16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

# AN ASSESSMENT OF SPEECH RELATED INFORMATION CONTAINED IN GEMS SIGNALS

*K. A. Jellyman*[1], *W. M. Liu*[1], *J. S. D. Mason*[1] *and N. W. D. Evans*[2,1]

[1]School of Engineering, Swansea University, UK
[2]Institut Eurécom, Sophia Antipolis, France
{174869, 199997, j.s.d.mason}@swansea.ac.uk, nicholas.evans@eurecom.fr

## ABSTRACT

As the essence of communication speech intelligibility, rather than more general speech quality, can be of paramount importance when communications systems operate in high noise environments. This paper considers applications where the acoustic signal is degraded by noise so as to be effectively lost and applications where it is simply not available. With such applications in mind we report experiments to assess the use of non-acoustic general electromagnetic motion sensors (GEMS). Whilst GEMS signals are essentially immune to background noise they are incomprehensible to the human listener. We show that GEMS signals nonetheless contain meaningful speech information within a usable bandwidth in the region of 1 to 2 kHz and report the first comparison of GEMS signals to acoustic signals in the context of automatic speech recognition (ASR). For a small, isolated digit ASR task in a speaker-dependent mode results show word accuracies of 77% are achieved using GEMS signals alone.

## 1. INTRODUCTION

This paper addresses the problem of speech communications in situations where the importance of intelligibility exceeds all other aspects of general quality. Some examples include military, security and surveillance applications either for high noise conditions, where the acoustic signal is essentially lost, or for situations where the acoustic signal is inaccessible. The importance of intelligibility, which is of course the very essence of communication, has been recognised by the ITU who have recently initiated a programme to extend the standard quality measure known as PESQ (perceptual evaluation of speech quality) [1] to the measurement of intelligibility.

Under noisy conditions it is sometimes desirable to process speech with the aim of improving intelligibility. Speech enhancement, however, is a notoriously difficult task. Hu and Loizou [2] performed an extensive comparison of speech enhancement algorithms reporting that almost none improves intelligibility.

Instead of trying to recover intelligibility from the acoustic signal, an alternative approach is to augment the degraded acoustic signal with other forms of speech representation, for example visual speech, lip dynamics, or by using throat microphones as in the work of Graciarena *et al* [3]. All of these approaches have been extensively reported. A lesser known approach is the use of general electromagnetic motion sensors (GEMS)[1] [4]. GEMS signals come from a low powered radar device which reflects movements within the vocal tract region. As with visual speech, GEMS signals are largely immune to background acoustic noise [5, 6] making them especially appealing for high-noise applications and also as a direct substitute for the conventional acoustic signal.

This paper assesses the use of GEMS-derived signals as an alternative to acoustic signals and pertains to situations where intelligibility is paramount. Here we assume applications where either the noise level is so high that the acoustic signal is essentially lost or that it is inaccessible. In both cases 'speech'-to-text is required thus we report the first experiments to assess the use of GEMS-derived signals for automatic speech recognition.

The remainder of the paper is organised as follows. In the following section we illustrate an example GEMS signal. Section 3 reviews GEMS research published since the original doctoral works of Burnett [7] and Gable [8]. Section 4 presents parallel ASR experiments which compare GEMS and corresponding acoustic signals. Finally our conclusions are presented in Section 5.

## 2. GEMS SIGNALS

GEMS signals come from a low powered radar device developed at the Lawrence Livermore National Laboratory (LLNL) [4]. The sensors are pointed toward the throat and the resulting GEMS signals reflect movements within the vocal tract region.

Example GEMS and acoustic signals are illustrated in Figure 1, top and bottom respectively, for the digit-string utterance "2 1 5 4 6". The SNR is in the order of 20 dB and both signals are captured simultaneously, hence there is a high degree of correlation. For each spectrogram the horizontal axis indicates time between 0 and 5.3 seconds and the vertical axis indicates frequency from 0 to 4 kHz. Pitch harmonics are visible in the GEMS signal spectrogram of Figure 1(a) although only in the band below 1 kHz. Energy correlation (high energy is signified by darker colour) can be seen between the GEMS and acoustic spectrogram, illustrated in Figure 1(c).

