

A Wearable Silent Speech Interface based on Magnetic Sensors with Motion-Artefact Removal

Lam A. Cheah¹, James M. Gilbert¹, Jose A. Gonzalez², Phil D. Green², Stephen R. Ell³,
Roger K. Moore² and Ed Holdsworth⁴

¹*School of Engineering, The University of Hull, Kingston upon Hull, U.K.*

²*Department of Computer Science, The University of Sheffield, Sheffield, U.K.*

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

⁴*Practical Control Ltd, Sheffield, U.K*

Keywords: Silent Speech, Permanent Magnet Articulography, Assistive Technology, Motion Artefact Removal, Magnetic Sensors.

Abstract: For a silent speech interface (SSI) to be truly practical, it has to be able to tolerate motion artefacts generated by the user while engaging in normal activities of everyday living. This paper presents a wearable speech restoration system based on magnetic sensors with an integrated background cancellation technique to eliminate the effect of motion-induced interference. The background cancellation technique is assessed when the user makes no intentional movement, when they perform a set of defined head movements and when they make more natural, conversational head movements. The performance is measured for the TIDigits corpus in terms of whole word recognition rate using a Hidden Markov Model and through Mel Cepstral Distortion for a Direct Synthesis of speech using Deep Neural Networks. The results indicate the robustness of the sensing system with background cancellation against the undesirable motion-induced artefacts.

1 INTRODUCTION

Speech is perhaps the most convenient and natural way of communication between humans. People whose larynx has been surgically removed following throat cancer, trauma or destructive throat infection find themselves struggling with vocal communication after losing their voice. This often has a severe impact on people's lives, which can lead to social isolation, feelings loss of identity and depression (Fagan et al., 2008; Danker et al., 2010). Unfortunately, existing post-laryngectomy voice restoration methods (i.e. oesophageal speech, the electrolarynx and speech valves) are often limited by their usability and abnormal voice quality (e.g. robotic and masculine voice), which may be unsatisfactory especially for female patients (Fagan et al., 2008). Typing-based alternative and augmentative communication (AAC) devices can also be employed, but are limited by slow manual text input.

To overcome the limitation of existing methods, assistive technologies (ATs) such as silent speech

interfaces (SSIs) have emerged and shown promising potentials in recent years. A SSI is a system that enables speech communication in the absence of audible speech by exploiting other non-audible signals associated with speech production (Denby et al., 2010). Because of their unique feature, SSIs can also be deployed in other scenarios such as spoken communication aid in noisy environments or where privacy/confidentiality is desirable. To date, there are several types of SSIs using different modalities, such as the electrical activity produced by the articulator muscles (Brumberg et al., 2010; Herff et al., 2015), the brain's electrical activity (Schultz and Wand, 2010; Wand et al., 2014), or the movement of speech articulators (Toda et al., 2008; Gilbert et al., 2010; Hueber et al., 2010). Despite the attractive attributes of SSIs, many are still deemed as impractical and ineffective outside laboratory environment. Factors, such as a high degree of intrusiveness, discomfort, unattractive appearance, unintelligible speech quality and artefacts/noise interference, affect their real-world implementation (Denby et al., 2010).

The present work builds upon the permanent magnet articulography (PMA), which is a sensing technique for articulator motion capture (Fagan et al., 2008; Gilbert et al., 2010; Hofe et al., 2013; Cheah et al., 2015). In previous work, progresses were reported in terms of the hardware, user-centric design (Cheah et al., 2015) and speech processing (Gonzalez et al., 2016). However, measurements from wearable devices (including PMA) are known to be susceptible to motion-induced interference (Such, 2007). Comparing to measurements acquired within laboratory settings, which are generally conducted when subjects are in steady positions, measuring outside the laboratory faces the problem of motion artefacts arising from unrestricted head movements (i.e. corruption by varying earth's magnetic field) by the subjects. For a PMA based device to be practical and effective in the field, integration of motion artefact cancellation into current prototype is therefore critical.

