SPEAKER VERIFICATION USING COMBINED ACOUSTIC AND EM SENSOR SIGNAL PROCESSING[1]

Ng, L. C., Gable, T. J., Holzrichter, J. F.
Lawrence Livermore National Laboratory and University of California, Davis
P.O. Box 808, L-491
Livermore, California 94550 USA

ABSTRACT

Low Power EM radar-like sensors have made it possible to measure properties of the human speech production system in real-time, without acoustic interference. This greatly enhances the quality and quantity of information for many speech related applications. See Holzrichter, Burnett, Ng, and Lea, J. Acoustic. Soc. Am. 103 (1) 622 (1998). By combining the Glottal-EM-Sensor (GEMS) with the Acoustic-signals, we've demonstrated an almost 10 fold reduction in error rates from a speaker verification system experiment under a moderate noisy environment (-10dB).

1. INTRODUCTION

Acoustic speech signals carry a great deal of information that can be automatically converted to text, coded for transmission, and many other applications. However, under conditions with a great deal of background noise, with speakers who do not speak clearly (e.g., who co-articulate, or incompletely articulate, etc.) or who speak with strong accents, such systems often do not work adequately. Many mechanisms, by which additional information, describing conditions of the vocal articulators as the speech signal is generated, have been examined to increase the accuracy of automated systems. Examples are TV images of the lip opening, jaw open-close sensors, electro-glottalgraph signals of the vocal fold conditions, etc.

Recently, it has been shown that very low power Electro Magnetic (EM) radar-like sensors can measure conditions of many of the internal (and external) vocal articulators and vocal tract parameters, in real-time, as speech is generated, Holzrichter (1). In particular, a voiced excitation function of speech has been obtained by associating EM sensor signals from the glottal region (i.e., Glottal Electro Magnetic Sensors, or GEMS) with sub- or supra-glottal air pressure pulsations, Burnett (2). These data, combined with corresponding acoustic data, enable robust methods for sampling background noise data, and vastly increase the quality and quantity of information for almost all applications involving speech processing and use.

In addition, these techniques enable accurate definitions of time periods of phonation, and using the statistics of the user's language (3) enable the definition of periods preceding and following phonation when unvoiced speech is likely to occur. In addition, they enable the determination of periods of no speech, when sampling of background noise signals can reliably take place. Along with robust speech presence determination, the timing and spectral content of the determined excitation function enable real-time filters to be constructed for purposes of denoising corresponding acoustic signal segments.

2. HOMODYNE SENSORS

EM radar-like sensors have been designed to transmit EM waves at 2.3 GHz with 0.2 mW of total power. This level is well below continuous international exposure standards for human use. The sensors use a homodyne field disturbance mode of operation that resembles an interferometer measuring the reflection of a transmitted wave against a local (phase reference) wave. As a reflecting interface moves, the phase of the reflected wave varies with respect to the stationary local wave, and a signal associated with this change is detected by a mixer and filter combination. The EM sensor positioned near the glottis in Figure 1 measures the positional changes of glottal tissues, as the air/tissue interface moves versus time, driven by air pressure waves from the glottis opening and closing. Also Figure 1 shows an EM sensor signal that characterizes the glottal tissue interface motion versus time, and which can be associated with the voice excitation function [2].

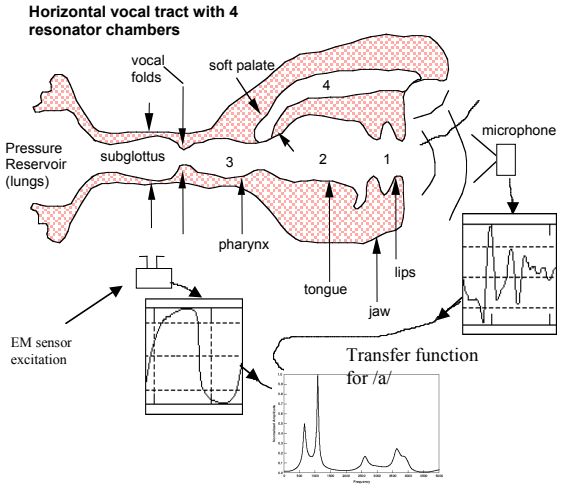

Figure 1. A linearized vocal tract showing locations of EM sensors, corresponding excitation and acoustic functions, and resulting transfer function.

