Multimodal Speaker Authentication using Nonacoustic Sensors*

W. M. Campbell, T. F. Quatieri, J. P. Campbell, C. J. Weinstein
MIT Lincoln Laboratory
{wcampbell, tfq, jpc, cjw}@ll.mit.edu

Abstract

Many nonacoustic sensors are now available to augment user authentication. Devices such as the GEMS (glottal electromagnetic micro-power sensor), the EGG (electroglottograph), and the P-mic (physiological mic) all have distinct methods of measuring physical processes associated with speech production. A potential exciting aspect of the application of these sensors is that they are less influenced by acoustic noise than a microphone. A drawback of having many sensors available is the need to develop features and classification technologies appropriate to each sensor. We therefore learn feature extraction based on data. State of the art classification with Gaussian Mixture Models and Support Vector Machines is then applied for multimodal authentication. We apply our techniques to two databases—the Lawrence Livermore GEMS corpus and the DARPA Advanced Speech Encoding Pilot corpus. We show the potential of nonacoustic sensors to increase authentication accuracy in realistic situations.

1. Introduction

Speaker authentication is a rich area for exploration of multimodality. Many facets of the speech production process are measurable through a variety of sensors. Traditionally, visual lip reading has been used to supplement speaker authentication and speech recognition [15,26]. These methods rely upon tracking the lip contour over time and then using the sequence of movements to supplement standard audio-only verification. These methods have been quite successful, leading to large gains in accuracy in high noise conditions.

Other methods of monitoring speech production are also available. Non-invasive sensors that are attached in the throat area have been available for many years; we call these nonacoustic sensors. These sensors nominally measure aspects of the speech production process related to the speech excitation. Typical sensors that we have explored in this study are the EGG (electroglottograph), the GEMS (glottal electromagnetic micro-power sensor), and the P-mic (physiological mic). Since traditional methods of verification [18] rely upon features designed to capture vocal tract information—e.g., mel-frequency cepstral coefficients—we would expect that multimodal fusing of excitation and vocal tract features would benefit recognition in *both* quiet and noisy conditions. An added benefit of nonacoustic sensors is that they are less influenced by acoustic noise. For the case of the EGG and the GEMS, the throat is exposed to RF signals; for the case of the P-mic, the sensor output is dominated by the vibrations sensed on the throat. These modes of measurement do not directly monitor air pressure in the ambient environment.

There has been several prior works on the use of glottal waveforms for recognition. Gable [8] used waveforms from the GEMS system for speaker verification; his work focused on using methods such as dynamic time warping for text-dependent verification. Plumpe [16] used inverse filtering techniques on the acoustic waveform to derive glottal waveform signals; speaker recognition was then performed. Both throat microphones [9] and the P-mic [1] have been used for automatic *speech* recognition. Our work is distinct in several aspects: 1) we consider both simulated and actual noise conditions, 2) we do not assume models for the glottal waveforms but instead use a learning approach, 3) we use late integration to combine *several* nonacoustic sensors, and 4) we consider integration accuracy of multiple nonacoustic sensors in low-noise conditions.

We attack the problem of authentication using nonacoustic sensors with a data-driven learning approach. We have chosen the data-driven approach as a baseline to future knowledge-based analysis. Sensor outputs can vary dramatically based on placement, sensor tuning, impedance matching, sensor design, etc. This variation can be captured easily with data-driven methods. Towards this end, we use standard feature transformation methods to find features which describe the speaker specific attributes of the different signals. We use various normali-

*This work is sponsored by the Defense Advanced Research Projects Agency under Air Force contract F19628-00-C-0002. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
DEC 2003		**00-00-2003 to 00-00-2003**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Multimodal Speaker Authentication using Nonacoustic Sensors	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Massachusetts Institute of Technology,Lincoln Laboratory,244 Wood Street,Lexington,MA,02420-9185**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
In Proc. Workshop on Multimodal User Authentication in Santa Barbara, California, pp. 215-222, 11-12 December 2003.

14. ABSTRACT
see report

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**8**	

zations based upon signal characteristics to improve accuracy.

After obtaining features for authentication, we use both Gaussian Mixture Models [18] and Support Vector Machines (SVM's) [25] for multimodal authentication. We combine the outputs of these different classification systems using late integration to achieve the final score. For the corpora explored in this paper, we consider only closed-set speaker identification. That is, given an utterance, identify an individual from a list of known individuals. Because of the limited number of speakers available in current corpora, other scenarios such as verification or open-set ID were impossible because of the lack of an adequate "background" population.

