Analysis of Phonetic Similarity in a Silent Speech Interface based on Permanent Magnetic Articulography

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

[1]Department of Computer Science, The University of Sheffield, Sheffield, UK
[2]School of Engineering, University of Hull, Kingston upon Hull, UK
[3]Hull and East Yorkshire Hospitals Trust, Castle Hill Hospital, Cottingham, UK
j.gonzalez@sheffield.ac.uk

Abstract

This paper investigates the potential of a silent speech interface (SSI) based on Permanent Magnetic Articulography (PMA) to be used in applications involving unconstrained, phonetically rich speech. In previous work the SSI was evaluated on isolated-word and connected-digits recognition tasks with promising results. Furthermore, it was shown that PMA data is enough to distinguish between minimal pairs of consonants with the same manner and place of articulation but different voicing. The study presented in this paper extends previous work by investigating to what extent information about the speech production process can be recovered from PMA data. In particular, the three main aspects of speech production are investigated here: voicing, place of articulation and manner of articulation. The results show that PMA achieves comparable accuracy to using the audio signal for discriminating the place of articulation of consonants, but it provides little information regarding the voicing and manner of articulation.

Index Terms: silent speech interfaces, assistive speech technology, permanent magnetic articulography, phonetic similarity

1. Introduction

The aim of the DiSArM (Digital Speech Recovery from Articulator Movement) project [1] is to restore the ability to communicate to patients who have undergone laryngectomy, normally following cancer. The technology developed in DiSArM is built upon PMA [2], a technique for monitoring the movements of the intact articulators (e.g. lips and tongue) by sensing the changes in magnetic field produced by a number of small magnets attached to these articulators.

In previous work [2, 3] the validity of this approach for speech recognition purposes on simple recognition tasks was shown. In particular, in [3], two different recognition tasks were evaluated. First, isolated word recognition using PMA data on a vocabulary consisting of 57 words[1] was reported to yield similar word accuracy results to recognition using audio (98.80 % vs. 99.70 %, respectively). The second recognition task was connected digits. In this case the word accuracy results for PMA and audio were 95.90 % and 96.30 %, respectively.

More recently, the performance across several speakers was investigated in [4]. The vocabulary in this case consisted of 71 single-syllable CVC words sharing the same central vowel. Two main conclusions were drawn in this study. First, it was shown that the performance of PMA-based speech recognition was high even for speakers unfamiliar with PMA technology. Second, more surprisingly, it was also shown that it is possible to distinguish between voiced and unvoiced consonant pairs using PMA data. This result was somehow unexpected since PMA does not have direct access to voicing information.

Finally, a feasibility study of direct speech synthesis from PMA data without an intermediate recognition step was presented in [5]. This study showed that it is possible to predict the speech formants F1/F2 from PMA data, thus opening up the prospect of a device that will allow laryngectomy patients to recover their own voice without a recognition stage.

In this paper we build upon our previous work investigating the potential of PMA to be used in applications involving phonetically rich speech (e.g. large-vocabulary speech recognition or speech synthesis). In particular, we are interested in determining to what extent aspects of the speech production process such as the voicing, place and manner of articulation can be identified from PMA data. Here, this problem is approached by performing systematic analysis involving synchronous recordings of audio and PMA data on a phonetically rich corpus. Our hypothesis is that, being a technology that does not have access to airflow information, PMA will be better at discriminating the place rather than the manner of articulation or voicing of the phones. The ultimate goal of this analysis is to study the feasibility of using PMA as a means of restoring the ability to communicate to laryngectomees.

This paper is organized as follows. The next section briefly describes the functional aspects of PMA. In Section 3, the experimental methodology used to assess the ability of PMA for discriminating between phones is presented. The experimental results are shown in section 4. Finally, the main conclusions are drawn in section 5.

2. Permanent Magnetic Articulography

PMA is a technique for capturing the movements of the vocal apparatus during speech. The technique is based on sensing the changes in the magnetic field generated by the movement of small magnets attached to the lips and tongue as the speaker 'mouths'. In the current set-up there are a total of six magnets, 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 three spatial components (x, y, z) of the magnetic field generated by the magnets is then acquired by four triaxial mag-

[1]The vocabulary consisted of the 10 English digits plus a word list intended to cover the ARPAbet phonetic inventory, with a high number of minimal pairs.