3. APPLICATIONS OF GEMS TO SPEAKER VERIFICATION

This section briefly outlines the method of extraction for each of the different speaker verification parameters. These parameters form a set of feature vectors used in the dynamic time warping algorithm to calculate the performance "distance" used to make the accept/reject decision on an identity claim. Verification parameters represent the individuality of the speaker, containing information about the timing, pitch, amplitude or

[1] Last update: November 10, 2000

spectral content of the speech. A conventional speaker verification used features derived from Cepstral coefficients [5]. However, the GEMS sensor provides additional, uncorrelated, acoustically noise free, features that include: GEMS pitch, GEMS shape parameter (GSP), auto-regressive and moving average (ARMA) coefficients.

GEMS Pitch Extraction

The use of the GEMS signal enables great speed and accuracy in pitch estimation. Figure 2 shows a sample three-glottal cycle graph of GEMS data. The smoothness of the GEMS signal and the linearity of the signal during the positive-to-negative zero crossing allows the use of a simple interpolated zero-crossing algorithm. The algorithm searches for the positive to negative crossing of the signal. Burnett (1999) developed the GEMS pitch algorithm used in support of this study [2].

Very little pre-processing or error checking needs to be carried out in the GEMS pitch extraction algorithm. First, the GEMS signal is bandpass filtered with an analog filter with 3-dB frequencies of 70 Hz-7 kHz, which produces the clean signal shown in Figure 2. Then, any linear trend is removed before the zero-crossing search is carried out. The algorithm uses 30 millisecond search windows with no overlap. An energy calculation is done to determine if the speech is voiced or unvoiced. If voiced, the first three zero crossings are calculated and the average pitch for two glottal cycles is determined. The next window begins after the second glottal cycle and the process is repeated. Any anomalous pitch values outside the typical pitch range of 50 Hz-400 Hz are zeroed out. The GEMS pitch algorithm also has the inherent benefit of yielding pitch-synchronous information. The pitch is found via the zero crossings, which are natural pitch cycle boundaries. The crossing locations can be used to do pitch synchronous processing, which increases the accuracy of FFTs. The fairly linear shape of the signal near the zero crossings also is conducive to linear interpolation for a further increase in accuracy for the pitch values. The algorithm also has the unique ability to specify how many glottal cycles are averaged to make a pitch estimate. In this work two cycles are used per pitch estimate, but any number of integer glottal cycles can be used. Two glottal cycles were found to be optimal in pitch estimation because it is long enough to get a smooth pitch contour, and yet short enough to capture natural pitch fluctuations.

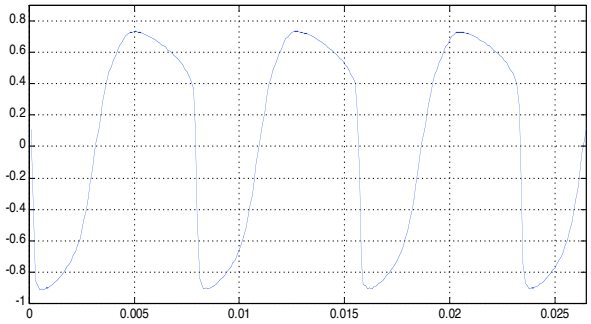

Figure 2. A typical GEMS waveform from a male speaker

GEMS Shape Parameter (GSP)

The shape of the GEMS output is very distinct for each individual. Figure 3 shows an example of a GEMS signal for a portion of speech for four different speakers. Each speaker's waveform is unique. Although they share certain qualities like the general shape, subtle differences are seen in the slope of the rise and fall of the waveform along with other variations in shape. Unlike the other parameters discussed in this chapter like spectral coefficient and pitch, the GEMS signal shape is not time varying.

