MULTISENSOR MELPE USING PARAMETER SUBSTITUTION*

Kevin Brady, Thomas F. Quatieri, Joseph P. Campbell, William M. Campbell, Michael Brandstein, Clifford J. Weinstein

MIT Lincoln Laboratory, 244 Wood St., Lexington, MA 02420-9185, USA
{kbrady, quatieri, jpc, wcampbell, msb, cjw}@ll.mit.edu

ABSTRACT

The estimation of speech parameters and the intelligibility of speech transmitted through low-rate coders, such as MELP, are severely degraded when there are high levels of acoustic noise in the speaking environment. The application of nonacoustic and nontraditional sensors, which are less sensitive to acoustic noise than the standard microphone, is being investigated as a means to address this problem. Sensors being investigated include the General Electromagnetic Motion Sensor (GEMS) and the Physiological Microphone (P-mic). As an initial effort in this direction, a multisensor MELPe coder using parameter substitution has been developed, where pitch and voicing parameters are obtained from GEMS and P-Mic sensors, respectively, and the remaining parameters are obtained as usual from a standard acoustic microphone. This parameter substitution technique is shown to produce significant and promising DRT intelligibility improvements over the standard 2400 bps MELPe coder in several high-noise military environments. Further work is in progress aimed at utilizing the nontraditional sensors for additional intelligibility improvements and for more effective lower-rate coding in noise.

1. INTRODUCTION

The Defense Advanced Research Projects Agency (DARPA) is currently sponsoring an Advanced Speech Encoding (ASE) program for addressing improvements to low-rate speech encoding in noisy military environments by using multiple, nonacoustic and nontraditional sensors to augment the acoustic microphones. Sensors being investigated include the General Electromagnetic Motion Sensor (GEMS) and the Physiological Microphone (P-mic). The long term goal of the program is to demonstrate high-performance, robust vocoding at rates less than 1000 bps. As a step toward this goal, some of the current efforts are aiming to exploit the multiple sensors to achieve 2400 bps speech intelligibility performance in a high-noise environment which approaches or matches the intelligibility of 2400 bps MELPe [9] performance in a substantially lower noise environment. This paper describes a first effort in this direction, in which a Multisensor MELPe coder using parameter substitution has been developed, where pitch and voicing parameters are obtained from GEMS and P-Mic sensors, respectively, and the remaining parameters are obtained from a standard acoustic microphone.

Section 2 will discuss a multiple sensor corpus (referred to as the DARPA ASE Pilot Corpus) that has been developed and is being used for investigating new approaches to low-rate voice coding. Section 3 will describe the architecture of a multisensor MELPe coder using parameter substitution. This new architecture is based on the use of speech excitation pitch and voicing information from the GEMS and P-mic sensors. Section 4 will discuss the Diagnostic Rhyme Test (DRT) [8] results corresponding to parameter substitution in 2400 bps MELPe on speech from the multisensor corpus, showing significant and promising DRT intelligibility improvements over the standard 2400 bps MELPe coder in several high-noise military environments. Section 5 will provide concluding remarks and discuss future work, including potential improvements to the parameter substitution approach, and a number of other approaches which can potentially utilize nontraditional sensors for additional intelligibility improvements and for more effective, lower-rate coding in noise.

2. CORPORA AND SENSOR MEASUREMENTS

A Pilot Corpus has recently been collected for the DARPA ASE program by ARCON Corporation, under subcontract to MIT Lincoln Laboratory. The corpus consists of ten male and ten female speakers. The content includes DRT word lists, Harvard Sentence lists, and Consonant Vowel Consonant (CVC) nonsense words in a carrier phrase. This content was recorded in a number of

* This work was 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.

different noise environments: Quiet, Office Environment, M2 Bradley Fighting Vehicle (M2) Environment, Military Operations in Urban Terrain (MOUT) Environment, UH-60 Blackhawk Helicopter Environment (BH), and a Military Command Enclosure (MCE) Environment. The M2, MOUT, and BH environments were recorded in low-noise and high-noise conditions that were offset by 40 dBC SPL. A total of six sensors were recorded during all sessions: a two channel glottally-located GEMS, a throat-located P-mic, a forehead-located P-mic, a glottally-located Electroglottograph (EGG), a resident microphone on the talker's helmet corresponding to the microphone used in that military environment, and a Bruel & Kjaer reference microphone located in front of the talker that was used in the original noise recording.