Being essentially immune to back ground noise note that between spoken digits, for example at 1 second, the GEMS signal energy is low whereas significant noise energy is seen in the corresponding acoustic spectrogram. In the limits, however, in high noise Lombard effects are likely to introduce variation in the GEMS signals as well as in the acoustic signals. These aspects remain topics for future research.

To the right of each spectrogram in Figure 1 are the corresponding time waveforms for the single digit "2" between 0.2 and 0.8 seconds. They also show good correlation in terms of

[1]There is more than one definition of GEMS. Others include glottal electromagnetic micropower sensor and glottal electromagnetic sensor.

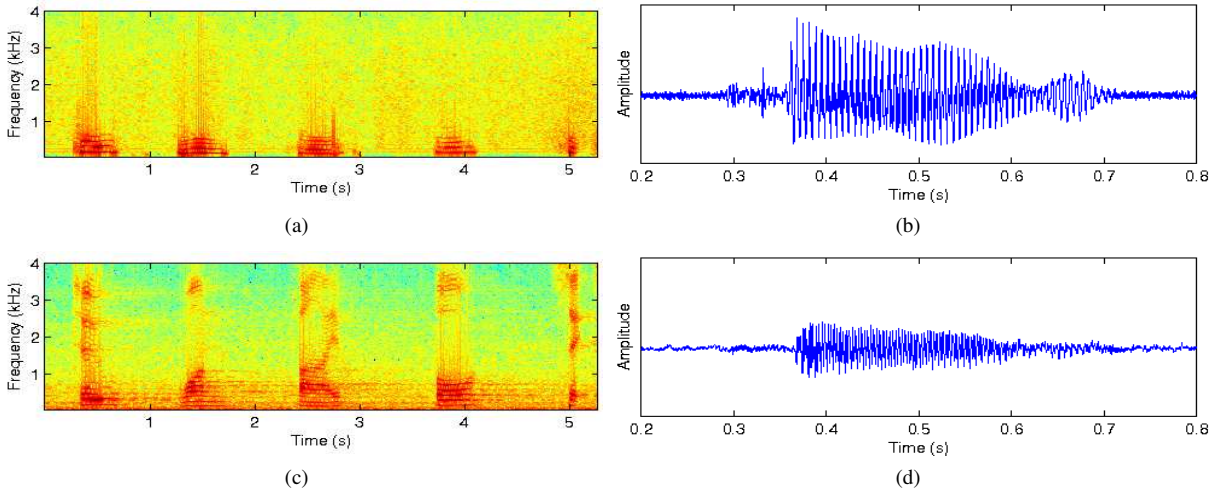


Figure 1: Example spectrograms and time waveforms of corresponding GEMS (top) and acoustic (bottom) signals for the utterance "2 1 5 4 6" at approximately 20 dB SNR. The horizontal axis is time in seconds (0 to 5.3 seconds for spectrograms (a) and (c), 0.2 to 0.8 seconds for time waveforms (b) and (d)). The vertical axes of the spectrograms (left) indicate frequency up to 4 kHz and for the time plots (right) indicate amplitude.

energy and pitch peaks. GEMS signals are audible but to the human listener they are incomprehensible. That is not to say, however, that they do not convey speech information and, given the appeal of GEMS signals for high noise-type applications and for cases where the acoustic signal is inaccessible, we seek in this paper to assess the use of GEMS-derived signals for automatic speech recognition. First though, we review related work.

## 3. PREVIOUS GEMS RESEARCH

Original work stems from the doctoral theses of Burnett [7] and Gable [8]. Burnett shows that GEMS signals, while associated with pitch, do not indicate vocal cord closure, but more generally indicate the dynamics of the trachea wall. Burnett's original work has been extended more recently by the works of Demiroglu and Anderson [9] and Quatieri *et al* [6] who both observe the presence of voice bars. Voice bars are low frequency periodic energy observed during the closure interval for voiced stops. They are detectable by conventional acoustic microphones, but can be easily lost as a result of background noise. Quatieri *et al* [6] shows that GEMS are able to detect the presence of what appear to be weak erratic pulses referred to as glottalisation. They show that glottalisation is partially prevalent in the acoustic signal, but because of the weak nature not all the glottalisation pulses are observed. An example of glottalisation is evident in Figure 1. The GEMS signal in (b) shows a small increase in energy between 0.6 and 0.7 seconds whilst the acoustic signal in (d) continues to decay.