The outline of the paper is as follows. In Section 2, we discuss the sensors in detail and describe their basic operation. In Section 3, we discuss our feature extraction methodology. Section 4 outlines the classifiers and fusion strategy used. Section 5 gives details on the corpora used and experiments. These corpora allow us to explore both the GEMS in quiet environments and multiple nonacoustic sensors in high noise (>110 dBC) situations. We show that our authentication strategy leads to gains in this challenging scenario. A complimentary method for achieving authentication accuracy gains is speech enhancement [27].

2. Nonacoustic sensors

We survey three nonacoustic sensors used for experiments—GEMS, EGG, and P-mic. These sensors have distinct methods of measuring speech production phenomena. Other sensors which would be of interest, but were not included due to corpus size and project focus, are accelerometers, "bone phones," in-ear microphones, video, etc.

2.1. GEMS

The GEMS (glottal electromagnetic micro-power sensor) is a novel sensor based upon transmitting electromagnetic (EM) waves into the glottal region. Two GEMS designs were used in the corpora in this paper. An earlier version was used in the LLNL Corpus [8], and Revision B, Version 1 created by Aliph Corporation (http://www.aliph.com) was used in the ASE Corpus of Section 5. The GEMS is also referred to as the "General Electromagnetic Movement Sensor" by Aliph Corporation.

During operation of the GEMS, a small antenna is placed on or near the throat at the level of the glottis. From this antenna is transmitted a 2.3 or 2.4 GHz low power (<1 mW) EM wave. Using these frequencies allows for EM waves to penetrate into the body and reflect back to the sensor with good signal levels. The receiver circuitry detects the reflected EM waves using a homodyne technique. Nominally, the sensor measures phenomena related to the opening and closing of the glottis [2]. Multiple theories have emerged on the exact phenomena occurring that generates the waveform—changing air-tissue interfaces as the glottis changes, vibration of the tracheal wall, and propagation along the vocal fold contact area, see [11, 21]. Although inferring the exact process that the GEMS is monitoring is challenging, the waveforms generated do provide speaker specific information which is related to the speech excitation.

2.2. EGG

The EGG (electroglottograph) is a device designed to measure contact between the vocal folds. The specific implementation used for this study was from Glottal Enterprises. This EGG is a multi-channel EGG device [19]; the multichannel feature allows for more precise placement on the neck to achieve higher signal to noise ratio.

The EGG nominally measures the vocal fold contact *area* (VFCA). This process is performed by using electrical signals in the MHz region. Two electrodes are placed on the subject's neck at the level of the thyroid cartilage. VFCA is measured by observing the variation in impedance over time. Since the EGG measures vocal fold contact, the sensor does not necessarily allow one to observe interesting phenomena during the open phase of the glottis. Note that the EGG is not an exact indicator of VFCA. For example, during transition to the open phase of the glottis, mucus can "short out" the device indicating that the glottis is closed when this is apparently not the case (the mucus bridging effect [4]).

2.3. P-mic

The P-mic (physiological microphone) is a non-invasive contact sensor for measuring sound [20]. The P-mic consists of a gel pad to provide acoustic impedance matching, a conical focusing aperture, and a piezoelectric element. Use of a gel pad minimizes interference from ambient noise.

The P-mic is typically placed in the throat area below the glottis. This placement insures that the P-mic signal can be simultaneously recorded with the GEMS and EGG signal. In our experiments, we found that the P-mic was most sensitive to ambient noise among nonacoustic sensors; presumably this is due to "leakage" of the ambient noise into the sensor element.

2.4. Comparison of the sensors

Figure 1 shows an example output from four sensors recorded simultaneously. In the figure, the top signal is a microphone recording of the /ao/ in "dog." The second signal represents the EGG signal (highpass filtered with a linear phase filter with a transition band from 64-80 Hz). We note that the EGG gives a very "smooth" waveform. The third waveform from the top is the P-mic signal. In this signal, we see more evidence of "leakage" of vocal tract information into the signal (as evidenced by ripple in the waveform). Finally, the fourth waveform is the GEMS signal. We can see this waveform has many of the same general characteristics as the EGG, but that there is additional structure in the waveform. Listening to the GEMS signal reveals little vocal tract information; therefore, this fine structured seems to represent supplementary excitation information not captured by the EGG.