The GEMS utilized in this corpus was developed by Aliph Corporation and is based on earlier work done at Lawrence Livermore National Laboratory [2]. It is an RF sensor that is placed in direct contact with the skin. In the Pilot Corpus it was placed on the throat directly over the glottis. The GEMS appears to measure vibrations of the tracheal wall [2] during voiced speech as well as during voice bars [6].

The P-mic is a piezoelectric sensor with a gel pack for contact with human skin. It was developed at the Army Research Laboratory for measuring physiological processes such as heart rate and respiration [7], and has since been utilized as a speech sensor [1]. For the Pilot Corpus, the P-mic located on the forehead provided vocal tract and excitation content with low SNR. The P-mic located on the throat provided good noise attenuation (~30 dB) with good excitation information, though little vocal tract content was available when the P-mic was located below the glottis. The P-mic signal tends to be low-pass, with significant roll off above 1-2 kHz.

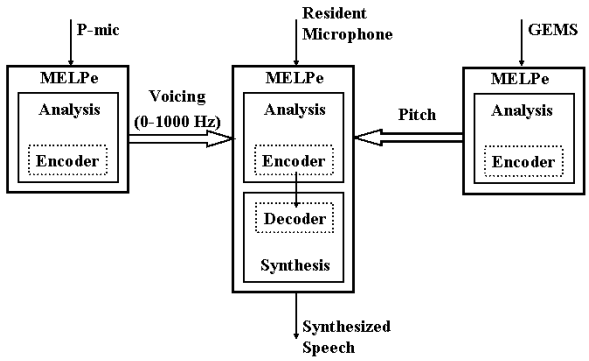

Figure 1. Multisensor MELPe architecture

3. MULTISENSOR MELPE

The Mixed Excitation Linear Prediction (MELP) [5] coder is the U.S. Federal Standard at 2400 bps. MELPe is based on MELP, with the addition of a noise preprocessor [4] using a minimum statistics approach to estimating the noise background and a harmonic synthesizer. In the 2400 bps version of MELP/MELPe an analysis/synthesis frame interval of 22.5 ms is used with 54 bits encoded per frame interval. The parameters encoded at each interval are five bandpass voicing decisions, ten line spectral frequencies, ten Fourier magnitudes, two gains, a pitch, and a pitch jitter flag. The five bandpass voicing regions are 0-500 Hz, 500-1000Hz, 1000-2000 Hz, 2000-3000 Hz, and 3000-4000 Hz. The encoding software used in the following experiments is the fixed point 2400 bps MELPe Version 8.0 [9] in its standard configuration with the noise preprocessor and postfilter activated.

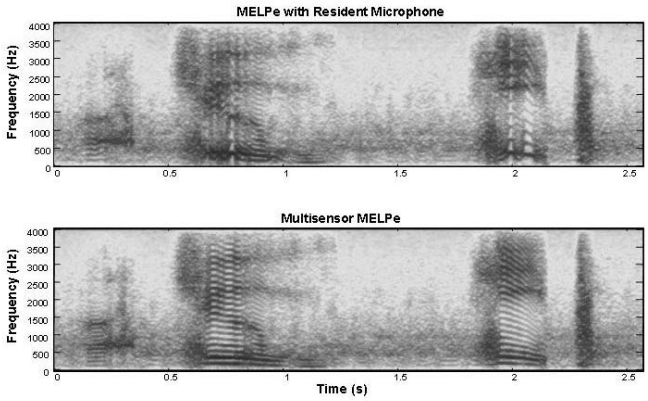

Figure 2. MELPe Spectrograms

A Multisensor MELPe architecture has been implemented as shown in Figure 1. The architecture utilizes the resident acoustic microphone, the P-mic, and a GEMS sensor. The latter two sensors are used to calculate MELPe parameters that can be substituted into the MELPe encoding of the resident acoustic microphone channel. The P-mic is used for encoding the bandpass voicing in the two lowest bandpass regions (0-1000 Hz). Since the pitch is calculated multiple times in the MELPe architecture it is necessary to substitute these intermediate pitch estimates into the encoding of the resident acoustic microphone channel. These intermediate pitches include the initial integer pitch estimate, the fractional pitch calculated during bandpass voicing calculations, and the final smoothed pitch. The standard MELPe noise preprocessor was utilized to preprocess the three sensor channels before the MELPe parameter estimation step. The MELPe postfilter was applied to the synthesized speech from the multisensor MELPe architecture. Example spectrograms of the MELPe and multisensor MELPe outputs are shown in Figure 2, illustrating the effect of improved pitch and voicing for a waveform in the M2 high-noise environment. Note the improved harmonic structure with Multisensor MELPe.

