USE OF LOW POWER EM RADAR SENSORS FOR SPEECH ARTICULATOR MEASUREMENTS

J.F. Holzrichter and G.C. Burnett
Lawrence Livermore National Laboratory, L-3
P.O. Box 808 Livermore, California 94550
email: holzrichter1@llnl.gov

ABSTRACT

Very low power electromagnetic (EM) wave sensors are being used to measure speech articulator motions such as the vocal fold oscillations, jaw, tongue, and the soft palate. Data on vocal fold motions, that correlate well with established laboratory techniques, as well as data on the jaw, tongue and soft palate are shown The vocal fold measurements together with a volume air flow model are being used to perform pitch synchronous estimates of the voiced transfer functions using ARMA techniques.

INTRODUCTION

Electromagnetic (EM) radar sensors (e.g. see designs by McEwan (1)) provide a capability for measuring EM wave reflections from speech organ interfaces in a non invasive, safe, fast, portable, and low cost fashion (2). See Fig. 1 below and Fig. 2 in the next column for positioning of these sensors.

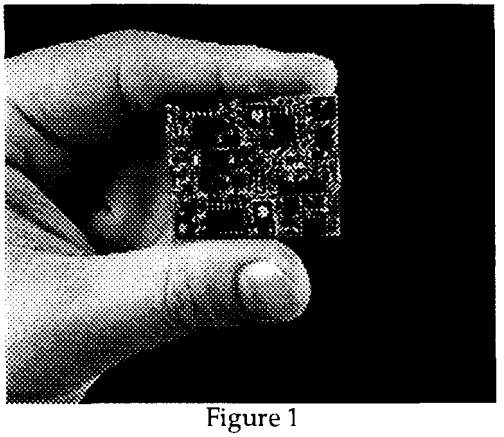

Figure 1

The sensor transmit mode generates a short pulse train of four to six cycles, lasting a few nanoseconds, that radiate from monopole antennas. A range gate is delayed, relative to the start of the transmitted pulse, by a fixed time of a few nsec. It "drives" a diode sample gate that detects the reflected EM waves received by a nearby receiver antenna. The sampler gate detects reflected energy from all wave-tissue reflections that meet the round-trip time delay conditions (i.e., about 4 cm into the neck for the leading edge of the pulse).

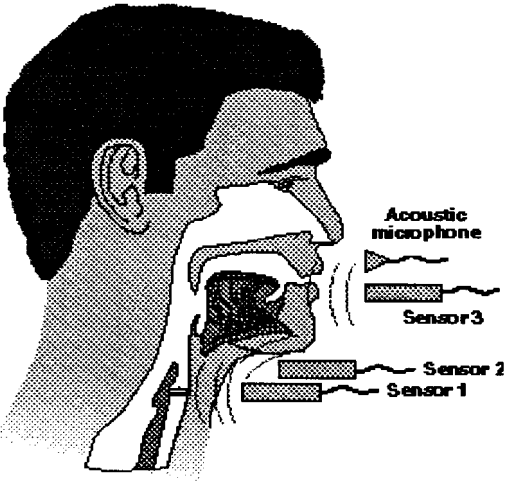

Figure 2

The transmitter generates 2×10^6 pulses per second, which means that in each 0.5 msec period, 1000 pulses are transmitted, received, and integrated into a high S/N analog signal. This signal is electronically filtered to remove reflected signals from non moving or slowly moving interfaces.

For glottal measurements in particular, the low frequency "clutter" from the skin/air interface and similar interfaces are removed using low frequency cut-off filtering below 80 Hz. For the jaw and tongue sensor approximately 0.5 Hz filtering is used.

DATA

Figure 3 shows typical data from the vocal folds, the simultaneously measured speech signal, and an EGG signal. These data indicate that the EM sensor is measuring tissue motions consistent with other instruments (3,4). In addition, high speed camera pictures taken through a laryngoscope (5) also confirm that the EM sensor is measuring signals associated with the glottal opening and closing.

