–3012.

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

Hueber, T., Bailly, G., Denby, B., 2012. Continuous articulatory-to-acoustic mapping using phone-baseda trajectory HMM for a silent speech interface. In: Proc. Interspeech. pp. 723–726.

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

Hueber, T., Chollet, G., Denby, B., Dreyfus, G., Stone, M., 2008. Phone recognition from ultrasound and optical video sequences for a silent speech interface. In: Proc. Interspeech. pp. 2032–2035.

Janke, M., Wand, M., Nakamura, K., Schultz, T., 2012. Further investigations on EMG-to-speech conversion. In: Proc. ICASSP. pp. 365–368.

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

Kawahara, H., Masuda-Katsuse, I., De Cheveigne, A., Apr. 1999. Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds. Speech communication 27 (3), 187–207.

Kubichek, R., 1993. Mel-cepstral distance measure for objective speech quality assessment. In: Proc. IEEE Pacific Rim Conference on Communications, Computers and Signal Processing. pp. 125–128.

Leonard, R., 1984. A database for speaker-independent digit recognition. In: Proc. ICASSP. pp. 328–331.

Maeda, S., 1982. A digital simulation method of the vocal-tract system. Speech Commun. 1 (3), 199–229.

Matthews, I., Cootes, T. F., Bangham, J. A., Cox, S., Harvey, R., Feb. 2002. Extraction of visual features for lipreading. IEEE Trans. Pattern Anal. Mach. Intell. 24 (2), 198–213.

Moriguchi, T., Toda, T., Sano, M., Sato, H., Neubig, G., Sakti, S., Nakamura, S., 2013. A digital signal processor implementation of silent/electrolaryngeal speech enhancement based on real-time statistical voice conversion. In: Proc. Interspeech. pp. 3072–3076.

Nakamura, K., Toda, T., Saruwatari, H., Shikano, K., Jan. 2012. Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech. Speech Commun. 54 (1), 134–146.

Neiberg, D., Ananthakrishnan, G., Engwall, O., 2008. The acoustic to articulation mapping: non-linear or non-unique? In: Proc. Interspeech. pp. 1485–1488.

Petajan, E., Bischoff, B., Bodoff, D., Brooke, N. M., 1988. An improved automatic lipreading system to enhance speech recognition. In: Proc. SIGCHI conference on Human factors in computing systems. pp. 19–25.

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

Qin, C., Carreira-Perpiñán, M. Á., 2007. An empirical investigation of the nonuniqueness in the acoustic-to-articulatory mapping. In: Proc. Interspeech. pp. 74–77.

Rubin, P., Baer, T., Mermelstein, P., 1981. An articulatory synthesizer for perceptual research. J. Ac. Soc. Am. 70 (2), 321–328.

Schroeter, J., Sondhi, M. M., 1994. Techniques for estimating vocal-tract shapes from the speech signal. IEEE Trans. Speech Audio Process. 2 (1), 133–150.

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

Stylianou, Y., Cappe, O., Moulines, E., Mar. 1998. Continuous probabilistic transform for voice conversion. IEEE Trans. Speech Audio Process. 6 (2), 131–142.

Toda, T., Black, A. W., Tokuda, K., 2005. Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter. In: Proc. ICASSP. pp. 9–12.

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

Toda, T., Muramatsu, T., Banno, H., 2012a. Implementation of computationally efficient real-time voice conversion. In: Proc. Interspeech. pp. 94–97.

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

Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., Kitamura, T., 2000. Speech parameter generation algorithms for HMM-based speech synthesis. In: Proc. ICASSP. pp. 1315–1318.

Toth, A. R., Kalgaonkar, K., Raj, B., Ezzat, T., 2010. Synthesizing speech from Doppler signals. In: Proc. ICASSP. pp. 4638–4641.

Toutios, A., Margaritis, K. G., 2005. A support vector approach to the acoustic-to-articulatory mapping. In: Proc. Interspeech. pp. 3221–3224.

Toutios, A., Narayanan, S., 2013. Articulatory synthesis of french connected speech from EMA data. In: Proc. Interspeech. pp. 2738–2742.

Toutios, A., Ouni, S., Laprie, Y., 2011. Estimating the control parameters of an articulatory model from electromagnetic articulograph data. J. Ac. Soc. Am. 129 (5), 3245–3257.

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

Wand, M., Schultz, T., 2011. Analysis of phone confusion in EMG-based speech recognition. In: Proc. ICASSP. pp. 757–760.

Wester, M., 2006. Unspoken speech: speech recognition based on electroencephalography. Master's thesis, Universität Karlsruhe.

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

Zen, H., Tokuda, K., Black, A. W., Nov. 2009. Statistical parametric speech synthesis. Speech Commun. 51 (11), 1039–1064.