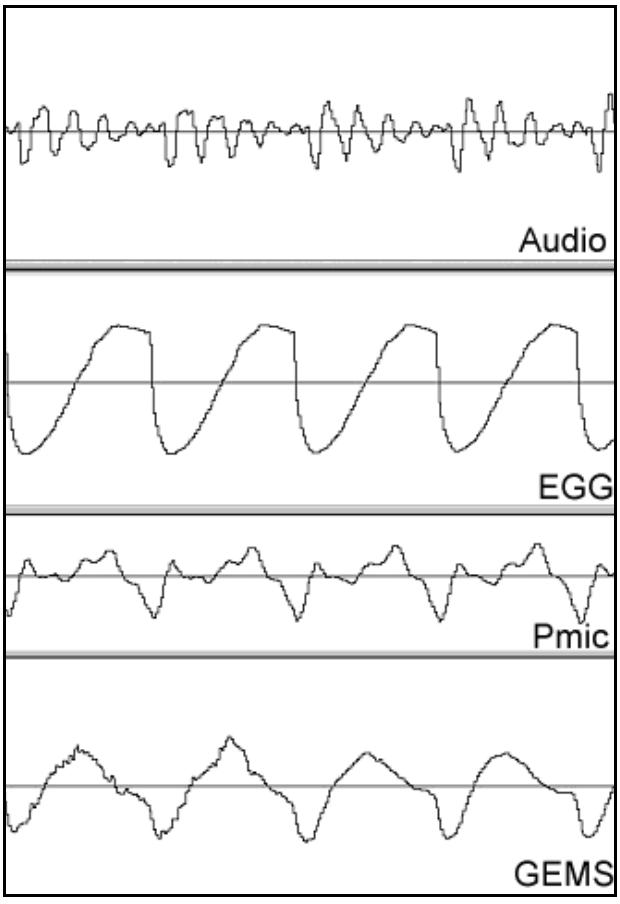

Figure 1. Comparison of different sensor waveforms for the /ao/ in "dog." From top to bottom—audio, EGG, P-mic, and GEMS. The length of time shown is approximately 30 ms.

3. Feature extraction

Our framework for feature extraction is shown in Figure 2. Our goal was to create a flexible architecture that incorporated linear matrix transformation for feature extraction. In the figure, the input signal is processed into frames creating a sequence of vectors. Each frame corresponds to a 30 ms time window with an overlap of 20 ms between consecutive frames. Since our sampling rate is 8 kHz, we obtain a sequence of vectors of dimension 240 (100 vectors per second).

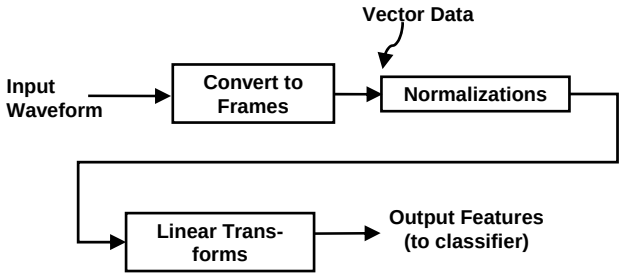

Figure 2. Framework for feature transformation.

We then applied several normalizations to the data; these normalizations are intended to provide invariances in the feature extraction to certain transforms—e.g., increasing the gain. We first remove the mean on a per frame basis; we then normalize the amplitude of the signal variance to 1. Finally, we introduce a transform to reduce a framing artifact; namely, a shift of the input should not matter in recognition. For this normalization, we calculate the discrete Fourier transform (DFT) of each frame, eliminate the phase of each component, and then calculate the inverse DFT. All of these normalizations are intended to throw out unnecessary signal information; potentially, they are too aggressive and could be modified. For example, the mean of the EGG signal carries information about the position of the larynx. In spite of drawbacks, these normalizations increased accuracy for all linear transforms we tried.

After appropriate normalization, the sequence of frames was used to calculate delta parameters [17]. This linear transform resulted in a sequence of vectors of dimension 480. We then wanted to design a linear transform to reduce this 480 component vector to a more reasonable dimension. There are multiple reasons for dimension reduction—obtaining compact representations of speaker specific features, avoiding excessively complex classifiers, discarding "uninformative" directions in feature space, and minimizing the "curse of dimensionality." For this paper, we explored several unsupervised methods of designing a linear transform—principal component analy-

sis (PCA) [7], random dimension reduction [6], and independent component analysis (ICA) [12].