The remainder of this paper is structured as follows. Section 2 describes the PMA prototype and the proposed background cancellation technique. Section 3 discusses about the performance of the PMA system, followed by the experimental results in Section 4. The final section concludes this paper and provides an outlook for future work.

2 MATERIAL AND METHOD

2.1 System Design

PMA is a technique for capturing the movement of the speech articulators by sensing changes in the magnetic field generated by a set of permanent magnets attached to the speech articulators (i.e. lips and tongue) by a set of magnetic sensors arranged around the mouth. The acquired data may then be used to determine the speech which the user wishes to produce, either by performing automatic speech recognition (ASR) on the PMA data (i.e. recognize-then-synthesis) (Hofe et al., 2013; Cheah et al., 2015), or by directly synthesizing audible speech from the articulatory data (i.e. direct synthesis) (Gonzalez et al., 2016). Contrary to other methods for articulator motion capture, PMA does not attempt to identify the Cartesian position or orientation of the magnets, but rather a composite of the magnetic fields from magnets that are associated with a particular

articulatory gesture. The current PMA system consists of six cylindrical Neodymium Iron Boron (NdFeB) magnets: four on the lips ($\text{\O}1\text{mm}\times 5\text{mm}$), one on the tongue tip ($\text{\O}2\text{mm}\times 4\text{mm}$) and one on the tongue blade ($\text{\O}5\text{mm}\times 1\text{mm}$), as illustrated in figure 1(a). These magnets are temporarily attached using Histoacryl surgical tissue adhesive (Braun, Melsungen, Germany) during experimental trials, but will be surgically implanted for long term usage. The remainder of the PMA system is composed of four tri-axial anisotropic magnetoresistive (AMR) sensors mounted on a bespoke wearable headset, a control unit, a rechargeable battery and a wireless link to a processing unit (e.g. computer/PC), as shown in figure 1(b).

The PMA has distinct advantages over other SSIs, such as being unobtrusive with no wires coming out of the mouth or electrodes attached to the skin, which may cause unwanted attention in public. Moreover, the PMA system is also relatively lightweight and highly portable. In addition, the current prototype has extensively improved over its predecessors particularly in terms of appearance, comfort and ergonomic factors for the users, but without compromising on the speech performances (Cheah et al., 2015).

2.2 Cancellation of Motion-induced Interferences

As illustrated in Fig, the first three sensors located closer to the mouth (Sensor1-3) are used to monitor the articulators. The data acquired by any of these sensors (S_A) is made up of the desired signal from one or more of the magnetic markers (A), and other 'background' signals (B_A), the most significant of which is generally the result of the earth's magnetic field. Hence:

$$S_A = A + B_A \quad (1)$$

Movement of the user's head results in significant interference to the field detected by the sensors and it has been found that the desired signal derived from the articulator movements are up to 10 times smaller than the changes resulting from typical head movements. This result in a poor signal-to-noise (SNR), which degrades the performance of the speech restoration algorithms, thus eliminating the head motion-induced interference is necessary.

