Micropower Electro-Magnetic Sensors for Speech Characterization, Recognition, Verification, and other applications

*Holzrichter, J.F., Burnett, G.C., Gable, T. J., Ng L.C.

Lawrence Livermore National Laboratory and University of California at Davis
Lawrence Livermore National Laboratory
P.O. Box L-3
Livermore, CA 94550
*email: holzrichter1@llnl.gov

ABSTRACT

Experiments have been conducted using a variety of very low power EM sensors that measure articulator motions occurring in two frequency bands, 1 Hz to 20Hz, and 70 Hz to 7 kHz. They enable noise free estimates of a voiced excitation function, accurate pitch measurements, generalized transfer function descriptions, and detection of vocal articulator motions.

1. INTRODUCTION

Experiments have been conducted using a variety of very low power EM sensors (< 0.1 milliwatt radiated power) to measure speech articular motions in real time (Holzrichter et al. 1998). These EM power levels are well below international safety standards for continuous public use (Polk 1996). Measurements of glottal tissue motions, associated with glottal opening and closing, were conducted using a 2 GHz homodyne field disturbance sensor attached to a monopole antenna. This system measures tissue motions which are located within a fixed range (e.g., within a 5 cm "bubble"), and which occurring in a frequency band from 70 Hz to 7 kHz.

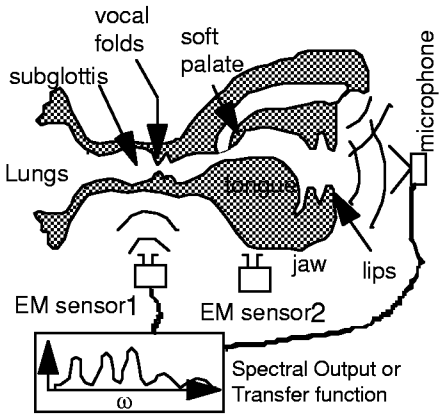


Figure 1: Illustrative vocal system and EM sensor measurement positions. Also, shown is the transfer function instrumentation schematic.

A second survey sensor has been used to measure jaw, tongue, and palate motions, using an impulse sensor coupled to a cavity backed monopole antenna. It can measure interface motion rates between 1 Hz and 20 Hz, and within a 10-cm distance. A third type of sensor, a 5 GHz impulse transmitter, with a directional horn antenna, has been used to measure glottal motions at distances up to a meter.

2. VOICED EXCITATION

EM sensor 1 measures tissue interface motions in the subglottal and vocal fold region. It is called GEMS for glottal electromagnetic sensor. Experiments by Burnett et al. (1997) show that the air tissue interfaces, measured by the GEMS, are moving synchronously with the opening and closing of the vocal folds.

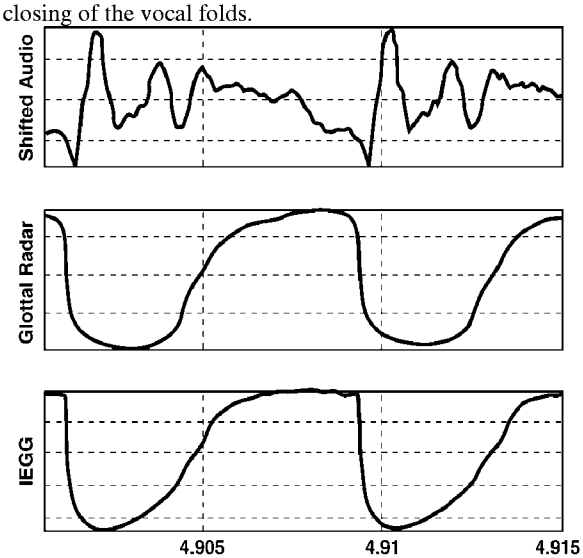


Figure 2: Comparison of acoustic signal, glottal EM sensor (i.e., radar), and EGG signal. The glottal signal is used to

estimate a voiced excitation, and the rapid fall as the folds close, is used for glottal pitch period timing.

The spectral content of the EM glottal signal contains information describing both the subglottal air pressure (and thus related airflow information) and high frequencies associated with rapid vocal fold closure. Its frequency response is from 70 Hz to 7 kHz, and is AC coupled to the A/D converter. The EM sensor data is first corrected for sensor internal filter characteristics to yield signals like those shown in trace 2 in Fig. 2. These signals are then used as an approximate voiced excitation function to estimate vocal tract transfer functions.