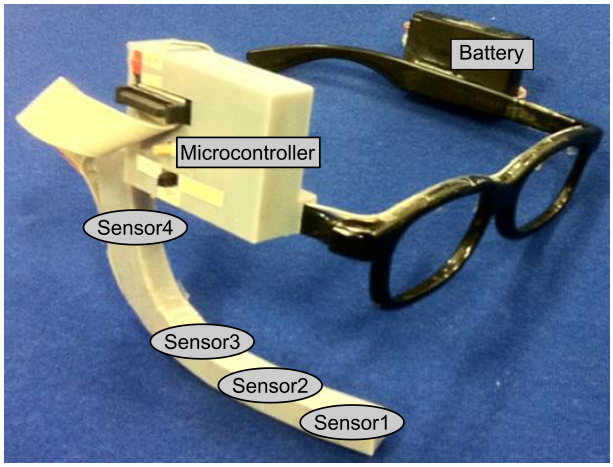

Figure 1: Sensor frame used to record the magnetic field generated by the movements of the magnets attached to the lips and tongue. Three channels of data with the (x, y, z) components of the magnetic field are acquired by each sensor.

netic sensors mounted on a rigid frame, as illustrated in figure 1. Only three of the sensors are used for capturing articulatory movements - those closer to the mouth. The remaining sensor, Sensor4 in the figure, is used for background cancellation, i.e. compensating the effects of the earth's magnetic field. Hence, in total, 9 channels of data are available for monitoring the movements of the vocal tract. In must be pointed out that contrary to other mechanisms for capturing articulatory data, PMA does not provide the exact position of the magnets attached to the articulators, as the magnetic field arriving at each sensor is a combination of the fields generated by all the magnets. More details of PMA can be found in [2].

3. Methodology

3.1. Data acquisition

A set of 420 utterances was selected at random from the CMU Arctic corpus of phonetically balanced sentences [6] and uttered by a male native British English subject with a normal speaking ability. To prevent speaker fatigue, the acquisition was split into eight subsets with 60 sentences each, allowing short breaks in between. After endpointing, the total amount of data recorded was approximately 22 minutes.

The audio signal and 9-channel PMA data were recorded simultaneously at sampling frequencies of 16 kHz and 100 Hz, respectively. Later, synchronisation of the two streams was carried out to compensate for small deviations from the ideal sampling frequencies of the analog-to-digital converters. This was done automatically using a timing alignment mechanism in which start-stop markers were generated and recorded simultaneously with both data streams (audio and PMA data). These markers were then used to realign both streams, thus minimising any potential timing error.

Because the speaker's head was not restrained during the recordings, background cancellation was applied when necessary to counter the possibility that the earth's magnetic field could distort the data recorded by the sensors during head movements. To do this, the data recorded by Sensor4 in figure 1 was used as the reference for removing the influence of earth's magnetic field.

3.2. Feature extraction

In order to reduce the dimensionality of the data and facilitate statistical modelling, feature extraction was carried out on both the audio and PMA signals. For the audio, 25 Mel-frequency cepstral coefficients (MFCCs) were extracted from 25 ms analysis windows with 10 ms overlap. The PMA data was oversampled from 100 Hz to 200 Hz to be compatible with the 25 ms analysis window used for the audio. Then, 45-dimensional feature vectors were computed by taking a 5-dimensional analysis window (25 ms) over the 9 channels with 2 samples overlap (10 ms). Next, principal components analysis (PCA) [7] was applied to reduce the dimensionality of the feature vectors from 45 to 25. The final dimensionality was chosen to match that of the audio features. Finally, after feature extraction, the MFCC and PMA features were normalized in mean and variance for a fairer comparison between them when computing the phonetic similarity for both streams.

3.3. Analysis of phonetic similarity

In order to investigate the suitability of PMA for applications involving phonetically rich speech, an analysis of phonetic similarity for PMA data is reported in this paper. In brief, the similarity between each pair of phones is measured by computing the similarity between the probability distributions estimated for the phones in the PMA feature space. The following procedure was used to obtain these distributions. First, a speaker-independent recognition model based on cross-word triphones was adapted to the speaker's voice using 71.2 minutes of speech material recorded for the speaker. The adaptation technique used was maximum a posteriori (MAP) [8] and it was performed using HTK [9]. Then, the adapted model was used to obtain force-aligned phone-level transcriptions of the audio signals.