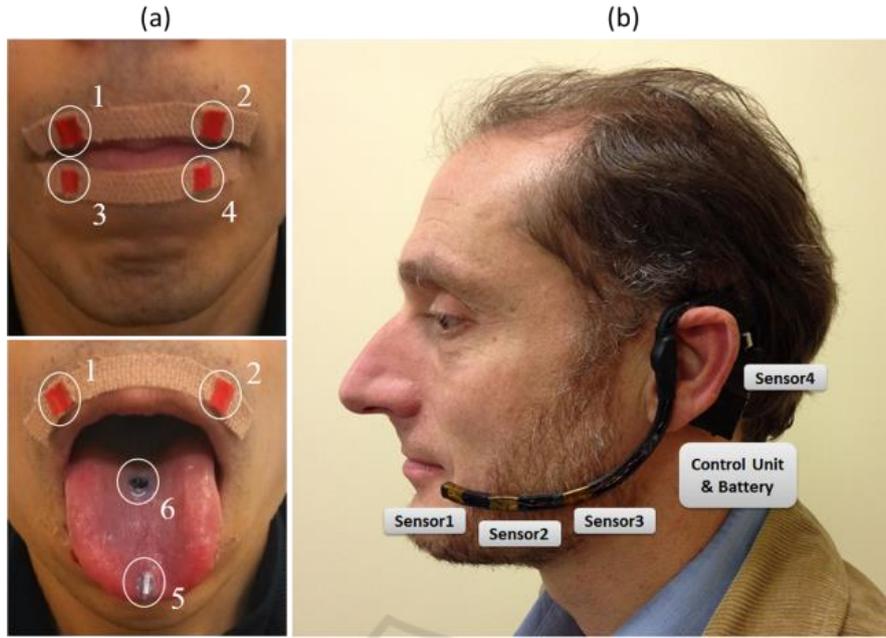


Figure 1: (a) Arrangement of magnetic markers placed on the tongue and lips. (b) Wearable PMA headset with 4 tri-axial magnetic sensors, control unit and battery.

Although head movements typically result in lower frequency signals than articulatory movement, there is significant overlap in their respective frequency spectral, thus making conventional filtering ineffective.

The method proposed here is to reduce the effect of these background signals by utilizing a reference sensor (S_R) located farther from the mouth (Sensor4) as shown in figure 1(b). This is placed at a distance far from the magnetic markers but rigidly attached to the PMA headset so that it moves with the articulator sensors (S_A) and so measures the ambient magnetic field. Hence, the aim is to estimate B_A and cancel out the effect of the background field on the articulator sensor signal in (1). The estimate of B_A may be calculated as:

$$\hat{B}_A = S_R \hat{T}_{RA} \quad (2)$$

where T_{RA} is a transformation between reference sensor and the articulator sensors. The estimate from (2) is then used to remove the background field from an articulator sensor and leave only the desired articulatory signal:

$$\begin{aligned} \hat{S}_A &= A + B_A - \hat{B}_A \\ \hat{S}_A &\approx A \end{aligned} \quad (3)$$

The required transformation T_{RA} could be calculated from the relative orientation of reference and articulator sensors but this is difficult to measure and will change if the headset becomes distorted. Instead,

it can be estimated by taking a series of measurement of S_A and S_R while rotating the PMA headset in the absence of any articulator movement i.e. $A = 0$, where (1) substitute into (3):

$$S_A = S_R \hat{T}_{RA} \quad (4)$$

From (4), T_{RA} is estimated via the least square method to determine the best cancellation coefficients. Assuming a model $Y = X\beta + \varepsilon$ and a set of M measurements of Y and X , then using the least squares method, the best estimate $\hat{\beta}$ of β is

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad (5)$$

where

$$Y = \begin{bmatrix} S_{A1}^x(1) & S_{A1}^y(1) & S_{A1}^z(1) & \dots & S_{AK}^z(1) \\ S_{A1}^x(2) & S_{A1}^y(2) & S_{A1}^z(2) & \dots & S_{AK}^z(2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ S_{A1}^x(M) & S_{A1}^y(M) & S_{A1}^z(M) & \dots & S_{AK}^z(M) \end{bmatrix}$$

is a set of M samples of the K articulator sensor outputs (each with x , y and z components) and

$$X = \begin{bmatrix} 1 & S_R^x(1) & S_R^y(1) & S_R^z(1) \\ 1 & S_R^x(2) & S_R^y(2) & S_R^z(2) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & S_R^x(M) & S_R^y(M) & S_R^z(M) \end{bmatrix}$$

is a set of M samples of the x , y and z components of the magnetic field at the reference sensor, β is a $4 \times K$

matrix of cancellation coefficients and ε is the estimation error which we seek to minimize.

$$\beta = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix} \quad (6)$$

The first row in (6) is a constant offset cancellation term followed by 3 cancellation coefficients corresponding to the estimated transformation matrix T_{RA} . Hence, in addition to removing the effect of the background field, any offset on the articulator sensor output is also removed.