The EM sensors provide robust glottal structure motion detection under conditions of both complete closure (i.e., modal speech as in Fig. 2) and incomplete closure (i.e., breathy and falsetto speech) where EGG sensor signals become very small. In addition, the time domain characteristics of the GEMS signals enable very accurate detection of voicing onset, of voiced/unvoiced speech boundaries, and of the glottal pitch period.

It is worth noting that these EM sensor measurements are unaffected by acoustic noise, and thus can be used as an "oracle" in denoising algorithms. For example, they indicate amplitude and phase continuity of voiced speech, they indicate periods of un-voiced speech when sufficient acoustic signal is available to identify periods of acoustics with no voicing, and they indicate periods of no voiced speech. For onset- and end-of-speech detection simple algorithms use the near simultaneous EM and acoustic information to test for unvoiced speech time frames (when possible), they can use language statistics to allot fixed time windows for unvoiced speech units, and should be able to use EM sensed vocal fold retraction and pharynx dimensional changes to resolve articulator changes associated with unvoiced speech units.

2.1. Pitch Measurement

The EM sensor measurements of glottal opening and closing provide very accurate pitch period measurements (accuracy < 1 Hz). The algorithm to process the GEMS data uses positive to negative zero crossing to obtain a consistent time of glottal closure. Commonly two pitch periods are averaged to obtain time domain data that is accurate to less than 1 Hz for each pitch period (i.e., 0.1 ms accuracy in each10 ms frame). The EM signal method shows a >10 fold increase in pitch accuracy and a 100 fold reduction in processing time compared to two all-acoustic techniques (see Burnett 1998) that were chosen as references.

In Figure 3, this method of pitch measurement is compared to traditional pitch measurements using established cepstral and autocorrelation techniques (Rabiner 1978). To illustrate the noise immunity properties of the GEMS approach to pitch measurement a two-speaker experiment was conducted. A second speaker, located about twice as far as the primary speaker from the primary speaker's microphone, began speaking at a time 1.2 seconds after the primary speaker started (and stopped 3 seconds after start). The GEMS data in trace two is unaffected by the noise (second speakers signal), while the 3rd (Cepstral) and 4th trace (autocorrelation) show serious deterioration of the pitch accuracy.

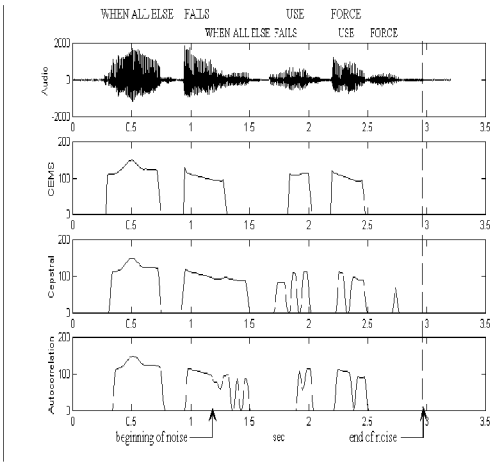


Figure 3: The pitch from speaker 1 is measured three ways while speaking the phrase "when all else fails use force." The acoustic signal is in trace 1. Using an EM sensor (called GEMS on trace 2) and using two all acoustic methods: Cepstral, (trace 3) and Auto-correlation (trace 4). Upon noise input from a 2nd speaking voice at 1.2 seconds (see vertical dashed lines), the Cepstral and the Auto correlation methods fail, the EM sensor signal is unaffected.

2.2. Pitch Synchronous Processing

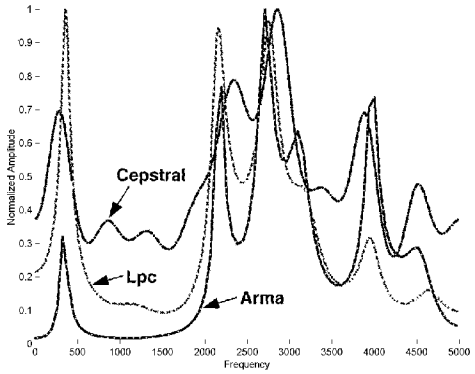
The capacity to measure instant glottal time frame boundaries and the corresponding voiced excitation and acoustic signal, enables the users of these methods to do pitch synchronous processing of the speech signal. There is a hierarchy of methods that can be employed using the additional EM sensor information provided. It can be used to define the time frames over which the acoustic signal is to be characterized by standard spectral methods such as cepstral or LPC. As shown below, the excitation function can be removed from the acoustic signal to estimate a transfer function which can provide good spectral information (such as is possible using ARMA techniques), it enables the removal of excitation function variations (a form of pitch normalization), and the data can be stored for subsequent speech synthesis.