A strong link between pitch and GEMS is well established and this is to be expected given the high dynamics of the vocal cords, especially during voiced speech. However, there is plenty of evidence to suggest other dynamic components exist in the GEMS signals and that these components might convey useful information. This raises the question of whether or not the GEMS signals contain any useful information beyond the obvious link to pitch.

Previous related work has suggested this to be the case. However, the literature on the subject is, perhaps surprisingly, little especially given the appeal of GEMS for high-noise applications. Below we summarise published work which has investigated the use of GEMS in four areas of research.

**Speaker recognition:** Early work to investigate the speaker information in GEMS signals was reported by Gable [8] in 2000 who, with Burnett [7], collected the Lawrence Livermore National Laboratory (LLNL) database. The LLNL database is comprised of 15 male speakers with up to 4 sessions. Gable combined GEMS and acoustic signals and investigated speaker verification in a text-dependent mode at noise levels as low as -10 dB where the acoustic signal is essentially lost due to noise. Campbell *et al* [10] (2003) investigated speaker identification experiments using GEMS among two other non-acoustic sensors. They report experimental work performed on the LLNL and DARPA ASE pilot speech database [10] which is comprised of 10 male and 10 female speakers. Speaker identification performance was investigated using GEMS signals alone and also in combination with acoustic signals. They used both time domain features and conventional cepstral features with various normalisation strategies and dimensionality reduction. Experiments using GEMS signals alone showed 64% accuracy for a 1-in-15 speaker identification task. The combining of GEMS with acoustic signals under relatively low noise conditions was shown not to offer any improvement in performance. However, in high noise conditions, the use of GEMS signals improved scores and time domain features were found to be better than conventional cepstral coefficients.

**Speech enhancement:** In 2000 Ng *et al* [11] investigated speech enhancement using two filters that fuse information from acoustic and GEMS signals. The first filter, referred to as a glottal windowing (GWIN) filter, implements a form of comb filtering which retains only harmonics at specific frequencies related to the excitation, as derived from GEMS

signals. The second filter, referred to as a glottal correlation filter (GCOR), is based on the spectral subtraction of noise from the degraded acoustic signal enhanced by the use of GEMS. The GEMS signals are used in two ways, firstly to determine periods of noise that can be used to form the noise estimate, and secondly as a noise free excitation signal. Both filters are limited to voiced speech; unvoiced speech is Wiener filtered. Assessment is performed on the LLNL database in white noise conditions down to 3 dB. The processed speech is reported to be of high quality. Raj and Singh [12] reported some successful experiments using GEMS signals in a speech enhancement mode, prior to automatic speech recognition.

**Speech recognition:** In 2004 Demiroglu and Anderson [5] reported a syllable based automatic speech recognition experiment which combines acoustic and three GEMS features, namely energy, delta and double delta. Using the DARPA ASE pilot speech database they show an improvement in ASR performance of 9% over an error rate of 40% in noisy conditions when GEMS features are combined with acosutic features. In clean conditions, however, the improvement over the acoustic-only system is not significant with error rates of 8% with acoustic-only and combined features. Experiments using GEMS-only features are not reported.

**Speech coding:** In 2006, Quatieri *et al* [6] investigated the use of GEMS with other non-acoustic and acoustic signals to improve intelligibility performance of the standard mixed excitation linear predictive (MELP) coder [13] in various degraded conditions. Two sensor fusion procedures are proposed that involve the use of GEMS, and for both procedures GEMS are used to provide pitch information. Experiments using the DARPA ASE pilot speech database show improvements in intelligibility ranging from approximately 2% to 9% depending on the noise environment and speakers used.

The potential of GEMS signals is thus clear. Whilst the benefit of combining GEMS signals with conventional acoustic signals has been demonstrated previously, with the exception of Campbell's speaker identification work [10], there is no published literature which assesses their use as a direct substitute. The contribution of this paper relates to the first comparison of GEMS and acoustic signals when used independently for automatic speech recognition.