4. INTELLIGIBILITY TESTING

Intelligibility can be thought of as having dual natures [8]. The perceptual aspect addresses the discrimination of phonetic sound based on the isolated acoustic sound without any context. The nonperceptual portion addresses the inference of speaker intent using inference-based context. The DRT [8] is often used to evaluate the perceptual aspect of speech intelligibility and has been adopted by the DARPA ASE Program for benchmarking the intelligibility of tested vocoders. The DRT test is further refined into 6 attribute tests (voicing, nasality, sustention, sibilation, graveness, and compactness) that measure qualitative differences in intelligibility. The test consists of word pairs that differ only in their leading consonant sound (eg. 'veal', 'feel'). Listeners are presented with a choice from a word pair, and select which of the 2 words that they perceive to have heard. A total of 8 listeners were used at ARCON Corporation for the testing in this section.

The Multisensor MELPe processed speech of three male (M1, M3, M6) and three female speakers (F4, F5, F7) from the DARPA ASE Pilot Corpus were submitted to ARCON Corporation for DRT evaluation. Four noise environments were evaluated: MCE environment (79 dBC SPL), Blackhawk (BH) high-noise environment (110 dBC SPL), M2 high-noise environment (114 dBC SPL), and the MOUT high-noise environment (113 dBC SPL). These results were compared with the DRT results obtained by ARCON Corporation for these same speakers using the standard MELPe Fixed Point encoder with noise preprocessing and postfiltering. The results can be seen in Table 1. In addition to the results corresponding to a high-noise condition are results corresponding to a low-noise condition using the same speakers and a noise field at a level down 40 dBC SPL. The MCE environment was only collected at one SPL. The results from Table 1 have been broken out for male and female speakers, respectively, in Tables 2 and 3.

Improvements for Multisensor MELPe over MELPe exceeding the magnitude of the standard errors are seen in all four noise environments. Particularly impressive absolute gains are seen in the M2 and MOUT environments. The MOUT environment shows the most impressive relative gains since the DRT gap for the MELPe low-noise and high-noise environments is smaller. This is true for both male and female speakers. The intelligibility gains in the M2 environment appear strongest for male speakers.

An analysis of the average DRT attribute scores for the investigated noise environments is shown in Table 4. A strong performance improvement is exhibited for the voicing attribute. An example of a DRT voicing word pair is 'veal'-'feel'. The improvement in the voicing attribute is not surprising due to the voicing substitution aspect of the Multisensor MELPe architecture. Typically, the voicing attribute is not strongly affected by acoustic noise [3], though strong degradation was seen for this attribute for MELPe in the Blackhawk, M2, and MOUT high-noise environments.

The sustention attribute also exhibits a strong improvement across all of the noise environments. This attribute distinguishes consonants by their temporal length or by the gradualness of onset for sustained consonants [3]. An example of a DRT sustention word pair is 'vee'-'bee'. The success of the multisensor architecture is noteworthy since the sustention attribute is typically strongly degraded by acoustic noise and vocoding [3, 8]. The other four DRT attributes did not show consistent improvement or degradation across the noise environments. Their average scores were quite close to their baseline scores using MELPe in a high-noise environment, as seen in Table 4.

5. CONCLUDING REMARKS

A Multisensor MELPe using parameter substitution has shown DRT intelligibility improvements across four tested noise environments for both male and female speakers. The architecture utilized two auxiliary sensors that are robust in providing speech information in acoustic noise. The P-mic provided low-frequency voicing information, while the GEMS provided pitch information. The Multisensor MELPe architecture show strong absolute DRT intelligibility gains in the MOUT and M2 environments. The voicing and sustention DRT attributes showed strong gains in all environments.