Random dimension reduction (i.e., generating the analysis matrix using random independent components) was used for multiple purposes. We preprocessed all of the normalized outputs (with delta components) from dimension 480 down to dimension 100 using random dimension reduction. As shown in [6], random dimension reduction tends to preserve distances and make clusters of data more spherical which improves problem conditioning. We found that for both PCA and ICA that this improved accuracy. Random dimension reduction also reduces the size of the problem making methods such as ICA and PCA more practical for large problems. Finally, random dimension reduction was also used as an analysis method to compare to other unsupervised methods.

We note that our feature transformation method is very similar to the standard filter bank approach for generating mel-cepstral coefficients. In a coarse sense, our approach could be thought of as applying a filter bank "tuned" to the glottal response.

4. Classification and fusion

Gaussian mixture models have been very successful for the speaker recognition task [18]. We use Gaussian mixture models to model the speaker specific distribution only (i.e., no background modeling is performed since our task is closed-set identification). For each speaker, we create a mixture model

$$f(\mathbf{x}) = \sum_{i=1}^{n} \lambda_i g_i(\mathbf{x})$$

where g_i is a single Gaussian. Training is accomplished using the EM algorithm with a small number of components—typically less that 256.

We also use support vector machines (SVM's) for classification [25]. Support vector machines are discriminatively trained classifiers and thus give excellent performance on closed set tasks. For our experiments, we use a polynomial basis of monomials in our SVM kernel up to and including a certain degree—typically degree 2 or 3, see [25]. Our SVM kernel is based upon comparing sequences of data and providing an inner product in a large dimensional space which captures speaker specific information. One interesting aspect of using support vector machines for our work is that it is possible to bypass the feature transformation process and perform classification directly in high dimensions. Although this is computationally intense, it gives a baseline for feature transformed classification systems which work in lower dimensions.

All of our reported experiments use late integration for fusion [3]. Fusion is accomplished by using a linear combination of scores from each of the classifiers applied to the different modalities. Methods involving construction of new SVM kernels based upon sums of kernels for each of the modalities were also tried, but these did not perform as well as late integration.

5. Corpora and experiments

5.1. LLNL GEMS corpus and experimental setup

The first corpus used for experiments was the Lawrence Livermore National Lab GEMS corpus collected by G. Burnett and T. Gable [8]. This corpus consists of 15 male speakers with up to 4 sessions per speaker. Both sentences from TIMIT and number/letter/{Yes,No,Zero} sequences were recorded. For the purposes of our experiments, we focused on the number/letter/short-word sequences. Typical utterances were a combination of 10 items; e.g., "T 60 YES 3 U R E 8 W P."

We used the initial session of 20 utterances as enrollment. The remaining 3 sessions of 20 utterances each were used for speaker identification. This resulted in 15*60=900 tests for speaker identification. Both audio and GEMS data were originally sampled at 10 kHz. We resampled to 8 kHz and then bandlimited the speech to 200-4000 Hz.

Noise was electronically added to the audio signal with noises from the NOISEX database [23]. (In Section 5.3 and 5.4, we consider a corpus where the noise environment is not electrically added.) The NOISEX noise signals were resampled to 8 kHz and also bandlimited to 200-4000 Hz. This insured that SNR was measured only in the band containing speech. All 24 NOISEX noises were used. When adding speech to noise, we generated a random offset into the noise file and then extracted a segment of noise the same length as the speech file. The resulting output signal was $x = x_{\text{speech}} + c * x_{\text{noise}}$, where

$$c = \frac{\sigma_{speech}}{\sigma_{noise}} 10^{-\frac{\text{SNR}}{10}}$$

and the standard deviations are calculated over non-silence regions.

5.2. LLNL corpus results

Our first set of experiments compared feature transformation methods. As indicated in Section 4, we explored random dimension reduction, PCA, and ICA. We initially considered closed-set speaker identification accuracy based upon the GEMS signal only. Each feature vector was reduced from dimension 480 to 100 using

Table 1. Comparison of accuracy of feature transformation methods for GEMS-only closed-set speaker identification on the LLNL database.

Feature Extraction Method	Speaker Identification Accuracy (%)
Random Projection	62.7 %
PCA	59.7 %
ICA	51.9 %
None	64.3 %

random dimension reduction. A linear transform was then designed and applied to reduce the dimension from 100 to 32 for input to the classifier. Dimension 32 was chosen since the accuracy typically plateaued at this dimension. A SVM classifier with a degree 2 polynomial kernel (full covariance) was used, see [25].

