LOW BANDWIDTH VOCODING USING EM SENSOR AND ACOUSTIC SIGNAL PROCESSING

Lawrence C. Ng, John F. Holzrichter, and Peder Larson

Lawrence Livermore National Laboratory, Livermore, California, U.S.A.
larry.ng@llnl.gov, holzrichter1@llnl.gov, larson40@llnl.gov

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 [1]. By combining these data with the corresponding acoustic signal, we've demonstrated an almost 10-fold bandwidth reduction in speech compression, compared to a standard 2.4 kbps LPC10 protocol used in the STU-III (Secure Terminal Unit, third generation) telephone. This paper describes a potential EM sensor/acoustic based vocoder implementation.

1. INTRODUCTION

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 [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 glottal air pressure pulsations [1,2]. In particular, 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, during which no coding is needed, or when sampling and removal of background noise signals can reliably take place [4,5].

In this paper, however, we present results from our recent study on the application of a specific EM sensor called GEMS, for low bandwidth vocoding or GEMS based coding (GBC). We first established equivalence performance comparison to the LPC10 standard, and then repeatedly decreased the speech transmission bandwidth while maintaining intelligibility of the speech quality. Our experiments demonstrated that speech quality and intelligibility can be maintained even at a rate as low as 300 bps.

2. APPLICATIONS OF GEMS TO A LOW BANDWIDTH VOCODER

This section briefly outlines the GBC coding method. The EM sensor signal is used in three critical areas: (1) Speech detection and typing, which separates the speech signal into voiced, unvoiced, and silence segments. (2) For voiced speech, the GEMS signal provides information to construct an input excitation to compute the short term transfer function using the autoregressive and moving average (ARMA) model. The resulting ARMA model can be converted into a small number of appropriate poles and zeros.

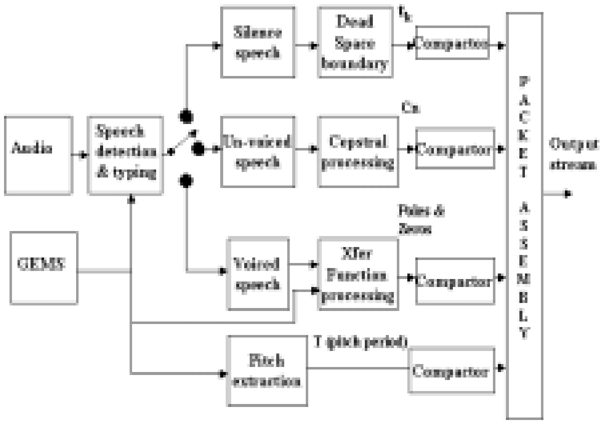

Figure 1. GBC based coding process.

(3) To determine pitch, one or more glottal cycles serve as a basic signal processing window or frame. For unvoiced speech segments, a cepstral description is used. For silence, the dead space boundaries are detected. The coefficients generated from each frame are accumulated and then processed by a compactor. The compactor will be discussed later. The output of the compactor units are then assembled into a bandwidth compressed output data packet stream along with a header and possibly a trailer for error detection. Note that when the communication link is first established, repetitive data such as pitch and the glottal waveform, characteristic of each speaker, can be transmitted in a header. At the receiver, the GBC

decoder expands the silence, voiced and unvoiced segments, into normal speech.

2.1. GEMS Pitch Extraction

The use of the GEMS signal enables great speed and accuracy in pitch estimation, pitch period determination, and an excitation function. 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 time-domain, interpolated, zero-crossing algorithm. The algorithm searches for the positive to negative crossing of the signal [2], thereby rapidly and accurately obtaining the pitch period.

The algorithm also has the unique ability to adaptively specify the number of glottal cycles over which an average pitch is estimated. 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. In addition, transfer functions remain constant over this two-cycle period.

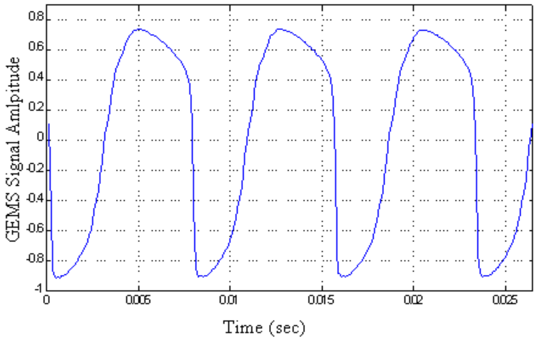

Figure 2. Typical GEMS waveform from a male speaker.