## 4. ASR EXPERIMENTS

The objective of the experiments reported here is to compare automatic speech recognition (ASR) performance using GEMS signals to that using acoustic signals. Thus a series of parallel experiments with corresponding acoustic and GEMS signals are conducted individually with common configurations across the two modes.

### 4.1 Database

Acoustic and GEMS signals were recorded simultaneously from 4 speakers who read aloud the entries of a completed Sudoku puzzle. Each speaker read the same 9-by-9 full matrix of 81 digits once in 5 sessions in a more-or-less isolated word manner. Each speaker therefore contributed 405 digits with 45 versions of each digit. With a sampling frequency of 8 kHz in all cases, recordings were taken simultaneously from one conventional acoustic microphone and one GEMS which captures two orthogonal signals, referred to as g1 (in-phase) and g2 (quadrature-phase) throughout the remainder of the paper. Each recording thus produces a triplet of signals resulting in a total of 1215 recordings per speaker. Speech end-pointing is performed on the acoustic signal and the same timings were used for the GEMS signals. The resulting triplet of signals are therefore all of the same length. We have not attempted to combine the two GEMS signals and have instead assessed each of them independently.

| # Filters (NF) | # Ceps (CC) |
|---|---|
| 33 | 24 |
| 17 | 12 |
| 9 | 8 |
| 5 | 4 |

Table 1: The number of filters and cepstral coefficients used for feature extraction.

### 4.2 Feature extraction

It is of interest to investigate suitable features for the GEMS signals and the useful bandwidth. Based on the spectrograms of Figure 1 this would appear to be in the region of 1 to 2 kHz thus here we investigate 4 different feature extraction parameterisations which correspond to different numbers of linear cepstral coefficients, CC, and filter banks, NF, as illustrated in Table 1. Each row of Table 1 corresponds to roughly halving NF and CC values. Similarly we assess ASR performance using different bandwidths, namely 4, 2, 1 and 0.5 kHz which then correspond to different filterbank widths when assessed in conjunction with each of the different feature parameterisations. This arrangement is intended to provide for a reasonably fair comparison of the GEMS and acoustic signals given that they occupy different bandwidths and as such will have different optimal feature parameterisations. Feature extraction was performed using G. Gravier's SPro toolkit version 4.0 [14] which provides the required functionality.

### 4.3 ASR configuration

The recogniser is that of the standard Aurora2 HTK reference system [15] which is modified for isolated rather than connected digits. The same recognition system is used to investigate acoustic and GEMS signals, the only difference being the features. We refer to the GEMS ASR case as G-ASR. Results are presented in terms of percentage word accuracy and, motivated by previous work in speaker recognition, we consider both speaker-independent and speaker-dependent configurations which we investigate by changing the training data as described below.

### 4.4 Speaker-independent recognition

For these experiments training is performed on data from 3 out of the 4 speakers and tested on data from the remaining 1 speaker in a round robin procedure. The averaged results across the 4 speakers are shown in Figure 2 for both GEMS signals, g1 and g2, and the corresponding acoustic signal, a1. The four plots correspond to the four NF and CC configurations of Table 1, with 33 NF and 24 CC in the top left plot (a), 17 NF and 12 CC in the top right plot (b), 9 NF and 8 CC