The motions of the glottis and tracheal walls are not different for the different phonemes (Burnett 1999), although they can vary for different registers of voice. All of our speakers used their normal speaking voice, referred to as the modal or chest register, so this was not a factor in these experiments. The GEMS signal is related to the excitation pressure of the system, which is filtered by the vocal tract to produce the different phonemes. Instead of a parameter that varies in time with the speech, the GEMS shape is relatively constant throughout speech. The shape of the GEMS signal changes only briefly during the beginning and ending of phonation, but this was recognized and only samples from the middle of phonation were processed. This unique quality of the GEMS signal presents a new opportunity in extracting a parameter for verification use.

A new parameter was needed to characterize the shape of the waveform from the GEMS device and compare it to different speakers. Although many different characterizations were examined, such as wavelets, polynomial coefficients and the K-L expansion coefficients, a simple method using the GEMS signal shape directly worked best. The GEMS shape parameter (GSP) is a based on averaged two-glottal cycle waveforms from each sentence data file. Many two-glottal cycle waveforms from data file are averaged together to produce one two-cycle waveform - this waveform is the GSP. Many cycles were averaged as to smooth out any anomalous cycles. Since the GEMS signal is not stable at the onset and offset of speech, the algorithm did not sample any waveforms near the beginning or end of phonation, normally 6-10 windows from the boundaries by using the voiced/unvoiced boundary information from the GEMS pitch algorithm.

The GSP algorithm also separated the waveforms used in the average, so as not to use consecutive two-glottal cycle windows. This would eliminate any heavy use of anomalous waveforms in the GEMS signal due to speaker or device motion. As with the gain parameter, shape (and not amplitude) is the important information, so care had to be used when choosing a distance calculation for the GSP in the DTW algorithm. The correlation coefficient distance and standard DTW distance (Euclidean) were tested on the normalized GSP waveforms and the DTW distance was found to have consistently lower error rates, about 2% lower on average.

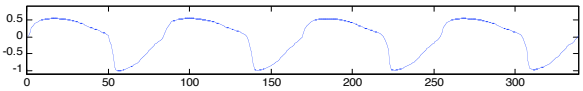

Figure 3. Example of GSP from 4 different speakers

Auto Regressive Moving Average (ARMA) Model

Spectral coefficients, like LPC and cepstral, estimate the transfer function using only the acoustic signal (output of the LTI system). These methods, while fast, are inherently inaccurate due to a lack of information about the input to the system (the excitation function). They make simple assumptions about the input to the LTI system; most commonly assuming the input to the system is white (spectrally flat). The GEMS signal yields information about the excitation function while the acoustic signal is the output signal. Together they can be used in an input-output or pole-zero model. These are often referred to as Auto Regressive Moving Average or ARMA models. The ARMA name comes from the models origin in statistics. The AR (Auto Regressive) part is information from the output; LPC is a very popular AR model. AR models are also called all-pole models, because it uses only poles to model the system. The MA (Moving Average) part is information about the input function and is conversely called an all-zero model because the denominator is a constant. An ARMA model uses poles and zeros, both input and output signals to model the system. The ability to measure both input and output signals gives access to the class of more accurate ARMA models.

Cepstral Coefficients

The cepstrum is a way to approximate the transfer function of the voiced speech system. By truncating the high frequency content of the cepstrum, we can retain the information from the transfer function, $H(\omega)$, and get rid of the information from the excitation function, $X(\omega)$. By keeping only the first 10 to 20 components of the real cepstrum, called the cepstral coefficients, an estimate of the transfer function with fewer coefficients is obtained. With the truncated cepstrum, the inverse transforms of can be used to obtain an approximation of the transfer function $H(\omega)$. The cepstral coefficients are used extensively not only in speaker verification system, but also for other applications including speech recognition [5].