2.2. Pitch Scaling of Excitation Function

In Code Excited Linear Predictive (CELP) coding [6], a set of Gaussian codes from a code book is used to extract the vocal tract model from the speech signal. A code that minimizes the residue is selected as the correct excitation function and the corresponding code index is then transmitted. For our application here, the GEMS signals are used to provide the excitation function. This excitation function can be very efficiently transmitted as follows. Since the pitch (i.e., duration) varies with time but the shape of GEMS for an individual does not, [5] the knowledge of pitch enables the algorithm to stretch or shrink the excitation signal to match that used by the speaker. Therefore, the excitation function shape needs only to be transmitted once. For example, Fig. 3 shows how the GEMS signals from four different speakers can be scaled to an identical pitch period while maintaining the individual shape while preserving the individual personality.

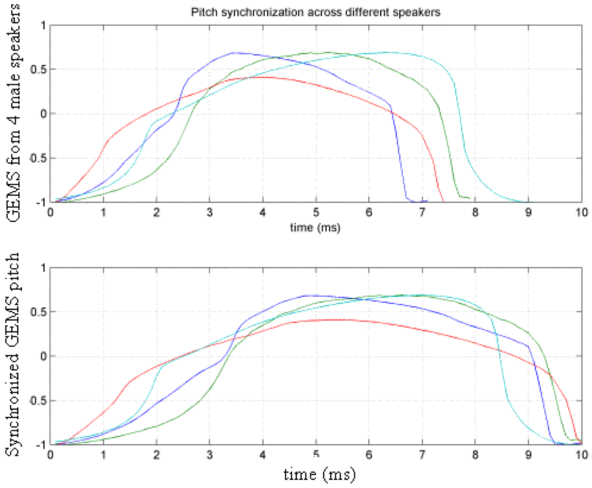

Figure 3. Scaling of different GEMS signals (top) to an identical pitch period (bottom).

2.3. GEMS Based Speech Type Detection and Timing

Figure 4 shows an example of how a typical speech utterance “Recognize Speech” can be detected and typed into silence, voiced and unvoiced segments. For a given utterance, this automated process can be used to develop statistics on voiced, unvoiced, and silence segments.

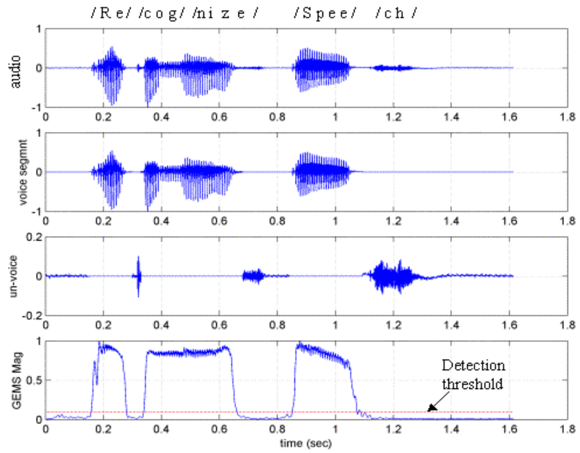

Figure 4. The GEMS signal is rectified and low pass filtered to produce a signal for speech detection and typing.

Note that any GEMS signal above a preset detection threshold will be classified as voiced speech segments (2^{nd} trace). Intervals below the threshold will be classified as either silence or unvoiced segments. The unvoiced segments are detected based on an audio signal energy threshold as shown in the 3^{rd} trace.

2.4. Unvoiced Speech Cepstral Processing

For unvoiced speech, such as leading or trailing fricatives, the signal spectrum are broadband-like with a slow variation in spectral envelope. Therefore, it is effective to model with cepstral coefficients. Figure 5 shows an example of the technique and the spectrum of a fricative.

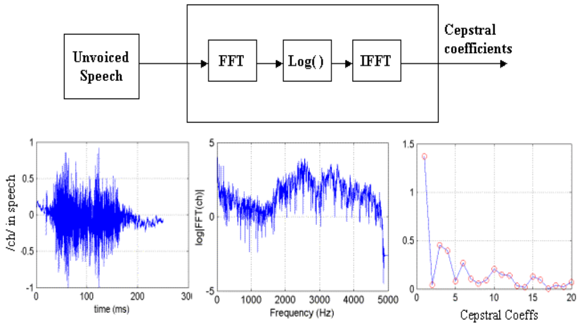

Figure 5. Example of unvoiced /ch/ cepstral processing.