3. TRANSFER FUNCTIONS

The excitation information can be removed from the acoustic signal by using a variety of deconvolving techniques (e.g., see Fig. 1), yielding well-defined transfer functions, formant locations, which are pitch normalized. Having the excitation enables the use of pole/zero approximation procedures, such as the autoregressive moving-average (ARMA) approach. Figure 4 below shows three different approximation methods for characterizing a segment of acoustic speech. The ARMA approach, using 16 poles and 12 zeros, provides very well defined lower and higher formants, that capture the individual qualities of the individual speaker. They provide quite useful filter parameters for subsequent convolution with an excitation function for speech synthesis, without the need for "residual" information. In the future, it is anticipated that model based characterization, using simplified areal functions, can be employed based upon these data.

3.1. Characterization of Speech

By using the procedures described above, naturally spoken speech can be characterized automatically. First the algorithm performs automated vocalization detection, then it defines the time frames for processing (2 glottal cycles at a time), and then it removes the excitation information via the ARMA transfer function process. The data are plotted below



for each 2-cycle frame of voiced speech.

Figure 4: Spectral and transfer function examples of the phoneme /i/, from an adult male speaker, using 20 coefficient LPC, 20 coefficient Cepstral, and 16 pole/12 zero ARMA using a 2 glottal cycle (16 ms) window. The cepstral represent the higher formants well, LPC the lower formants well, and ARMA 16 pole/12 zero provides uniform formant descriptions.

On the following page, figure 6 illustrates the automated processing methods (described above) based upon pitch synchronous processing and pole/zero characterization. The two phrases "wreck a nice beach" and "recognize speech" are commonly used to test speech recognition systems. The speech segments describing the underlined speech units in "wreck a nice" and "recognize speech" have very different excitations and transfer funtions.

4. EM SENSED ARTICULATORS

EM sensors can be used to characterize generalized articulator motions as shown below in Figure 5. The EM sensor 2 was positioned as shown in figure 1, but placed against the skin under the jaw. As a result it measured the relative positions of internal articulators relative to the jaw surface. The EM sensing of several articulators simultaneously, in concert with acoustic speech, can provide a great deal of information on the completeness of the articulation, the degree of co-articulation, and other information for use in special applications.

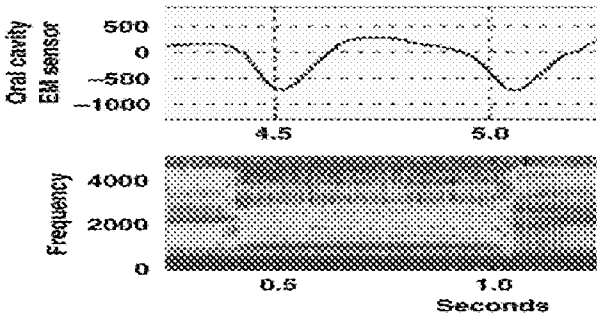


Figure 5: EM sensor 2, placed under the jaw, measures tongue and soft palate motions as the following sequence, /ng/, /a/, /ng/, /a/ was articulated. The corresponding spectrogram is shown.

5. SUMMARY

The information provide by low power, low cost EM sensors can be used to enhance the characterization of speech in many situations. The additional data appears to be especially valuable for those applications operating in noisy conditions, for those that can take advantage of the high quality transfer functions, and that can use generalized articulator information.

5.1. Acknowledgements

This work was performed under the auspices of the U.S. Department of Energy by the Lawrence Livermore National Laboratory under Contract W-7405-ENG-48 and the National Science Foundation by a SGER grant.