3 PERFORMANCE EVALUATION

3.1 Experimental Setup

The data used for this evaluation were collected from a male native English speaker who is proficient in the usage of the PMA device. Magnets were temporarily attached to the subject's lips using medical adhesive strips and the tongue using medical adhesive (Histoacryl®, Braun, Melsungen, Germany), as illustrated in figure 1(a). and the PMA headset is shown in figure 1(b). University research ethics committee approval was obtained for this procedure. It may be noted that the removal of the tongue magnets causes temporary, mild discomfort but that the intention is that magnets, with a biocompatible protective coating, would be permanently implanted into the articulators of long-term users.

For optimal sound quality, the recording was conducted in an acoustically isolated room using a shock-mounted AKG C1000S condenser microphone via a dedicated stereo Lexicon Lambda USB-sound card. The audio and PMA signals were recorded simultaneously at sampling frequencies of 16 kHz and 100 Hz. A bespoke Matlab-based GUI was created to provide visual prompt of randomized utterances to the subject at an interval of 5 seconds during the recording session.

3.2 Vocabulary and Data Recording

To evaluate the performance of the proposed background cancellation method, sentences from the TIDigits database (Leonard, 1984), which consists of sequences of up to seven connected English digits, were recorded. The vocabulary is made up of eleven individual digits, i.e. from 'one' to 'nine', plus 'zero'

and 'oh' (both representing digit 0). Each dataset consists of 77 digits sequences, and a total six datasets were recorded, giving 462 utterances containing 1518 individual digits.

For model training, 3 datasets (231 utterances) with no intentional head movements were employed, while the testing was carried out with 1 dataset (77 utterances) for each of three conditions: i) no intentional movement, ii) fixed angle movements where the subject was asked to rotate his head to look at a sequence of markers and speak as prompted. These markers were arranged to give 30° rotation left and right and 22° tilt upwards and downwards, iii) conversational movements where subject was asked to read the prompt and speak while simultaneously moving their head in gestures which might be made during conversation e.g. shaking from side to side, nodding up and down, and tilting side to side.

3.3 Evaluation

Speech recognition and direct synthesis (i.e. generation of audible speech from PMA data) experiments were used to evaluate the performance of our system. For speech recognition, whole-word hidden Markov models (HMMs) were trained on PMA data as described (Cheah et al., 2015). The word accuracy results (i.e. percentage of words correctly recognized after discounting the insertion errors) achieved by our PMA system are reported here as an objective measure of the system performance for speech recognition. For direct synthesis, a deep neural network (DNN) with 3 hidden layers and 128 sigmoid units per layer was trained on simultaneous recordings of PMA and speech data to predict the acoustic signals from the articulator movement. As described in (Gonzalez et al., 2017), the DNN was trained on feature vectors extracted from the raw PMA and speech signals. The speech signals were parameterized as 27 dimensional-feature vectors computed every 5ms from analysis windows spanning 25ms of data: 25 Mel-frequency cepstral coefficients (MFCCs), fundamental frequency (F_0) in log-scale and binary voicing decision. The PMA signals were parameterized as segmental features by applying the principal component analysis (PCA) technique for dimensionality reduction over symmetric windows spanning 105ms of articulator movement data. Performance on direct synthesis was evaluated by comparing the speech features extracted from the original signals with those predicted by the DNN from PMA data using the following objective measures: Mel-Cepstral distortion (MCD) (Kubichek et al., 1993) in dBs for the MFCCs, root mean square error (RMSE) for F_0 and voicing error rate.