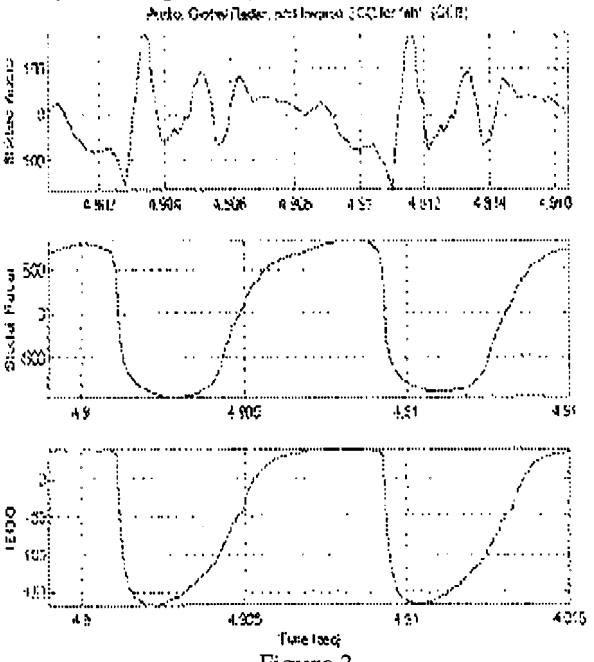

Figure 3

As discussed below, such glottal sensor signals are used to develop a volume air flow excitation function for real time estimation of the transfer functions for each speech time frame. In addition, the repetitive structure of the signals shown in the center tract of Figure 3 enable measureing accurate pitch period values.

Examples of EM sensor data, simultaneously recorded with acoustic data during the articulation of the word "print," are shown in Fig. 4 in the next column.

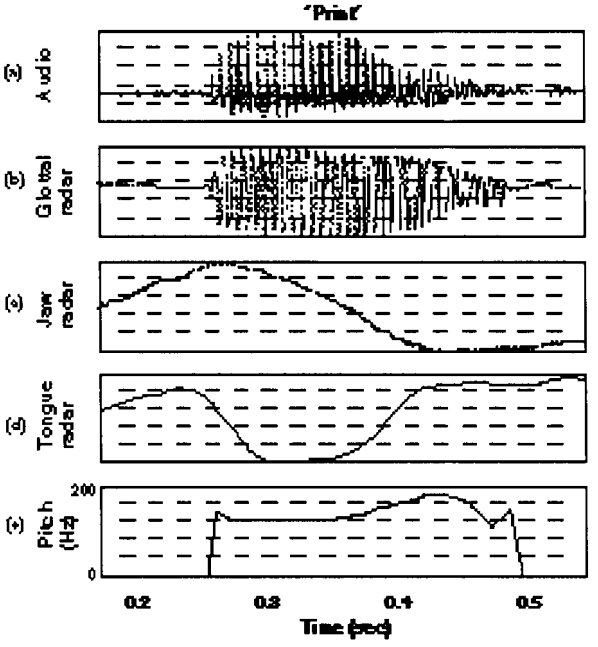

Figure 4

The sequence of windows in Fig. 4 show the acoustic signal 4a, a simultaneously recorded EM glottal sensor signal 4b, and a separate EM jaw sensor signal showing up/down jaw movements 4c. In a sequential experiment, EM sensor data from the oral cavity showing whole tongue motion was recorded for the same word by the same speaker and is shown in Fig. 4d. Fig. 4e also shows an example of simple information processing by displaying the instantaneous pitch period versus time.

The data in Fig. 4c,d are from organs that move more slowly than the vocal folds, and thus their motions take place over periods of 0.1 to 1 seconds. Time filtering of such slow signals is more difficult, especially for environments where the head may be moving relative to the sensors. However in the presently used survey mode, all organs that move in the time window defined by

the filtering are measured. For research topics on specifically chosen single-organ phonemes, accurate data is obtained. The multiple articulator information shown in Fig. 4d where, at the end of the word "print" the tongue body and tip have been raised for the /n/, the detection of tongue tip motion for the /t/ is not yet clear. For accurate measurements of such multiple articulator data, it is desirable to use ranged gated sensors (6) which can measure air/tissue interfaces determined by the timing of the range gate.

TRANSFER FUNCTIONS

Using simple models of the volume air flow excitation function (2), several of the voiced transfer functions have been estimated by Fourier transforming both the acoustic speech and approximated excitation signal (as estimated above), then deconvolving. Examples for 4 vowels are shown in Figure 5 below.