Table 1 compares accuracies for the different methods. Also included in the table is the case of no dimension reduction (with a diagonal covariance SVM kernel) which provides a baseline for reduced dimension methods. As can be seen from the table, random projection works as well as other transformation methods. Potentially, this is due to multiple factors. The classifier may be better matched to this feature extraction technique. Also, there could be spurious directions in the feature space data which are not relevant to speaker identification. One way to mitigate this problem (which we do not explore here) is to use supervised feature transformation methods, e.g. [22].

After using linear transform feature extraction methods for speaker identification, we investigated the use of fundamental frequency (F0) to augment the recognition process. The Entropic pitch extractor in Wavesurfer (http://www.speech.kth.se/wavesurfer) was used. A GMM was trained with 32 components to model each speaker from the F0 data. The resulting error rate for GEMS only recognition was 50.6%. Note that a similar rate of accuracy was also observed for the audio data using F0 only—49.1%.

We then fused (with equal weights) the GEMS F0 classifier scores with the linear transform feature extraction scores (random dimension reduction) to obtain a GEMS-only accuracy of 64.0%. The use of F0 information demonstrated two items. First, since F0-only classification accuracy is significantly below that of linear transform feature extraction accuracy, we are obtaining additional non-F0 information from our linear transform technique.

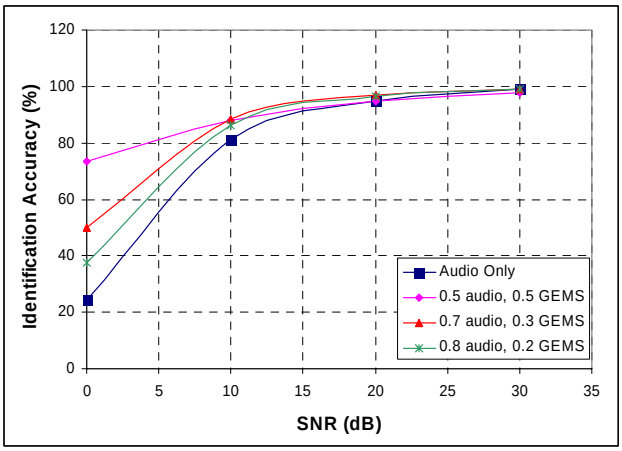

Figure 3. Comparison of speaker identification accuracy across noise type 3 (white noise) for different late integration strategies and random dimension reduction.

Second, because the accuracy improved from the fusion, there is complementary information in the two scores.

Finally, we considered the effect of late integration upon speaker identification in noise. We implemented an audio-only speaker recognition system using the system in [25] with a degree 3, diagonal covariance model; input features were 12 LP cepstral coefficients plus deltas. In addition, the MELPe noise preprocessor [24] was applied to the audio input signal. Figure 3 shows the performance of a late integration system which fuses an audio-based system with the GEMS-based system (both pitch and linear feature transformation were used). In the figure, at low SNR (0-10 dB) and for NOISEX white noise (noise type 3), significant increases in accuracy are obtained by late integration—greater than 50% in some cases.

We then considered the effect of late integration with a fixed weighting, 0.5*GEMS + 0.5*audio, as the type of noise varied for a fixed SNR (specific information on the noise types can be found in the NOISEX corpus documentation). The results for 0 dB SNR are shown in Figure 4. As can be seen from the figure, significant increases in accuracy over an audio-only system are achieved—greater than 25% average improvement. The best performing environments were NOISEX types 3 (white noise), 16 (machine gun), 18 (STI test signal), 19 (voice babble), and 21 (factory). The worst performing environments were NOISEX types 1 (sinusoid), 5 (colored, -12 dB/octave), 9 (Leopard 2), 23 (Car) and 24 (Car).