Although it is known that, due to anticipatory behaviour of the articulators, the acoustic and articulatory signals are delayed respect to the other, this delay is usually small (typically between 10-50 ms) and it depends on the specific articulator and the phone being articulated (see e.g. [10, 11]). For these reasons, the audio and PMA features are assumed to be synchronous in this work. Hence, the force-aligned transcriptions derived for the audio signals were also used to gather the PMA features belonging to each phone. Next, a probability distribution was estimated for each phone in the PMA feature space. A non-parametric kernel density estimation (KDE) approach was adopted to estimate the distributions. The reason for choosing the KDE approach instead of a parametric one (e.g. Gaussian mixture models) was to avoid having to decide the number of clusters (i.e. Gaussians) in the distribution. Thus, the KDE model is expected to better represent the feature space.

Let $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ be the PMA feature vectors belonging to a given phone. Then, the kernel density estimator of the distribution for the phone is given by

$$p(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} K(\boldsymbol{x} - \boldsymbol{x}_i), \tag{1}$$

where $K(\cdot)$ is a kernel function. In this paper an isotropic Gaussian kernel with bandwidth σ is chosen, i.e.

$$K(\boldsymbol{x}) = (2\pi)^{-\frac{d}{2}} \left|\sigma^2 \boldsymbol{I}\right|^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2} \boldsymbol{x}^T \boldsymbol{x}\right). \tag{2}$$

Finally, for each pair of phones, the symmetric Kullback-Leibler (SKL) divergence is computed in order to measure the

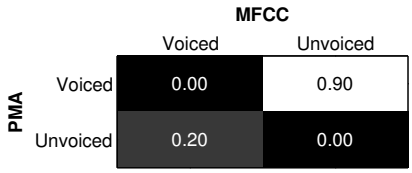

Figure 2: Comparison of the SKL divergences obtained for the MFCC and PMA features with respect to the voicing. The upper triangular part of the matrix corresponds to the results obtained for the MFCCs, and the lower triangular part are those computed for the PMA features.

similarity between their distributions. Let $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$ be the distributions for the two phones. Then, the SKL divergence between the distributions is computed as $C(p,q) = D(p\|q)+D(q\|p)$, where $D(q\|p)$ is the Kullback-Leibler (KL) divergence between $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$ and is given by

$$D(p\|q) = \int p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} d\boldsymbol{x}, \tag{3}$$

with $D(p\|q) \geq 0$ for all distributions p, q and $D(p\|q) = 0$ only if $p = q$.

Unfortunately, the KL divergence between two Gaussian mixture models or, as in our case, two KDE with Gaussian kernels, is not analytically tractable. Thus, we resort here to Monte Carlo integration to compute $D(p\|q)$ (see e.g. [12] for more details).

4. Results

In this section the SKL divergence results obtained for the PMA features are presented and compared to those obtained for the audio features. For the purposes of this work only the similarity between the consonants have been investigated. Instead of comparing all the consonants individually, three different analysis are presented in the following subsections, one for each aspect of speech production: voicing, manner and place of articulation. The phonetic categories used in the analysis are those defined in the International Phonetic Alphabet (IPA) chart [13]. Moreover, for those phones with several possible places of articulation in the IPA chart, the following decisions are taken: (i) [t d n ɹ] are all considered as alveolars and (ii) the affricates [tʃ dʒ] are considered as postalveolars. Finally, the value of the kernel bandwidth in (2) used in the experiments is $\sigma = 0.1$, although it was verified that this parameter has only a minor impact on the results.

4.1. Voicing

Figure 2 shows the SKL divergences obtained for the MFCC and PMA features with respect to the voicing of the phones. As mentioned before, small values for the divergence indicate higher similarity between the phonetic classes since their probability distributions will be more overlapped. For the particular case in which both distributions are the same, the divergence equals to zero, as appears in the diagonal of the matrix. For the sake of avoiding redundancy when presenting the results, and since the resulting matrices are symmetric, only a part of the matrix with the SKL divergences is shown for each type of features: the upper triangular part for the MFCCs and the lower triangular part for the PMA features. Hence, the value

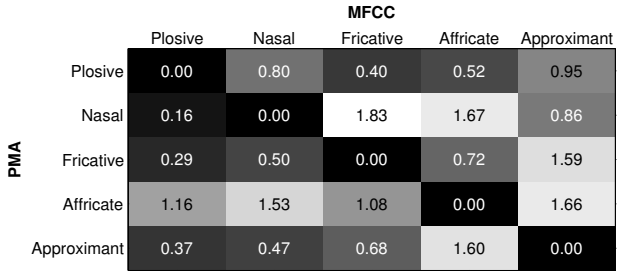

Figure 3: Comparison of the SKL divergences obtained for the MFCC and PMA features with respect to the manner of articulation.

0.90 in the figure represents the divergence between the voiced-unvoiced pairs for the MFCCs, while 0.20 is the result for the PMA features.