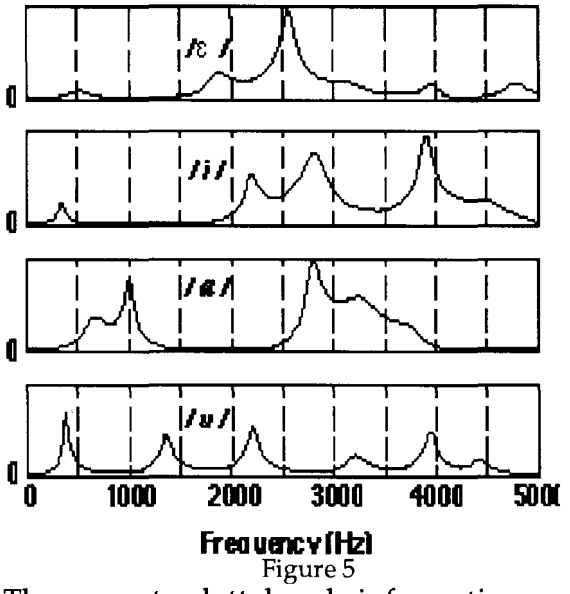

Figure 5

The accurate glottal cycle information illustrated in Figure 3 enables automatic algorithms to Fourier transform the excitation function and acoustic signal "pitch synchronously," over two glottal periods. Because a real time excitation function can be estimated, both poles and zeros can be used to approximate the transfer functions. An autoregressive, moving average (i.e., ARMA) approximation procedure was used, with 15 poles and 15 zeros, to fit a smooth analytic function to the numerical data for four American English vowels, /ɛ/, /i/, /a/ and /u/, as spoken by a young adult male. The data is very clear and they appear to approximate formants that match well the locations shown in previous publications, but with enhanced low intensity structure as expected by using "zeros," in addition to "poles."

APPLICATIONS

The signals from the glottal motions enable the user to determine a great deal of information about the types of speech being spoken in real time and the EM sensors and algorithms are very economical. Examples of the information include the types of excitation function used for the speech unit (e.g., voiced or unvoiced), the onset of speech, the presence of noise, pitch synchronous frame definition, as well as estimates of the volume air flow excitation estimation and consequent transfer function for each one or several glottal cycles.

The signals for the jaw and tongue positions versus time shown above in Fig. 4 (plot c and d) were measured to submillimeter relative accuracies, showing the upward motion of the organs well before the plosive /p/, where upon the jaw drops continuously but the tongue drops for /i/ and lifts for /n/. Such positional articulator information may be used in speech recognition by testing measured sequences of motions against those for a trial sequence of phonemes (e.g., jaw motions for /silence/-/p/-/r/, and tongue for /r/-/i/ -/n/), then generating a

"closeness score" using a recognition algorithm.

Preliminary experiments indicate also that very personalized speech synthesis can be effected using the procedures described above and in the referenced documents. In addition to the accurately defined pairs of excitation functions and transfer functions, relatively rapid library formation of individualized sets of basic phoneme, diphone, and other speech units is expected using these procedures. The procedures for synthesis, are the reverse of those described under the above heading "Transfer Functions." The measured pair of excitation and transfer functions for each voiced speech units, when convolved together, matches the original speech signals well. For unvoiced speech units, which are relatively small in number, established procedures can be used for their synthesis. Preliminary experiments have validated this approach.

CONCLUSION

When these techniques are more completely explored, with improved sensor calibrations and validated on statistically significant groups of speakers, these new sensor technologies should add value to many speech research, diagnostic, and technology applications. Similarly, EM sensors can be optimized for research, for medical diagnostics, or miniaturized for technology applications such as insertion into microphone housings, telephone headsets, and other devices. The noninvasive aspects appear useful for routinely diagnosing articulator problems and providing feedback to speakers, singers, language learners, and the speech disabled. Other promising applications are to use the estimated excitation functions and calculated transfer functions for speech coding as used in telephony transmission, for storage and subsequent synthesis as "personalized" speech, and for speech recognition and speaker verification.

ACKNOWLEDGMENTS

We thank L.C. Ng for signal processing advice, and T. E. McEwan for supplying several low power EM radar sensors. We are grateful to Dr. Wayne Lea for helpful discussions, and Prof. I. Titze and Dr. B. Story at the National Center for Voice and Speech, University of Iowa for discussions on the glottal structures and for preliminary EGG measurements. Work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.

REFERENCES

1. McEwan, T. E. , US Patents 5,345,471 (1994), 5,361,070 (1994), and 5,573,012 (1996).

2. Holzrichter, J. F., Lea, W. A., McEwan, T. E., Ng, L. C., Burnett, G. C. "Speech Coding, Recognition, and Synthesis using Radar and Acoustic Sensors," University of California Report UCRL-ID-123687, (1996).

3. Rothenberg, M. "Some relations between glottal flow and vocal fold contact area," ASHA Rep. 11, 88-96 (1981) .

4. Titze, I. R. "Parametrization of the glottal area, glottal flow, and vocal fold area," J. Acoust. Soc. Am. 74(2), 570-580 (1984)

5. Leonard, R.J. et al (1997). "A comparison of laryngoscopic and EM glottal sensor data," to be published.

6. Skolnik, M. . *"Radar Handbook,"* 2nd edition, McGraw-Hill, New York, (1990).