2.5. Voiced Speech Auto Regressive Moving Average (ARMA) Processing

To process voiced speech, a processing frame of three glottal cycles with two overlapping cycles for each successive frame, was selected. The key computation steps are summarized in Fig. 6.

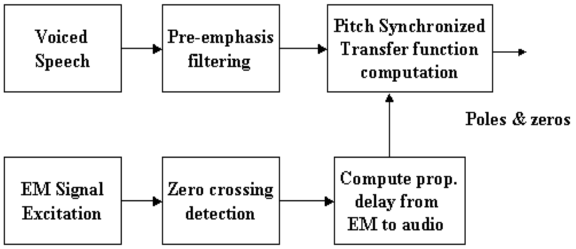

Figure 6. GBC based voiced speech processing.

In essence, the GEMS signal provides precision pitch extraction and pitch synchronized computation of transfer function. An ARMA process is used to represent the transfer function because the pole-zero representation provides a direct physical mapping to the spectral formants. Since a physical vocal motion is slowly varying, the resulting poles and zeros also behave similarly. The slowly varying motion allows us to compact the information into a small number of bits. Experiments have found that a minimum of 4 poles and 2 zeros are needed to model each phoneme as shown in Fig. 7. For compression, a pole-zero model is superior to an all-pole LPC model, since many poles (and bandwidth) are needed to model the presence of zeros in the transfer function.

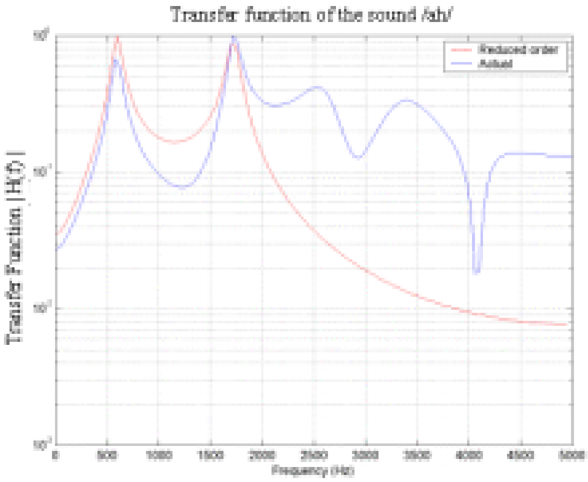

Figure 7. A minimum of 4 poles and 2 zeros per phoneme are needed for acceptable speech intelligibility.

2.6. The Compactor

The “compactor” operation as shown in Fig. 1, is at the heart of the low bandwidth GBC vocoder. As a comparison, the LPC-10 used 10 reflection coefficients for 41 bits, 7 for pitch, 1 voiced and unvoiced bit, 5 for a gain including a bit for synchronization for a total of 54 bits per frame at 44.44 frames/sec for a total of 2.4 kbps [8,9]. Starting with no compactors, we used a processing frame of two glottal cycles or about 20 ms per frame. During each frame we use 6 poles and 2 zeros for a total of 8 coefficients for 32 bits, 7 for pitch, 5 for a gain, 2 for voiced, unvoiced or silence, and 1 synchronization bit for a total of 2.4 kbps (48 bits/frame x 50 frames/sec). Thus without the compactor, it appears there is no advantage. Perhaps a small gain can be realized by using 4 poles instead of 6 poles per frame. This would reduce the bit rate by 400 bps.

The compactor operation first compresses the poles and zeros information by utilizing their relatively slow motion on the complex Z plane over a packet transmission interval say one second. Figure 8 shows the motion of the poles over the voicing portion of the word “PRINT.” Note that over a period of ~300ms the trajectory of the poles can be fitted with a cubic polynominal. Thus instead of transmitting 30 coefficients/pole, only 6 coefficients are needed, representing a 5-fold reduction in bandwidth. Similarly, pitch can be coded by a header and small changes using 3 bps. Unvoiced segments can be represented by a catalog of 8 fricatives. Thus one arrives at 480 bps for voiced, 280 bps for unvoiced, and 20 bps for silence. Now assuming an occurrence probability of 0.5, 0.2, and 0.3 for each type of speech respectively, then the average

GBC bandwidth (for this example) becomes approximately 312 bps.