The promising results reported here are viewed as just a first step in the application of multisensor techniques to noise-robust low-rate coding. Several improvements to the Multisensor MELPe architecture are under consideration. A dynamic programming-based pitch estimation algorithm using GEMS and P-Mic is under consideration. A multisensor noise preprocessor [6] is under development, which can potentially be used in conjunction with parameter substitution to obtain additional intelligibility improvements. More detailed analysis of the information inherent in the GEMS and P-Mic signals is underway, with the aim of using these signals in advanced coders both to reduce sensitivity to noise and to achieve high performance at lower bit rates.

6. REFERENCES

[1] J.D. Bass, M. V. Scanlon, T. K. Mills, John J. Morgan, "Getting Two Birds with One Phone: An Acoustic Sensor for Both Speech Recognition and Medical Monitoring", *Acoustic Society of America*, Columbus, Ohio, Nov. 1999

[2] G. C. Burnett, J. F. Holzrichter, T. J. Gable, and L. C. Ng, "Denoising of Human Speech using Combined Acoustic and EM

Sensor Signal Processing", *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Istanbul Turkey, June 2000

[3] Dynastat Document, "Interpretation of Diagnostic Rhyme Test Scores for Voice Communication Systems"

[4] R. Martin, I. Wittke, and P. Jax, "Optimized Estimation of Spectral Parameters for the Coding of Noisy Speech", *ICASSP 2000*, Istanbul Turkey, pp. 1479-1482, June 2000

[5] A. McCree, K. Truong, E. B. George, T. P. Barnwell, V. Viswanathan, "A 2.4 kbit/s MELP Coder Candidate for the New U.S. Federal Standard", *ICASSP 1996*, Atlanta GA

[6] T. F. Quatieri, D. Messing, K. Brady, W. B. Campbell, J. P. Campbell, M. Brandstein, C. J. Weinstein, J. D. Tardelli, and P. D. Gatewood, "Exploiting Nonacoustic Sensors for Speech Enhancement", submitted to *Workshop on Multimodal User Authentication*, Santa Barbara, CA, 11-12 December 2003

[7] M. V. Scanlon, "Acoustic Sensor for Health Status Monitoring", *Proceedings of IRIS Acoustic and Seismic Sensing*, Vol. 2, pp. 205-222, 1998

[8] W. D. Voiers, "Diagnostic Evaluation of Speech Intelligibility", *Benchmark Papers in Acoustics*, Vol. 11: *Speech Intelligibility and Speaker Recognition* (M. Hawley, ed.) Dowden, Hutchinson and Ross, Stroudsburg (1977)

[9] T. Wang, K. Koishida, V. Cuperman, A. Gersho, and J. Collura, "A 1200/2400 bps Coding Suite Based on MELP", 2002 *IEEE Workshop on Speech Coding*, Tsukuba, Japan, 6-9 October 2002

Table 1. Composite DRT results for three male and three female speakers

	MCE	BH	M2	MOUT
MELPe (Low Noise)		92.69	93.75	93.38
Multisensor MELPe (High Noise) (Standard Error)	90.15 (0.38)	83.64 (0.51)	79.36 (0.87)	89.61 (0.63)
MELPe (High Noise)	88.65	82.87	76.67	86.63

Table 2. Composite DRT results for three male speakers

	MCE	BH	M2	MOUT
MELPe (Low Noise)		91.54	93.97	92.84
Multisensor MELPe (High Noise)	90.19	82.86	79.82	87.63
MELPe (High Noise)	89.28	81.50	75.74	84.94

Table 3. Composite DRT results for three female speakers

	MCE	BH	M2	MOUT
MELPe (Low Noise)		93.84	93.53	93.92
Multisensor MELPe (High Noise)	90.10	84.42	78.91	91.58
MELPe (High Noise)	88.02	84.24	77.60	88.32

Table 4. Average DRT attribute results for three male and three female speakers

	Voicing	Nasality	Sustention	Sibilation	Graveness	Compactness
MELPe (Low Noise)	95.62	97.83	92.02	91.71	87.07	95.40
Multisensor MELPe (High Noise)	89.18	89.65	81.09	86.56	77.05	90.85
MELPe (High Noise)	82.85	89.23	75.82	85.68	76.99	90.76