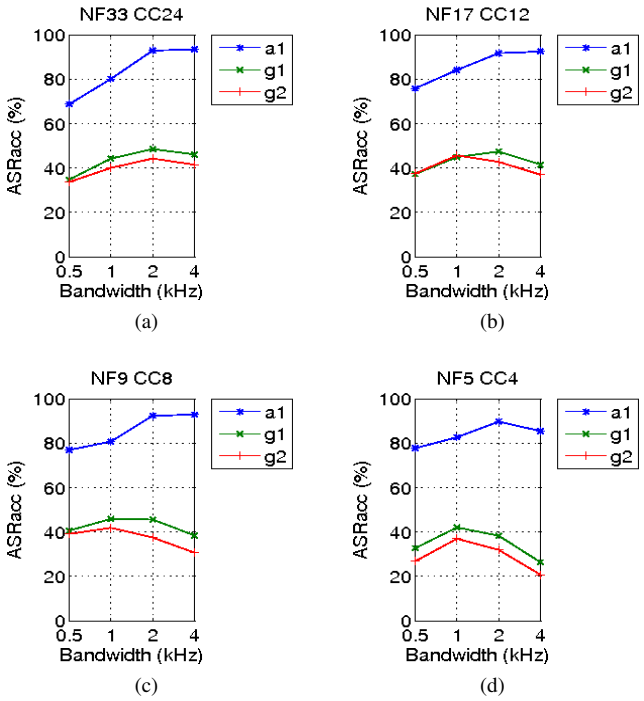


Figure 2: Speaker-independent ASR results averaged over 4 test speakers for GEMS signals g1 and g2 and the single acoustic signal a1 for (a) 33 filters and 24 cepstral coefficients, (b) 17 filters and 12 cepstral coefficients, (c) 9 filters and 8 cepstral coefficients and (d) 5 filters and 4 cepstral coefficients.

to the bottom left (c) and finally 5 NF and 4 CC to the bottom right (d). The horizontal axes indicate the total width of the filterbank and range between 0.5 and 4 kHz for each feature parameterisation considered. The vertical axes indicate percentage word accuracy.

The best GEMS result is in the top left graph (NF=33, CC=24) with a word accuracy of approximately 49% for g1 and 47% for g2 at a bandwith of 2 kHz. This corresponds to 93% for the acoustic signal. For the other three configurations the highest scores for GEMS are all at 1 kHz with the exception of g1 at 2 kHz and 47% word accuracy in Figure 2(b) for NF=17 and CC=12. Decreasing the number of filters, NF, and the number of ceptral coefficients, CC, has little effect and the word accuracy remains at aproximately 40% for a bandwidth of 1 kHz in all cases. This observation would seem to agree with the GEMS spectrogram of Figure 1 which similarly shows a bandwidth of approximately 1 to 2 kHz.

In conclusion, whilst the best score obtained with GEMS signals is approximately 50% worse than that for the corresponding acoustic signals, they are shown to contain meaningful information with a useful bandwidth in the region of 1 to 2kHz. Note that these experiments relate to SNRs in the order of 20 dB and above. For very high noise conditions the acoustic signals will be effectively lost, whereas the GEMS signals are essentially immune to noise.

| Speaker | g1 | g2 | a1 |
|---|---|---|---|
| 1 | 74% | 72% | 97% |
| 2 | 76% | 83% | 99% |
| 3 | 72% | 77% | 94% |
| 4 | 85% | 77% | 99% |
| Average | 77% | 77% | 97% |

Table 2: Speaker-dependent results for NF=33 filters, CC=24 cepstra and 2 kHz bandwidth.

### 4.5 Speaker-dependent recognition

For these experiments the ASR structure remains the same as that for the speaker-independent case, only we use speaker-dependent training data. Now the ASR system is trained for each speaker on 8 versions of each digit and tested on the remaining 1 version for that speaker, in a round robin procedure. Experiments are repeated for each of the 4 speakers using a bandwidth of 2 kHz and NF=33 and CC=24. Results are shown in Table 2, for both GEMS signals, g1 and g2, and the corresponding acoustic signal, a1.

The scores for GEMS signals are particularly interesting and show reasonable performances ranging from 72% up to 85% across the 4 speakers with averages of 77% for both GEMS signals g1 and g2. Corresponding acoustic results are all above 90% with an average of 97% across the 4 speakers.

It is well known that in higher noise conditions ASR performance falls rapidly and previous research has shown how GEMS signals can assist in offsetting this fall [5, 12]. The results reported here suggest that in high noise conditions where the acoustic signal is effectively lost, or for applications where the acoustic signal is inaccessible, a viable alternative is to replace the acoustic signals with GEMS-derived signals. Word accuracies in the order of 75% are achievable for a speaker-dependent, isolated digit ASR task and this level of performance is essentially independent of SNR.

## 5. CONCLUSIONS

Previous work reports the use of GEMS signals for speech and speaker recognition among other applications. The contribution in this paper is the first side-by-side comparison of GEMS signals to conventional acoustic signals when used independently for automatic speech recognition (ASR), i.e. when the GEMS signals are used as a substitute for acoustic signals.

