Exploiting Nonacoustic Sensors for Speech Encoding

Thomas F. Quatieri, *Fellow IEEE*, Kevin Brady, Dave Messing, Joseph P. Campbell, *Fellow IEEE*, William M. Campbell, Michael S. Brandstein, *Senior Member IEEE*, Clifford J. Weinstein, *Fellow IEEE*, John D. Tardelli, and Paul D. Gatewood

***Abstract*—The intelligibility of speech transmitted through low-rate coders is severely degraded when high levels of acoustic noise are present in the acoustic environment. Recent advances in nonacoustic sensors, including microwave radar, skin vibration, and bone conduction sensors, provide the exciting possibility of both glottal excitation and, more generally, vocal tract measurements that are relatively immune to acoustic disturbances and can supplement the acoustic speech waveform. We are currently investigating methods of combining the output of these sensors for use in low-rate encoding according to their capability in representing specific speech characteristics in different frequency bands. Nonacoustic sensors have the ability to reveal certain speech attributes lost in the noisy acoustic signal; for example, low-energy consonant voice bars, nasality, and glottalized excitation. By fusing nonacoustic low-frequency and pitch content with acoustic-microphone content, we have achieved significant intelligibility performance gains using the DRT across a variety of environments over the government standard 2400-bps MELPe coder. By fusing quantized high-band 4-to-8-kHz speech, requiring only an additional 116 bps, we obtain further DRT performance gains by exploiting the ear's insensitivity to fine spectral detail in this frequency region.**

***Index Terms*—Intelligibility, low-rate coding, nonacoustic sensors.**

I. Introduction

NONACOUSTIC sensors provide an exciting opportunity for multimodal speech processing with application to areas such as speech enhancement and coding. These sensors provide measurements of functions of the glottal excitation and, more generally, of the vocal tract articulator movements that are relatively immune to acoustic disturbances and can supplement the acoustic speech waveform. One sensor under consideration is the general electromagnetic motion sensor (GEMS) [2], [5], [6]. This microwave radar sensor was originally developed at Lawrence Livermore National Laboratory and is currently under development by Aliph Corporation. A second sensor under consideration is a piezo-electric-based vibrometer known as the physiological microphone (P-mic) [17] that was developed at the Army Research Laboratory. Several other sensors were also considered in the context of this paper's work, including the electroglottograph (EGG) [16] and a bone-conduction microphone [23].

This paper describes an approach to speech processing, with particular application to low-rate speech coding, that exploits these nonacoustic sensors according to their capability in representing specific speech characteristics in different frequency bands in harsh acoustic environments. This effort is a component of the ongoing Defense Advanced Research Projects Agency (DARPA) Advanced Speech Encoding (ASE) program in exploiting nonacoustic sensors for low-rate speech encoding in harsh environments. One objective of the first phase of the ASE program, and the focus of this paper, was to improve the intelligibility of the government standard 2400-bps Mixed Excitation Linear Prediction speech coder (MELPe) in harsh acoustic environments. Intelligibility of this voice coder was measured by the Diagnostic Rhyme Test (DRT) [21]. The specific corpus used for testing is the multisensor DARPA ASE Pilot Speech Corpus, collected in a variety of harsh military acoustic noise environments.

There has been previous work in the speech community that has shown speech coder intelligibility gains by fusing low-frequency spectral information from nonacoustic sensors. Viswanathan *et al.* [20] have shown that low-frequency spectral content from an accelerometer, combined with the degraded acoustic signal from a gradient microphone, is able to improve intelligibility of an early 2400-bps linear predictive coder (LPC). Our work extends and generalizes this technique by exploiting the frequency dependence of the more recent GEMS and P-mic, as well as a bone-conduction microphone. The sensor outputs are fused according to the quality of their signals in different frequency bands, both with respect to the excitation and vocal tract information, as well as noise level [12], [15]. We show that these nonacoustic sensors can reveal certain speech attributes lost in the noisy acoustic signal; for example, low-energy consonant voice bars, nasality, and glottalized excitation. In addition, we explicitly use the pitch of nonacoustic sensors in the synthesis and encoding stage of the MELPe coder, providing additional robustness.

Previous work has also been performed in exploiting high-frequency content above the 4000-Hz bandwidth of the synthesis of typical low-rate coders. Kang and Everett [7] have shown that spectral mirror-imaging near the 4000-Hz bandwidth cutoff can give added intelligibility, also in an early LPC coder. More recently, McCree has demonstrated quality gains with high-band parametric coding at 14 kbps [11]. In our work, we propose a simple scheme of coding only the temporal energy variations over a 4000- to 8000-Hz region, with a fixed representative spectrum, resulting in significant intelligibility gain.