Contrary to our previous work [3, 4] where it was reported that PMA data was sufficient to distinguish between voiced and unvoiced fricatives and plosives, the results in figure 2 suggest that PMA provides little information regarding the voicing information in comparison with the audio features. As PMA does not explicitly capture the movement of the vocal cords, there are two possible explanations to this apparent contradiction. First, while the results presented here are computed in a frame-by-frame basis, those reported in [4, 3] are for speech recognition using PMA data. Thus, in addition to static articulatory data, the hidden Markov models (HMMs) used in [4, 3] offer the extra flexibility of modelling both the temporal dynamics and duration of the articulatory data. Second, possible coarticulation effects between phone boundaries can be better represented using HMMs rather than the KDE approach adopted in this paper. Therefore, the extra modelling flexibility offered by the HMMs could be helpful for distinguishing between minimal voice-unvoiced pairs.

4.2. Manner of articulation

Figure 3 shows the SKL divergences obtained for the audio and articulatory data when considering the manner of articulation of the consonants. As expected, PMA data is less informative than the MFCCs for distinguishing this aspect of speech production. This limitation can be attributed to the following. First, as mentioned before, PMA does not have direct access to the airflow information, so those phones with the same place but different manner of articulation (e.g. [t d n s z ɹ l]) are hard to differentiate. Second, the current magnet arrangement limits the ability of PMA for discrimination between some phones. For example, no magnet is attached to the velum in the current set-up and this is probably the reason that nasal consonants are easily confused with other consonants in figure 3.

From the figure we also can see that the plosives are hard to distinguish from other consonants not only in the PMA feature space, but also in the MFCC space. Plosives are among the most complex sounds in English and, as also observed by other authors [14, 15], one important aspect of their articulation refers to the dynamics of the articulators during their pronunciation. This could explain why in our previous study [4] speech recognition using HMMs trained for each phone achieved good results even for the plosive consonants.

Finally, it is worth noting the ability of PMA to correctly discriminate the affricate sounds (i.e. [tʃ dʒ]). One possible ex-

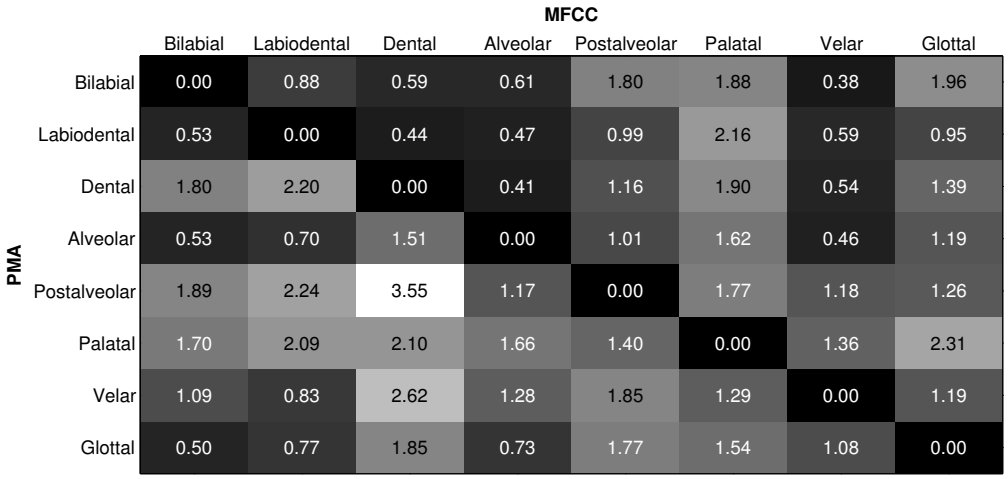

Figure 4: Comparison of the SKL divergences obtained for the MFCC and PMA features with respect to the place of articulation.

planation for this could be that the characteristic pattern of articulation of these sounds, i.e. they begin as plosives articulated in the alveolar area but are released as fricatives with closure in the postalveolar area. They therefore have distinctive place characteristics which PMA can detect.

4.3. Place of articulation

To conclude the analysis, figure 4 shows the results obtained for the different places of articulation of the phones. Surprisingly, the SKL divergences reported in the figure are slightly better for PMA than for the audio features: the average excluding the main diagonal is 1.16 for the MFCCs and 1.51 for the PMA features. This result supports our hypothesis that PMA is better at discriminating the place of articulation rather than the other aspects of speech production. In other words, with PMA we are basically detecting patterns in the magnetic field fluctuations associated with specific articulatory gestures.