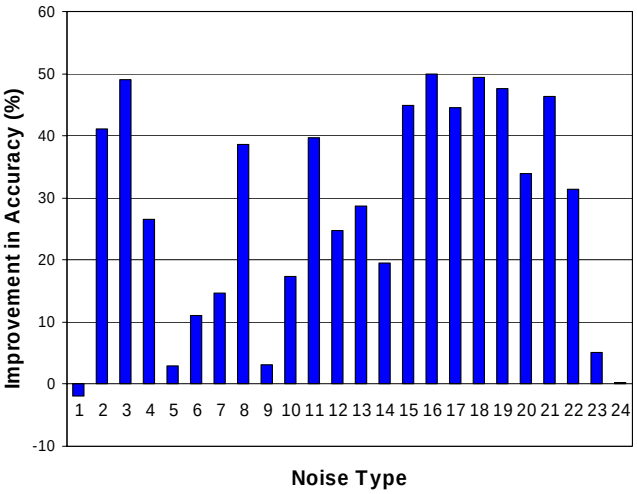

Figure 4. Improvement in speaker identification accuracy of a late-integration fusion system over an audio-only system by noise type (NOISEX database) at 0 dB SNR.

5.3. ASE corpus and experimental setup

The Advanced Speech Encoding Pilot Corpus (ASE Pilot Corpus) is a multisensor corpus collected for the purpose of studying viability of multiple sensors for speech enhancement, speech coding, and speaker characterization. Sensors recorded simultaneously include a resident microphone (the microphone typically used in the environment), two channels of a GEMS device, an EGG, a high quality reference microphone (B&K), and P-mics positioned on the forehead and the throat region. The corpus was collected in two sessions (on two different days). Speakers were exposed to a variety of noise environments—-quiet, office (56 dBC), MCE (mobile command enclosure, 79 dBC), M2 Bradley Fighting Vehicle (74 dBC and 114 dBC), MOUT (military operations in urban terrain, 73 dBC and 113 dBC), and a Blackhawk helicopter (70 dBC and 110 dBC). We call these environments (with L indicating low noise and H indicating high noise) quiet, office, MCE, M2L, M2H, MOUTL, MOUTH, BHL and BHH, respectively. To protect our subjects and realistically simulate Lombard effects, all talkers used the hearing protection systems typical of each environment. This normally consisted of a communication headset with approximately 20 dB noise attenuation. Human subject testing procedures were followed carefully and noise exposure was monitored.

For speaker identification experiments, we partitioned the corpus by session. The initial sessions—quiet, office, and MCE—were used for enrollment. Identification was then performed using the data from the remaining sessions; we grouped these into low noise—M2L, MOUTL, BHL—and high noise—M2H, MOUTH, BHH—conditions. The corpus had phrases in both sessions drawn from a variety of material—conversations, DRT lists, vowels, Harvard phonetically balanced sentences, and CVC nonsense words. Typical utterance lengths ranged from 1-5 minutes. A total of 20 speakers were available, 10 males and 10 females. The total number of enrollment utterance available per speaker was 12. The total number of tests for identification performance was 360 per noise condition (low, high). Cross-gender testing was allowed since it was not clear if the nonacoustic sensors would distinguish this well; cross-gender tests do not bias identification accuracy (as they would in speaker verification).

5.4. ASE corpus results

The feature extraction methods from Section 3 were applied to the ASE pilot corpus. As for the experiments in Section 5.2, we used a SVM with diagonal covariance and degree 3 polynomials for the audio modality. For the nonacoustic modalities, we used a full covariance SVM of degree 2 with random dimension reduction. Both the MELPe noise preprocessor and high-pass filtering above 200 Hz were applied to the audio signal. The MELPe noise preprocessor was applied to the non-acoustic modalities, since noise from the ambient environment did effect the sensor outputs (possibly through tissue vibration). The EGG was highpass filtered with a linear phase filter with transition band from 64-80 Hz. Results are shown in Table 2.

Since the P-mic has some vocal tract information (as evidenced by listening), we also applied a standard LP cepstral coefficient front end to the data; i.e., we applied the audio recognition system to all sensors. Results for this set of experiments are shown in Table 3. As can be seen from the table, accuracy results for both the EGG and GEMS are generally lower for LP cepstral coefficients than with data driven methods shown in Table 2. For the P-mic, the identification accuracy is higher for LPCC's; this illustrates that standard methods are tuned to extracting vocal tract information.

Table 2. Identification accuracy in both low and high noise situations for multiple modalities using random dimension reduction.

Modality	Low Noise Accuracy	High Noise Accuracy
EGG	73.0 %	43.3 %
GEMS	64.7 %	43.6 %
P-mic	66.7 %	41.4 %

Table 3. Identification accuracy in both low and high noise situations for multiple modalities using LP cepstral coefficients.