Manuscript received July 30, 2004; revised January 3, 2005. This work was supported 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. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Gerald Schuller.

T. F. Quatieri, K. Brady, D. Messing, J. P. Campbell, W. M. Campbell, M. S. Brandstein, and C. J. Weinstein are with the MIT Lincoln Laboratory, Lexington, MA 02420 (e-mail: quatieri@ll.mit.edu; kbrady@ll.mit.edu; dmessing@mit.edu; jpc@ll.mit.edu; wcampbell@ll.mit.edu; msb@ll.mit.edu; cjw@ll.mit.edu).

J. D. Tardelli and P. D. Gatewood are with ARCON Corporation, Waltham, MA USA (e-mail: jdtardelli@arcon.com; pdg@arcon.com).

Digital Object Identifier 10.1109/TSA.2005.855838

1558-7916/$20.00 © 2006 IEEE

In Section II of this paper, we first describe the GEMS, P-mic, bone-conduction microphone, and EGG nonacoustic sensors, as well as the DARPA ASE Pilot Speech Corpus recorded in a variety of harsh noise environments. In Section III, we show the nature of the sensor outputs by example and illustrate their frequency dependence. Section IV describes elements of an enhanced multisensor 2400-bps MELPe coder, based on the results of Section III, and introduces a composite system based on these elements. DRT evaluations are provided for system components as well the composite system. These multisensor schemes use the GEMS, P-mic, bone-conduction microphone, and acoustic sensors in different frequency bands. In Section IV, we also discuss the implications of specific DRT attributes given in our evaluation and, in Section V, we further enhance our system with high-frequency fusion. Finally, in Section VI, we summarize and give future directions.

II. Nonacoustic Sensors and Corpus Collection

In this section, we overview the numerous sensors of interest and describe the sensor corpus collection and environmental conditions.

A. GEMS

The GEMS measures tissue movement during voiced speech; i.e., speech involving vocal chord vibrations [2], [5], [6]. An antenna is typically strapped or taped on the throat at the laryngeal notch, but also can be attached at other facial locations. This sensor emits a 2-GHz electromagnetic signal that penetrates the skin and reflects off the speech production anatomy, such as the tracheal wall, the vocal folds, or the vocal tract wall. Because signals collected from a GEMS device depend on the tissue movement in the speech production anatomy, it is relatively immune to degradation from external acoustic noise sources.

During voiced speech activity, GEMS records quasi-periodic electromagnetic signals due to vibration of the speech production anatomy. When placed at the larynx, quasi-periodic measurements are found during vowels, nasals, and voiced consonants, including before and after the burst in voiced plosives, i.e., during voice bars. There are also precursor signals seen with the GEMS that precede the acoustic signal at the lips for both plosive and fricative consonant voice bars and for nasals, voiced affricates, and voice approximants. The GEMS also sometimes shows quasi-periodicity following the acoustic revelation of these events. These GEMS signals are seen in both benign and severe acoustic background noise cases, but moreso in the later. They are not always present for a specific talker and their strength and duration appear to be talker dependent. Single pulses have also been observed sporadically at the burst in unvoiced plosive consonants, as well as with irregular vibration from, for example, a glottalized source.

B. P-Mic

The P-mic is composed of a gel-filled chamber and a piezo-electric sensor behind the chamber [17]. Vibrations permeating the liquid-filled chamber are measured by the piezo-electric sensor that provides an output signal in response to applied forces generated by movement, converting vibrations traveling through the liquid-filled chamber into electrical signals. The liquid-filled chamber is designed to have poor coupling between ambient background noise and the fluid-filled pad, thus attenuating vibrations of unwanted ambient background noise.

Like the GEMS sensor, the P-mic can be strapped or taped on various facial locations. The P-mic on the throat below the laryngeal notch primarily measures vocal fold vibrations with quasi-periodic measurements similar to that of GEMS. The P-mic signal at the throat above the laryngeal notch, however, contains some low-pass vocal tract formants with bandwidths wider than normal. Other facial locations can provide additional vocal tract characteristics, though the signal-to-noise ratio (SNR) is typically not as good as at the throat. The P-mic located on the forehead, for example, gives significant vocal tract information, but is less noise-immune than the P-mic at the throat in severe environments. The precursor signals seen with the GEMS are also present with the P-mic when it is positioned at the throat.

C. Bone-Conduction Microphone

In a bone-conduction microphone (mic), voice content is transmitted by way of bone vibrations [23]. As with the P-mic, bone-conduction mics comprise a piezo-electric-like material. In the context of the ASE program, the bone-conduction mic is located in the top of the helmet used in one particular acoustic environment (see Section II-E). As such, bone vibrations are picked up at the top of the skull. At this skull location, the bone-conduction mic is found to provide strong voicing, as well as significant vocal tract content in a midfrequency range up to about 1000 Hz. Precursor signals have also been seen with the bone-conduction mic.