Not surprisingly, a large proportion of the phonetic confusions for PMA data occur for the glottal consonants (i.e. [h]). Although adding an extra magnet in the glottis might help to reduce part of these confusions, some practical problems (e.g. gag reflex, danger of swallowing the magnet) along with the relative low frequency of occurrence of glottal consonants in English have prevented us for doing that until now.

It also can be seen in the figure that bilabial-labiodental and bilabial-alveolar are among the pairs with lowest SKL divergences in the PMA feature space. For the bilabial-labiodental case, it is easy to see that the lips in e.g. the phones [p] and [f] adopt a similar position. For the bilabial-alveolar case, one possible explanation for the confusion between these consonants is that the tongue can adopt any position before the airflow is released in the bilabial plosives [p b] and, hence, this can led to confusion with other consonants when looking to the PMA data during the holding phase.

As can be seen, PMA achieves the best result in discrimination for the dental-postalveolar consonants and this is a direct consequence of the different positions that the tongue adopts for these sounds: the dental consonants are articulated with the tongue against the upper teeth, whereas for the postalveolar consonants the tongue is further back, behind the alveolar ridge.

5. Conclusions

In this paper, a study of phonetic similarity for a silent speech interface based on Permanent Magnetic Articulography has been presented. Our results suggest that PMA could potentially be used in applications involving unconstrained, phonetically-rich speech. In particular, it is shown that PMA provides enough information to discriminate the place of articulation of English consonants. On the other hand, other aspects of speech production such as the voicing and manner of articulation are shown to be less accurately modelled by PMA. Furthermore, comparisons between the results reported here and those in previous work suggest that the articulator dynamics could play an important role in phone identification and, thus, it should be taken into account when developing applications involving articulatory data.

Future work includes the extension of this study to other phones, notably the vowels. Furthermore, based on the results obtained in this paper, other future research will address the improvement of the current PMA set-up by modifying the sensor and magnets arrangement for a better phone discrimination.

6. 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.

7. References

[1] The DiSArM project website, http://www2.hull.ac.uk/science/engineering/research/medicalengineering/devicesdiagnosticstherapeutics/speech/disarm-1.aspx

[2] J. M. Gilbert, S. I. Rybchenko, R. Hofe, S. R. Ell, M. J. Fagan, R. K. Moore, and P. D. Green, "Isolated word recognition of silent speech using magnetic implants and sensors," *Med. Eng. Phys.*, vol 32, no. 10, pp. 1189-1197, 2010.

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

[4] R. Hofe, J. Bai, L. A. Cheah, S. R. Ell, J. M. Gilbert, R. K. Moore, and P. D. Green, "Performance of the MVOCA silent speech interface across multiple speakers," in *Proc. Interspeech 2013*, pp. 1140-1143.

[5] R. Hofe, S. R. Ell, M. J. Fagan, J. M. Gilbert, P. D. Green, R. K. Moore, and S. I. Rybchenko, "Speech synthesis parameter generation for the assistive silent speech interface MVOCA," in *Proc. Interspeech 2011*, pp. 3009-3012.

[6] J. Kominek and A. Black, "The CMU Arctic speech databases," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[7] I. T. Jolliffe, "Principal Component Analysis," *Springer-Verlag New York*, 2002.

[8] J. L. Gauvain and C. H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 291-298, 1994.

[9] S. Young, et al., "The HTK Book (for HTK Version 3.4)," *Cambridge University Engineering Department*, 2009.

[10] S-C. Jou, L. Maier-Hein, T. Schultz, and A. Waibel, "Articulatory feature classification using surface electromyography," in *Proc. ICASSP 2006*, pp. 605-608.

[11] S-C. Jou, T. Schultz, M. Walliczek, F. Kraft, and A. Waibel, "Towards continuous speech recognition using surface electromyography," in *Proc. Interspeech 2006*, pp. 573-576.

[12] J. R. Hershey and P. A. Olsen, "Approximating the Kullback Leibler divergence between Gaussian mixture models," in *Proc. ICASSP 2007*, pp. 317-320.

[13] "The International Phonetic Alphabet (revised to 2005)," *International Phonetic Association*, 2005.

[14] D. Neiberg, G. Ananthakrishnan, and O. Engwal, "The acoustic to articulation mapping: non-linear or non-unique?," in *Proc. Interspeech 2008*, pp. 1485-1488.

[15] G. Ananthakrishnan, O. Engwall, and D. Neiberg, "Exploring the predictability of non-unique acoustic-to-articulatory mappings," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 10, pp. 2672-2682, 2012.