Modality	Low Noise Accuracy	High Noise Accuracy
Resident Mic	89.4 %	81.9 %
EGG	61.1 %	38.0 %
GEMS	50.3 %	43.6 %
P-mic	77.5 %	55.0 %

Table 4. Identification accuracy in both low and high noise situations for late integration fusion.

Modalities Fused	Low Noise Accuracy	High Noise Accuracy
Audio (Resident Mic)	89.4 %	81.9 %
0.8*Audio+0.2*EGG	93.1 %	86.7 %
0.8*Audio+0.2*GEMS	92.5 %	85.8 %
0.5*Audio+0.5*P-mic	95.8 %	87.2 %
All	95.8 %	89.4 %

Two items should be noted about the results in Tables 2 and 3. First, the accuracy of the resident microphone is somewhat low in low noise situations. This result is probably due to mismatch in microphones between training and testing. Second, high-noise accuracy of the resident microphone is quite good. The MELPe noise preprocessor and associated processing is fairly robust to noise.

Another observation from Tables 2 and 3 is the degradation of nonacoustic sensors in noise. For the GEMS modeling in Section 5.2, we assumed the ideal case of no degradation due to noise. It is well known in the literature [5], that even if acoustic noise is not present in the sensor data, a human speaker responds to the environment, e.g. Lombard effect [13]. This response to stress will cause degradation in the speaker identification performance of the nonacoustic modalities. An open research question is how to compensate for the effects of stress in the excitation parameterization. Although we do not explore methods here, the ASE pilot corpus provides a realistic scenario for studying methods of noise compensation of the speech excitation waveform.

Table 4 shows the results of late integration. For the EGG and GEMS, fusion with the weights shown and random dimension reduction yielded the best results. For the P-mic, LPCC's performed the best with equal weighting of audio and P-mic modalities. For the fusion of all modalities, we tried a variety of weightings; the best performing weighting was 0.5*audio, 0.2*EGG, 0*GEMS, and 0.3*P-mic (labeled "All" in Table 4). Unfortunately, a cross-validation data set was not available to validate the fusion process.

As indicated in Table 4, we obtain substantial gains of 7.5% in speaker identification accuracy in noise, over the resident-microphone-only case by combining nonacoustic and acoustic scores. This result shows the potential of these methods for noise robust speaker authentication.

6. Conclusions

We have demonstrated the use of nonacoustic sensors for speaker authentication. A data-driven approach was used to derive features of different modalities. Powerful classification techniques such as support vector machines and Gaussian mixture models were then applied. Results in both simulated and actual noisy conditions showed the success of the techniques for dramatically improving speaker authentication in noise. Future work should explore methods on statistically-significant larger speaker populations to further validate results.

Acknowledgements

We thank John Tardelli and Paul Gatewood of ARCON Corporation for their excellent work in collecting the ASE multisensor corpus used for experiments in this paper. We thank Kevin Brady of MIT Lincoln Laboratory for his extensive support in this corpus collection. We thank Doug Reynolds for consultation on speaker recognition.

References

[1] Bass, J. D., M. V. Scanlon, T. K. Mills and J. J. Morgan, "Getting two birds with one phone: an acoustic sensor for both speech recognition and medical monitoring," *presentation at 138th meeting of the Acoustical Society of America, Columbus, OH,* 1999.

[2] Burnett, G. C., The physiological basis of Glottal Electromagnetic Sensors (GEMS) and their use in defining an excitation function for the human vocal tract, PhD Thesis, University of California, Davis, 1999.

[3] Chen, T. and R. R. Rao, "Audio-Visual integration in multimodal communication," *Proceedings of the IEEE,* 1998, pp. 837-852.

[4] Childers, D.G and A. K. Krishnamurthy, "A critical review of electroglottography", CRC Critical Reviews in Biomedical Engineering, 12, 1985, pp. 131-161.

[5] Cummings, K. E. and M. A. Clements, "Estimation and comparison of the glottal source waveform across stress styles using glottal inverse filtering," *Proceedings Southeastcon,* 1989, pp. 776-781.

[6] Dasgupta, S., "Experiments with Random Projection," *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*, 2000, pp. 143-151.

[7] Fukunaga, K., *Introduction to Statistical Pattern Recognition*, Academic Press, Inc., San Diego, CA, 1990.

[8] Gable, T.J., *Speaker Verification Using Acoustic and Glottal Electromagnetic Micro-power Sensor (GEMS) Data*, PhD Thesis, University of California, Davis, 2000.