4 RESULTS AND DISCUSSION

4.1 Performance of the Background Cancellation Scheme

The performance of the background cancellation scheme is illustrated in figure 2 for a single articulator sensor output. In this trial, a participant made a series of head movements (tilting the head from side to side) and the uttered a digit sequences (i.e. ‘8428136’) without head movement followed by simultaneous utterance with same sequence of head movement. It can be seen in the raw data as shown in figure 2, that the effect of the head movement is approximately 5 times larger than the signals resulting from speech. After background cancellation, the effect of head movement has been almost entirely removed (during the ‘Movement’ phase in figure 2).

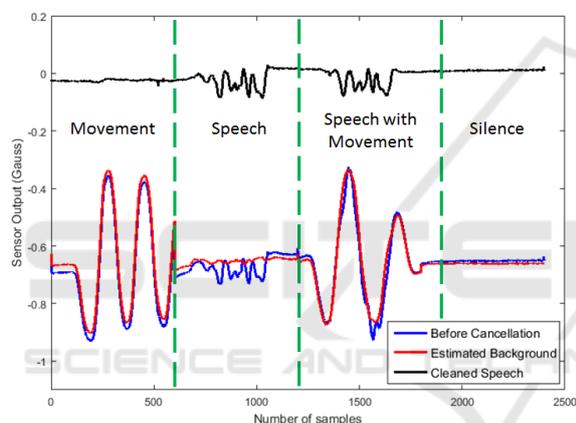


Figure 2: Articulator sensor signal (Sensor3, y-axis) with and without cancellation.

The effect of the background cancellation can be quantified by comparing the rms of magnetic flux density variations resulting from articulation and head movement, before and after cancellation. In addition, the variation in measured flux density in the absence of both head movement and articulation (‘Silence’ as illustrated in figure 2) is also calculated as this represents the ‘noise floor’ below which signals would be undetectable. The rms values of these quantities for a dataset of 100s speech and 80s of movements are given, along with relevant ratios in table 1.

From table 1 we see that the mean rms during articulation before cancellation is 2-3 times smaller than the signal resulting from head movement. In other words, the signal is buried in the background. After cancellation, the articulation signal amplitude remains largely unchanged but the amplitude of the

signal resulting from movement has reduced by a factor of approximately 15-20, thus the desired articulation signal is now 6 times larger than the background. On the other hand, the process of background cancellation has not significantly affected the level of noise during silence and so the articulation signal remains at least 10 times larger than the noise floor.

Table 1: Articulation, background and noise level before and after cancellation.

		Before Cancellation	After Cancellation
Articulation (mG)		20.96	20.69
Movement (mG)	Fixed Angle	38.57	2.8
	Conversational	59.13	3.37
Silence		1.79	1.46
Articulation	Fixed Angle	0.54	6.85
Movement	Conversational	0.35	6.14
<u>Articulation</u> Silence		11.39	13.97

4.2 Performance of PMA System in Speech Recognition

Figure 3 shows the word accuracy results on the speech recognition experiment with and without background cancellation applied to the PMA data. Without background cancellation, the recognition performance in the presence of movement deteriorates significantly (down to about 2%).

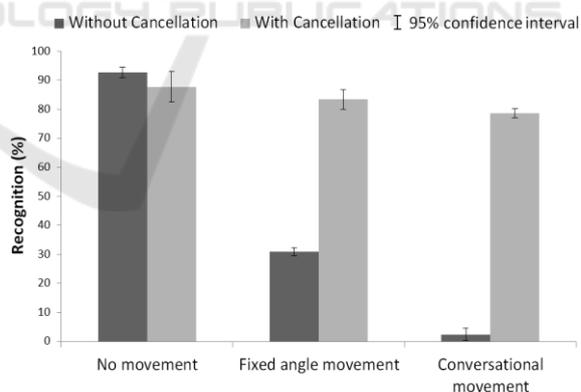


Figure 3: Word recognition rates with and without background cancellation.