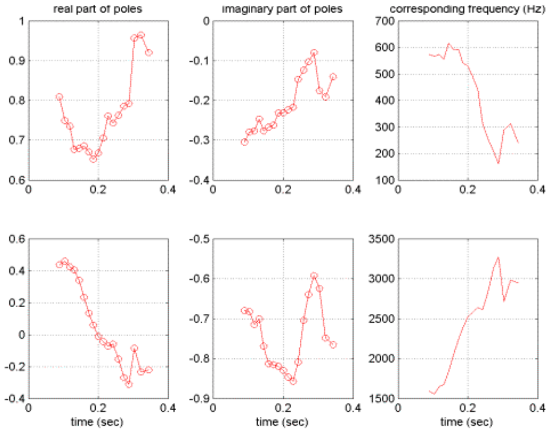

Figure 8. Example of temporal variation of real and imagery parts of poles.

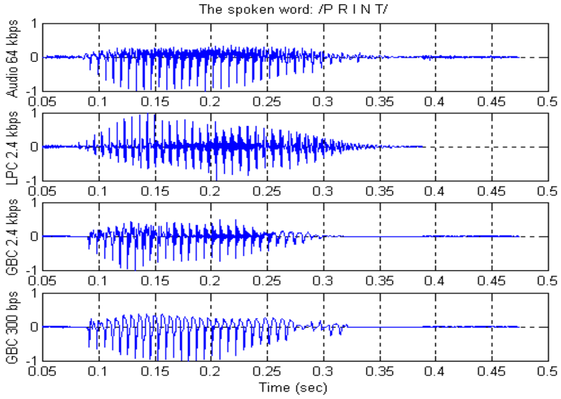

Figure 9. GBC vocoding of the word “PRINT” at 300 bps.

3. EXPERIMENTAL RESULTS

The GEMS Based Coded (GBC) vocoder was implemented along with the LPC-10 2.4 kbps standard. The GBC implementation allows us to vary the number of coefficients and the quantization bits. The output stream was written out to a data file, and used by the decoder to reconstruct the utterances. We have tested the vocoder for ten different sentences and ten different single words drawn from our data set [3]. Figure 9 shows four traces, the results for the spoken word “PRINT”: the original recording, the LPC-10, the GBC at 2.4 kbps, and finally the low bandwidth GBC at 300 bps. Comparative listenings of the waveforms led us to conclude that even at 300 bps, the intelligibility or quality of the sound is good.

4. CONCLUSION

We have demonstrated a low bandwidth vocoder using an EM sensor based coding approach. We found that even at 300 bps, the quality of the speech is more than adequate for communication. The heart of this approach is the compactor's ability to track the poles and zeros of the transfer function which is generated by using the input excitation and the audio signal. In addition, the compactor robustly identifies the two major types of speech segments and the non-speech segments.

5. ACKNOWLEDGMENTS

We thank the National Science Foundation for their support. This work was performed under the auspices of the U.S. Department of Energy by the University of California Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48.

6. 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, University Microfilms, Inc., Ann Arbor, Michigan, document #9925723, January 15, 1999.

[3] 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 database) Lawrence Livermore National Laboratory, UCRL MI-132776.

[4] Ng, L. C., Burnett, G. C., Holzrichter, J. F. and Gable, T. J. "Denoising of Human Speech Using Combined Acoustic and EM Sensor Signal Processing," Lawrence Livermore National Laboratory, UCRL-JC-136631, presented at IEEE ICASSP-2000, Istanbul, Turkey, June 6, 2000.

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

[6] Deller, J.R., Hansen, J.H., and Proakis, J.G., *Discrete-Time Processing of Speech Signals*, IEEE Press, 2000.

[7] Tremain, T.E. "The government standard linear predictive coding algorithm:LPC-10," *Speech Technology*, vol. 1, pp. 40-49, April 1982.

[8] Rahikka, D.J., Krebs, R.E. and et. El., "STU-III Secure Bearer Service Option Invocation Reliability Over Enhanced IS-136 TDMA Digital Cellular ACELP IS-641 Vocoder," *Proc. Of IEEE Vehiclar Technology Conference*, pp. 657-661, May 1997,