[9] Graciarena, M., H. Franco, K. Sonmez and H. Bratt, "Combining standard and throat microphones for robust speech recognition," *IEEE Signal Processing Letters*, vol. 10, no. 3, 2001, pp. 72-74.

[10] Holzrichter, J. F., G. C. Burnett, L. C. Ng and W. A. Lea, "Speech articulator measurements using low power EM-wave sensors," *Journal of the Acoustical Society of America*, 1998, 103(1), pp. 622-625.

[11] Holzrichter, J. F., L. C. Ng, G. J. Burke, N. J. Champagne II, J. S. Kallman, R. M. Sharpe, J. B. Kobler, R. E. Hillman and J. J. Rosowski, "EM wave measurements of glottal structure dynamics," *University of California, Lawrence Livermore Laboratory Report*, UCRL-JC-147775 , 2002.

[12] Hyvarinen, A., "Fast and Robust Fixed-Point Algorithms for Independent Component Analysis," *IEEE Trans. On Neural Networks*, vol. 10, no. 3, 1999, pp. 626-634.

[13] Lippmann, R. P., E. A. Martin, "Multi-Style Training for Robust Isolated-Word Speech Recognition," *Proceedings IEEE International Conference on Acoustics Speech and Signal Processing*, 1987, pp. 705-708.

[14] Lawrence Livermore National Lab, Glottal Electromagnetic Micropower Sensor and Acoustic Data, http://speech.llnl.gov, 1999.

[15] Luettin, J., N. Thacker and S. Beer, "Speaker Identification by Lipreading," *Proc. ICSLP*, 1996, pp. 62-64.

[16] Plumpe, M. D., T. F. Quatieri and D. A. Reynolds, "Modeling of the Glottal Flow Derivative Waveform with Application to Speaker Identfication," *IEEE Trans. Speech and Audio Processing*, vol. 7, no. 5, 1999, pp. 569-586.

[17] Rabiner, L. and B.-H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, Englewood Cliffs, NJ, 1993.

[18] Reynolds, D. A., "Speaker Identification and Verification using Gaussian Mixture Speaker Models," *Speech Communication*, vol. 17, 1995, pp. 91-108.

[19] Rothenberg, M. "A Multichannel Electroglottograph," *J. of Voice*, 1992, vol. 6, no.1, pp. 36-43.

[20] Scanlon, M. V., "Acoustic Sensor for Health Status Monitoring," *Proceeding of IRIS Acoustic and Seismic Sensing*, 1998, Volume II, pages 205-222.

[21] Titze, I. R., B. H. Story, G. Burnett, J. F. Holzrichter, L. C. Ng, W. A. Lea, "Comparison between electroglottography and electromagnetic glottography," *J. Acoust. Soc. Am.*, vol. 107, no. 1, 2000, pp. 581-588.

[22] Torkkola, K. and W. Campbell, "Mutual information in learning feature transformations," *Seventeenth International Conference on Machine Learning*, 2000, pp. 1015-1022.

[23] A.P. Varga, H.J.M Steenekan, M. Tomlinson, and D. Jones, "The NOISEX-92 study on the effect of additive noise on automatic speech recognition," *Tech. Rep., DRA Speech Research Unit*, 1992.

[24] Wang, T., K. Koishida, V. Cuperman, A. Gersho and J. S. Collura, "A 1200/2400 BPS Coding Suite Based on MELP," NATO AC/322(SC/6-AHWG/3) AD HOC Working Group on Narrow Band Voice Coding, 2002 IEEE Workshop on Speech Coding, Special Session 1: Topics on NATO Standardization, Tsukuba, Ibaraki, Japan, October 6-9, 2002.

[25] Campbell, W. M., "Generalized Linear Discriminant Sequence Kernels for Speaker Recognition," *Proceedings of ICASSP*, 2002, pp. 161-164.

[26] Zhang, X., C. C. Broun, R. M. Mersereau and M. A. Clements, "Automatic Speechreading with Applications to Human-Computer Interfaces," *Eurasip Journal on Applied Signal Processing*, 2002, pp. 1228-1247.

[27] Quatieri, T. F., D. Messing, K. Brady, W. M. Campbell, J. Campbell, M. Brandstein, C. Weinstein, J. Tardelli, and P. Gatewood, "Exploiting non-acoustic sensors for speech enhancement," *submitted to Workshop on Multimodal User Authentication.*