The use of background cancellation deteriorates slightly the results in the absence of movement but causes a large improvement in recognition in the presence of movement. It should also be noted that asking the participant to move and speak at the same time causes some difficulties since, for instance, the subject tend to hesitate and stumble over words more because they are not continually looking at the

Table 2: Articulation, background and noise level before and after cancellation.

Movement	Cancellation	MCD (dB)	F_0 RMSE(Hz)	Voicing error rate (%)
No movement	No	5.02	22.82	15.90
	Yes	3.91	19.48	15.73
Fixed Angle movement	No	5.53	27.80	26.40
	Yes	4.55	19.71	16.36
Conversational movement	No	6.41	26.73	28.37
	Yes	4.49	17.01	16.18

prompt. Hence, some deterioration in recognition may be due to changes in articulation rather than a failure of the background cancellation scheme.

4.3 Performance of PMA System in Direct Synthesis

Table 2 shows the results obtained on the direct synthesis experiment. For each type of speech feature (i.e. MFCC, F_0 , and voicing decision) the error made by the DNN when predicting the feature from PMA data is shown. As in the speech recognition experiment, significant improvements are obtained for all types of features when background cancellation is applied. Again, the most detrimental movement when no compensation is applied is the conversational movement. However, when cancellation is applied, the performance results obtained for both types of movements is similar. In any case, from informal listening, intelligible speech is obtained for all movements types when background cancellation is applied on the PMA data.

5 DISCUSSION

The background cancellation results presented in this paper are derived for a single male subject who is proficient in use of the PMA system. The performance of the Direct Synthesis approach for a small number of different speakers without intentional head movement, but with background cancellation active, is described in (Gilbert et al., 2017). In that paper, three male subjects and one female subject are considered and significant variations in performance are noted between them. Subjects who speak more slowly achieve better performance while the female subject achieves inferior performance. It is noted that the smaller size of the female subject's head means that the sensors are further from the articulators, and it is suggested that this may explain the inferior performance. The

effect of gender and age on background cancellation performance have not been assessed since this would require a large number of recordings which are time consuming and uncomfortable.

The approach to background cancellation proposed here is based on the assumption that the magnetic field experienced by the articulator sensors is the same as that experienced by the reference sensor. While this is expected to be true of the earth's magnetic field, it may not be the case for localised sources of magnetic fields or close to objects which distort the earth's field. Further testing will be required to assess whether the performance of the background cancellation method is maintained in other environments, although it may be noted that no measures were taken to control the background field in the trials reported here. Similarly, although the method proposed should be capable of cancelling the effects of alternating magnetic fields, this has not been verified experimentally.

Developing a speech rehabilitation system which is acceptable to patients who have undergone a laryngectomy involves a number of challenges. The ability of the PMA system to remove motion artefacts is one important element of this but work is also needed to improve the speech reconstruction achieved, to assess usability and potentially improve the comfort of the sensor frame and to prove the safety and viability of long-term implantation of magnets into the articulators. Work is underway to address all of these challenges and the results will be reported elsewhere.

6 CONCLUSIONS

Significant improvements on the silent speech restoration were achieved when using the proposed background cancellation scheme for removal of motion artefact induced by subject's head movement. This is an important step in our objective of developing a speech rehabilitation system for

everyday use by subjects who have undergone a laryngectomy. Encouraged by the results obtained so far, work is underway to further evaluate the background cancellation scheme and to develop other aspects of the system so that a speech rehabilitation system can be offered to individuals who have undergone a laryngectomy which they find preferable to existing methods.