4. EXPERIMENTAL RESULTS

The twelve sentences chosen for our database were all common speech application sentences, appearing in many speech corpuses, like the TIMIT corpus, which was developed for this experiment [7]. It was noticed during the experiment that the longer, more complex sentences produced lower error rates. This is not too surprising as they contain more information. After the decision was made to reduce the total number of sentences from 12 to 3 so that the processing could be completed in a reasonable amount of time, the logical choice was to choose the ones with the lowest EER. However, it is not always easy to determine which ones would have the lowest EER without testing them all, defeating the purpose of choosing only 3. A method was sought to give an indication of EER performance based on sentence length.

Calling it "sentence length" is a little misleading. It is not the length in time it takes to speak the sentence, that would be speaker specific due to different speech rates. Or is it how many words or syllables per sentence? A new term called sentence complexity was chosen to describe the parameter that was sought. The sentence complexity is a number, the larger the number, the more complex or more speech content in the sentence. The complexity number (C) is simply the number of syllables (S) added to the number of words (W), C = S+W. This method emphasizes sentences that have many small words and sentences that have few, many-syllable words. Both of these sentence types contain more speech information than small, short, one-syllable sentences.

A selected EER versus sentence complexity plot is shown below (Figure 4) for CC1(the first cepstral coefficient). Some sentences had the same complexity number, so some complexity values have multiple data points. There is a linear trend relationship between EER and sentence complexity, seen most clearly in the results for CC_1 in Figure 4. In general, as the sentences get more complex, the EER gets lower. Therefore, choosing the most complex sentences will yield the best EER. The explicit linear relationship between EER and sentence complexity is shown in the figures below. One can write in general

$$EER = \alpha C + \beta$$

where α and β are the slope and bias coefficients.

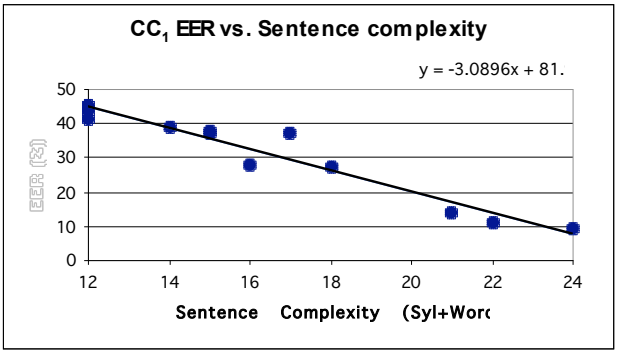

Figure 4. Equal error rate versus sentence complexity

Overall Verification EER Results

The results above demonstrate the ability of the individual verification parameters to distinguish and verify the identity of speakers. However, a working verification system would not rely on one parameter alone to make the accept/reject decision. The Bayes classifier is used to combine the statistics of the individual parameters and calculate a combined, single EER. As described earlier, the classifier maps the normalized DTW distances from all the participating parameters into one vector. This mapping allows for a calculation of the overall EER, which is lower than the EER

from any of the individual parameters. Parameters are chosen so that they compliment one another by being as statistically independent as possible.

Figure 5 shows the results of the EER using the Bayes classifier. The EER points on the far left use two verification parameters and an additional parameter is added as the curve moves to the right. The three lines on each graph represent different possible verification systems. The first two have been discussed thus far: the traditional (blue) and the GEMS enhanced (red) systems. Their performance is similar with the noiseless data. This is not completely surprising given that they both contain very similar information: gain, pitch, spectral coefficients and delta-spectral coefficients. Both sets of feature vectors have similar information content and both sets of data were recorded in a controlled laboratory setting. The third (green) line was constructed to show what the additional, purely GEMS based, GEMS shape parameter (GSP) can provide. An ultimate EER of 0.01% is obtained using the GEMS enhanced system with the GSP. This is a factor of seven lower than the traditional system. The motivation behind the third line is as follows. Every point in the first two curves (red and blue) adds an analogous parameter to the classifier. For example, the second point adds the CC_1 parameter in the traditional system and the analogous As_2 is added to the GEMS enhanced system. However, there is no acoustic based analog to the GSP. This additional pure GEMS based feature vector provides insight into how an optimized GEMS verification system would perform, even without the presences of noise.