D. EGG

The EGG sensor measures vocal fold contact area by providing an electrical potential (of about 1 V rms and 2–3 MHz) across the throat at the level of the larynx [16]. With a pair of gold-plated electrodes, the sensor measures the change of impedance over time. When the vocal folds are closed, the impedance is decreased; when they are open, the impedance is increased. Thus, the opening and closing of the vocal folds, present in voiced speech, are measured by the EGG. Precursor signals are evident with the EGG sensor.

E. Corpus Collection

An extensive multisensor speech corpus was collected at ARCON Corporation from ten male and ten female talkers. Scripted phonetic, word, and sentence material along with conversational material were generated by each talker. These materials were generated in nine different acoustic noise environments. For each speaker, the corpus was collected in two sessions on two different days. Speakers were exposed to a variety of noise environments, including both benign and severe cases. Three of the acoustic environments were each presented at two intensity levels that differed by 40 dBC sound pressure level (SPL). Specific environments are quiet, office (56 dBC), mobile command enclosure (MCE, 79 dBC), M2 Bradley Fighting Vehicle (74 dBC and 114 dBC), military operations in urban terrain (MOUT, 73 dB and 113 dBC), and a Blackhawk helicopter (BH, 70 dBC and 110 dBC). We will sometimes refer to these environments as (with L indicating low noise and H indicating high noise) quiet, office, MCE, M2L, M2H, MOUTL, MOUTH, BHL and BHH, respectively. The sound recordings

TABLE I
TRIANGULAR MATRIX SHOWING RELATIVE PAIR-WISE DELTA SPL dBC DIFFERENCES OF THE EIGHT CONDITIONS. SPL MEASUREMENTS FOR THE MOUT CONDITIONS (SHOWN WITH LIGHT GRAY BACKGROUND) WERE MADE IN SECTIONS OF ALMOST CONTINUOUS DISTURBANCE

Acoustic Noise Conditions – Delta SPLs (dBC)								
Condition	Office	MCE	M2 (High)	M2 (Low)	MOUT (High)	MOUT (Low)	BH (High)	BH (Low)
Office Environment		23	58	18	57	17	54	14
MCE Environment			35	5	34	6	31	9
M2 Environment (High)				40	1	41	4	44
M2 Environment (Low)					39	1	36	4
MOUT Environment (High)						40	3	43
MOUT Environment (Low)							37	3
BH Environment (High)								40
BH Environment (Low)								

for the quiet, office, MCE, M2H, MOUTH, and BHH are field recordings made by ARCON Corporation. The MOUTL and M2L are identical to the MOUTH and M2H environments, except for their presented SPL. The BHL corresponds to a field recording made inside the closed cockpit presented at 40 dBC SPL below the presentation level of the BHH, which corresponds to a different field recording outside of the cockpit. Our complete recording conditions are summarized in Table I. In addition to the 40-dBC differences within environmental conditions, there are intermediary high-/low-noise differences across conditions.

For each talker and environment combination, time-synchronous data was collected from up to seven separate sensors. These sensors consisted of the previously introduced GEMS, P-mic, and EGG. Data was also collected from two acoustic microphones, a high-quality B&K calibration microphone and an environment-specific microphone that is referred to as the *resident microphone*. The resident microphone (mic) was typically the first-order gradient noise-cancelling microphone used for normal communications in that specific military environment. In the MOUT condition, however, two resident mics were present: an acoustic gradient mic and a bone-conduction mic.

One GEMS and one EGG were located near the talker's larynx. Careful attention was given to tuning the GEMS sensor and in optimizing its placement. The GEMS was considered the primary sensor during the corpus collection. A specific talker's neck and shoulder geometry often required that tradeoffs be made in the placement of the secondary sensors in order to optimize the GEMS signal. Two P-mics were used: one located in the vicinity of the talker's larynx and the other on the talker's forehead.

Due to the acoustic presentation levels of some of the noise environments, all talkers used the acoustic protection systems typical of each specific noise environment. This normally consisted of some type of communication headset that provided noise attenuation on the order of 20 dB. Human subject procedures were followed carefully and noise exposure was monitored.

The complete corpus consists of up to eight channels of data from approximately 20 min of speech material in each of nine acoustic noise environments from each of the 20 talkers. All sensor signals were sampled at 48 kHz, though the nonacoustic data was downsampled to 16 kHz for space considerations. The full corpus takes nearly 70 GB of storage.