A series of parallel experiments show the useful frequency range of GEMS signals to be in the order of 1 to 2 kHz. On a small isolated digit task, ASR results using GEMS signals alone show word accuracies in the order of 50% and 75% in speaker-independent and speaker-dependent modes respectively. These findings corroborate those of previous works and serve to demonstrate that GEMS signals are of use, not only for augmenting the acoustic signal, but also as a substitute. Features specific to GEMS remain to be investigated.

Even though ASR results using GEMS signals do not compare favourably to those of acoustic signals (93% and 97% word accuracies for speaker-independent and speaker-dependent modes respectively), unlike acoustic signals, GEMS signals are essentially immune to background noise.

Thus GEMS are of use for high noise applications and for situations where the acoustic signal is inaccessible.

Finally, whilst our database is of similar size to those used in all previous work, it is nonetheless acknowledged that its size is small. The collection of a large, freely available database would help to stimulate greater effort in GEMS research.

## 6. ACKNOWLEDGEMENTS

This work was sponsored by Her Majesty's Government Communications Centre (HMGCC).

## REFERENCES

[1] J. G. Beerends, E. Larsen, N. Iyer, and J. M. V. Vugt. Measurement of speech intelligibility based on the PESQ approach. *Measurement of Speech and Audio Quality in Networks (MESAQIN)*, 2004.

[2] Y. Hu and P. C. Loizou. A comparative intelligibility study of speech enhancement algorithms. In *Proc. ICASSP*, 4(4):561–564, 2007.

[3] M. Graciarena, H. Franco, K. Sonmez, and H. Bratt. Combining standard and throat microphones for robust speech recognition. In *Proc. IEEE Signal Processing Letters*, 2003.

[4] Lawrence Livermore National Laboratory. Glottal electromagnetic micropower sensor and acoustic data, 1999. http://speech.llnl.gov.

[5] C. Demiroglu and D. V. Anderson. Broad phoneme recognition in noisy environments using the GEMS device. In *Proc. Asilomar Conf. on Signals, Systems and Computers*, volume 2, pages 1805–1808, 2004.

[6] T. F. Quatieri, K. Brady, D. Messing, J. P. Campbell, W. M. Campbell, M. S. Brandstein, C. J. Clifford, J. D. Tardelli, and P. D. Gatewood. Exploiting nonacoustic sensors for speech encoding. *IEEE Trans. on ASLP*, 14(2):533–544, 2006.

[7] G. C. Burnett. The physiological basis of glottal electromagnetic micropower sensor (GEMS) and their use in defining an excitation function for the human vocal tract. *PhD Thesis, University of California*, 1999.

[8] T. J. Gable. Speaker verification using acoustic and glottal electromagnetic micropower sensor (GEMS) data. *PhD Thesis, University of California*, 2000.

[9] C. Demiroglu, S. D. Kamath, and D. V. Anderson. Segmentation-based speech enhancement for intelligibility improvement in MELP coders using auxiliary sensors. In *Proc. ICASSP*, volume 1, pages 797–800, 2005.

[10] W. M. Campbell, T. F. Quatieri, J. P. Campbell, and C. J. Weinstein. Multimodal speaker authentication using nonacoustic sensors. In *Proc. Workshop Multimodal User Authenatication*, pages 215–222, 2003.

[11] L. C. Ng, G. C. Burnett, J. F. Holzrichter, and T. J. Gable. Denoising of human speech using combined acoustic and EM sensor signal processing. In *Proc. ICASSP*, volume 1, pages 229–232, 2000.

[12] B. Raj and R. Singh. Feature compensation with secondary sensor measurements for robust speech recognition. In *Proc. EUSIPCO*, 2005.

[13] L. N. Supplee, R. P. Cohn, J. S. Collura, and A. V. McCree. MELP: The new federal standard at 2400 bps. In *Proc. ICASSP*, volume 2, pages 1591–1594, 1997.

[14] G. Gravier. SPro: Speech signal processing toolkit v4. Available at http://gforge.inria.fr/projects/spro.

[15] H. G. Hirsch and D. Pearce. The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. *ISCA ITRW ASR2000 'Automatic Speech Recognition: Challenges for the next Millenium'*, 2000.