As seen in the figures below, the three lines are now well separated due to the addition of noise to the acoustic data, especially when the system includes more than two verification parameters. They differ by a factor of 1.7 with the added white noise and by over a factor of 3 with the color noise. The bottom line (green), which is the GEMS system with the GSP parameter, illustrates again how well the system can perform, even in the presence of noise. The third GSP augmented system shows almost a factor of 6 improvement over the traditional system with white noise and over a factor of 9 improvement with the color noise.

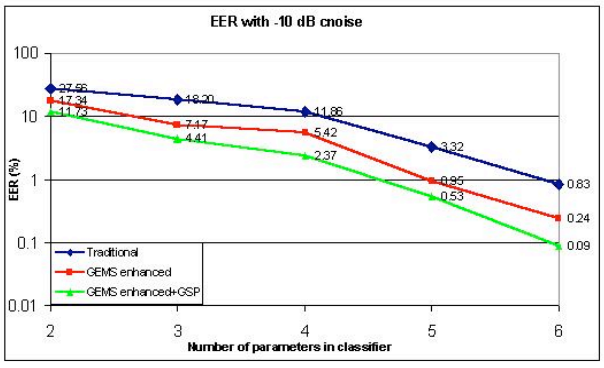

Figure 5. EER performance as a function of number of feature parameters.

5. CONCLUSION

Low power EM radar-like sensors can measure the internal properties of the human glottal regions safely and non-invasively. These data, together with the user's speech signal and reliable sampling of the acoustic noise signals, enable novel applicaton to a speaker verification system. Results of the experiments shown that error rate can be reduced by almost a factor of 10. This is due largely to the non-acoustic, independent nature of the GEMS measurements. It is anticipated that measurements of other EM sensor physiological conditions of the speech articulator, such as tongue and lips motions, will further reduce the verification error rate.

ACKNOWLEDGMENTS

We thank the U.S. Department of Energy and the National Science Foundation for their support. This work was sponsored by the U.S. Government and performed by the University of California Lawrence Livermore National Laboratory under Contract W-7405-ENG-48 with the Department of Energy.

REFERENCES

[1] Holzrichter, J. F.; Burnett, G. C.; Ng, L. C.; and Lea, W. A., *Speech Articulator Measurements using Low Power EM-wave Sensor,* J. Acoust Soc. Am. 103 (1) 622,1998. Also see the Web site http://speech.llnl.gov/ for related information.

[2] Burnett, G. C., *The Physiological Basis of Glottal Electromagnetic Micropower Sensors (GEMS) and Their Use in Defining an Excitation Function for the Human Vocal Tract,* Thesis UC Davis, Jan. 15th, 1999, document #9925723 available from University Microfilms, Inc., Ann Arbor, Michigan; also see {1]

[3] Gable, T.J., *Speaker Verification Using Acoustic and Glottal Electromagnetic Micropower Sensor (GEMS) Data*, UC Davis PhD dissertation, February, 2001.

[4] Herrnstein, A., Holzrichter, J. F., Burnett, G. C., Gable, T. J., and Ng, L.C., *Statistics of Unvoiced Time Period Duration Relative to EM Sensor Detected Voiced Onset and End Times,* unpublished. Statistics are based upon a corpus of 15 male speakers pronouncing excerpts from a TIMIT phoneme, numeral, and sentence data set is contained on 8 CDs, available as UCRL MI-132776, Lawrence Livermore National Laboratory.

[5] Rabiner, L. and Juang, B. W., *Fundamental of Speech Recognition,* Prentice-Hall, 1993.

[6] Rosenberg, A.E., "Automatic Speaker Verification: A Review," Proceedings of the IEEE, Vol. 64, No.4, April 1976.

[7] "Glottal Electromagnetic Micropower Sensor & Acoustic Data," LLNL-UCDavis speech data base containing seven CDs of 14 male speakers , UCRL-MI-132776, February 1, 1999. Also see Web site mentioned in [1].