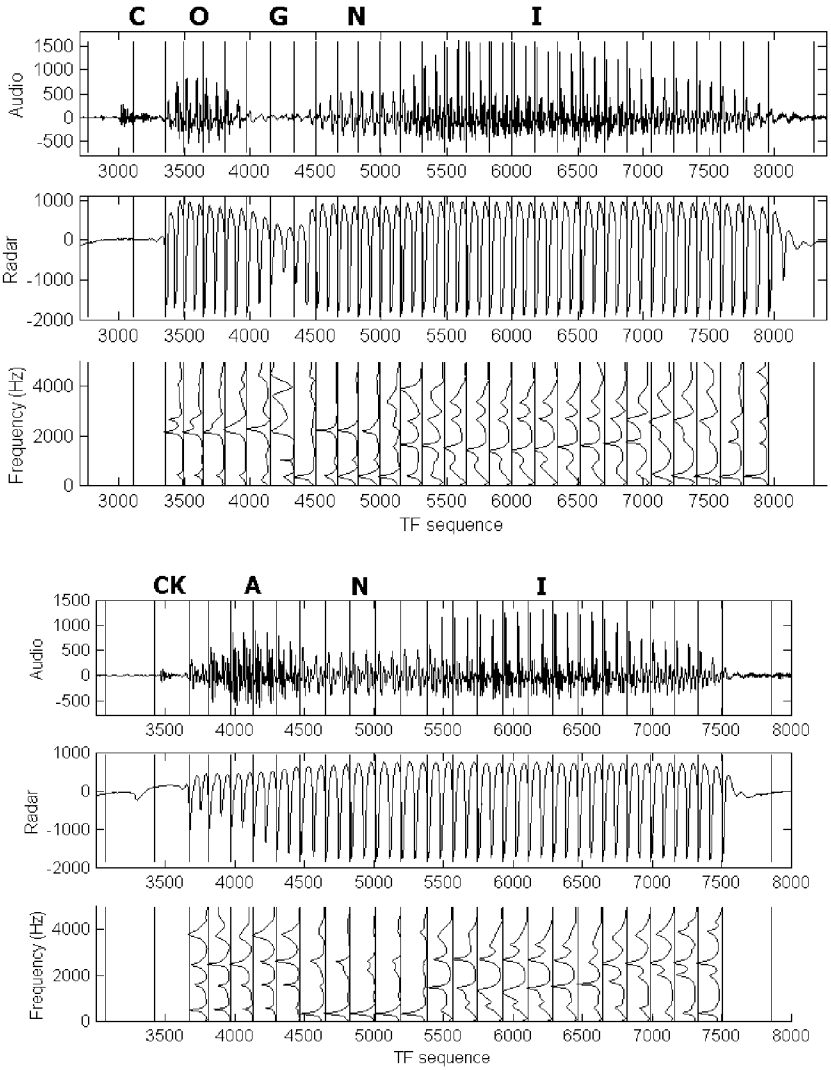


Figure 6: Pitch synchronous transfer functions and glottal excitation for short segments (underlined) of the phrase "wreck a nice beach" and "recognize speech." These are obtained automatically using pitch synchronous processing and ARMA procedures. Note differences in radar (GEMS) sensor signal at the location of the "g" for the two phrases. The transfer functions provide information on articulator motions, indirectly through area functions, and at the same time the glottal EM sensor is providing data on voicing/unvoicing, excitation energy versus acoustic energy, pitch, time duration, and many other properties.

6. REFERENCES

Burnett, G.C., Gable,T. G., Holzrichter,J.F., Ng, L.C., (**1997**) "Voiced excitation functions calculated from micropower impulse radar information." J.Acoust. Soc. Am. Vol. 102 (5) Pt. 2 , 3168 (Nov 97)

Burnett, G.C., (**1998**) "Accurate and Noise Robust Pitch Extraction using Low Power Electromagnetic Sensors" to be published

Gable T. G., Burnett, G.C., Holzrichter,J.F., Ng, L.C., Lea, W.A. (**1997**) "Comparison of conventional acoustic and MIR radar/acoustic processing of speech signals" J.Acoust. Soc. Am. Vol. 102 (5) Pt. 2 , 3168 (Nov 97)

Holzrichter J.F., Burnett G.C., Ng L.C., and Lea W.A., **1998** "Speech articulator measurements using low power EM-wave sensors," JASA 103 (1) 622, Jan., also see information on web at http://speech.llnl.gov/

Rabiner, L.R. & Schafer **1978** R.W. "Digital Processing of Speech Signals," Prentice Hall

Polk C.,& Postow, E., 1996 "Biological Effects of Electromagnetic Fields" 2nd ed, CRC Press