ACKNOWLEDGEMENTS

The report is an independent research funded by the National Institute for Health Research (NIHR)'s Invention for Innovation Programme (Grant Reference Number II-LB-0814-20007). The views stated are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

REFERENCES

- Brumberg, J. S., Nieto-Castanon, A., Kennedy, P. R., and Guenther, F. H. (2010). Brain-computer interfaces for speech communication. *Speech Communication*, 52(4):367-379.
- Cheah, L. A., Bai, J., Gonzalez, J. A., Ell, S. R., Gilbert, J. M., Moore, R. K., and Green, P. D. (2015). A user-centric design of permanent magnetic articulography based assistive speech technology. In *Proc. BIOSIGNALS*, pages 109-116, Lisbon, Portugal.
- Denby, B., Schultz, T., Honda, K., Hueber, T., Gilbert, J. M., and Brumberg, J. S. (2010). Silent speech interfaces. *Speech Communication*, 52(4):270-287.
- Danker, H., Wollbrück, D., Singer, S., Fuchs, M., Brähler, B., and Meyer, A. (2010). Social withdrawal after laryngectomy. *Eur Arch Otorhinolaryngol*, 267(4):593-600.
- Fagan, M. J., Ell, S. R., Gilbert, J. M., Sarrazin, E., and Chapman, P. M. (2008). Development of a (silent) speech recognition system for patients following laryngectomy. *Medical Engineering & Physics*, 30(4):419-425.
- Gilbert, J. M., Rybchenko, S. I., Hofe, R., Ell, S. R., Fagan, M. J., Moore, R. K. and Green, P. D. (2010). Isolated word recognition of silent speech using magnetic implants and sensors. *Medical Engineering & Physics*, 32(10):1189-1197.
- Gilbert J. M., Gonzalez J.A., Cheah L.A., Ell, S.R., Green P., Moore R.K. and Holdsworth E., Restoring speech following total removal of the larynx by a learned transformation from sensor data to acoustics, *The Journal of the Acoustical Society of America* 141, EL307 (2017);
- Gonzalez, J. A., Cheah, L. A., Gilbert, J. M., Bai, J., Ell, S. R., Green, P. D., and Moore, R. K. (2016). A silent speech system based on permanent magnet articulography and direct synthesis. *Computer Speech & Language*, 39:67-87
- Gonzalez, J. A., Cheah, L. A., Green, P. D., Gilbert, J. M., Ell, S. R., Moore, R. K., and Holdsworth, E. (2017). Evaluation of a silent speech interface based on magnetic sensing and deep learning for a phonetically rich vocabulary. In *Proc. Interspeech*, pages 3986-3990, Stockholm, Sweden.
- Herff, C., Heger, D., de Pestors, A., Telaar, D., Brunner, P., Schalk, G., and Schultz, T. (2015). Brain-to-text: decoding spoken phrases from phone representations in the brain. *Frontiers in Neuroscience*, 9(217).
- Hofe, R., Ell, S. R., Fagan, M. J., Gilbert, J. M., Green, P. D., Moore, R. K., and Rybchenko, S. I. (2013). Small-vocabulary speech recognition using silent speech interface based on magnetic sensing. *Speech Communication*, 55(1):22-32.
- Hueber, T., Benaroya, E.-L., Chollet, G., Denby, B., Dreyfus, G., and Stone, M. (2010). Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips. *Speech Communication*, 52(4):288-300.
- Kubichek, R. (1993). Mel-cepstral distance measure for objective speech quality assessment. In *Proc. IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, pages 125-128, Victoria, Canada.
- Leonard, R. G. (1984). A database for speaker-independent digit recognition. In *Proc. ICASSP*, pages 328-331, San Diego, USA.
- Such, O. (2007). Motion tolerance in wearable sensors—the challenge of motion artifact. In *Proc. IEEE EMBC*, pages 1542-1545, Lyon, France.
- Schultz, T., and Wand, M. (2010). Modeling coarticulation in EMG-based continuous speech recognition. *Speech Communication*, 52(4):341-353.
- Toda, T., Black, A. W., and Tokuda, K. (2008). Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model. *Speech Communication*, 50(3): 215-227.
- Wand, M., Janke, M., and Schultz, T. (2014). Tackling speaking mode varieties in EMG-based speech recognition. *IEEE Transactions on Biomedical Engineering*, 61(10):2515-2526.