III. NATURE AND FUSION OF SENSOR MEASUREMENTS IN HARSH NOISE CONDITIONS

There are numerous speech events that may be characterized by the pattern of signal energy across specific frequency bands. For example, unvoiced fricative and plosive sounds are high in frequency, while nasals and the voice-bar components of voiced plosives are low in frequency. These events are not always measured by the acoustic (resident) microphone, or by the nonacoustic sensors, due to the harsh noise environment or due to the characteristics of the sensor. Fortunately, sensor measurements can be somewhat complementary. Consequently, we developed an approach that exploits the propensity of the various sensors to detect speech events in different frequency bands. In this section, we present qualitative discussion on the nature of the sensor measurements. Then, based on this understanding, in Sections IV and V, we introduce schemes for fusing the sensor outputs used in a variety of low-rate speech coding scenarios.

A. Frequency Dependence

For each sensor (GEMS, P-mic, and bone-conduction mic), we investigated properties of its signal output in the time and frequency domains. For each sensor output, we studied source components with respect to voicing (including voice bars), frication, consonant bursts, and glottalization. In addition, we looked at vocal tract components with respect to formant location and bandwidth, as well as low-frequency contributions, such as nasality. Time-frequency properties were studied in relation to those of the corresponding resident acoustic microphone signal in harsh acoustic backgrounds.

As we alluded to earlier, signal quality and presence of speech events of nonacoustic and acoustic outputs are band-dependent. For the resident-mic acoustic signal, we found the following properties.

- Weak low-frequency content with voice bars and nasal energy severely attenuated or degraded.
- Strong high-frequency content, particularly in noise fields with strong roll offs at high frequencies.

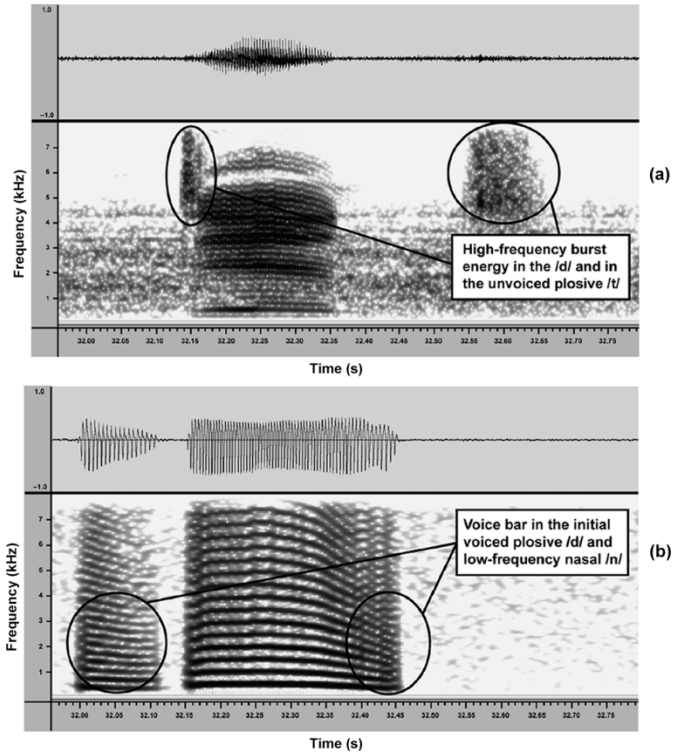

Fig. 1. Waveforms (from the M2H environment) and spectrograms of the (a) resident-mic and (b) GEMS outputs for the DRT word "dint."

- Weak irregular voiced source content, particularly in low-energy; for example, with glottalization at word endings often lost.
- Moderate noise immunity, but significantly greater than the B&K reference microphone.

An example of the first two properties is shown in Fig. 1(a), with a waveform (from the M2H environment) and spectrogram of the resident-mic output for the word "dint" taken from a DRT word list of the ASE Pilot Speech Corpus. The resident mic shows the high-frequency burst energy in the /d/ and in the unvoiced plosive /t/, but poor low-frequency content. For certain harsh conditions of interest, the SNR is high for frequencies within about a 3-kHz to 8-kHz range. This is advantageous because, typically, several consonants, such as an /s/, /sh/, /ch/, and /th/, contain significant energy in this frequency region. For these same conditions, however, the resident mic has a poor SNR for low frequencies, particularly below about 500 Hz, and thus fails at representing low-frequency events such as nasals and voice bars.

On the other hand, we find more reliable low-frequency energy in the GEMS and, typically, other nonacoustic sensors. For the GEMS, we found the following signal properties.

- Strong low-frequency source content in low-frequency voicing, including voice bars and nasality.
- Glottalization in low-energy regions.
- Essentially no vocal tract content.
- Strong noise immunity.

An example is shown in Fig. 1(b) where the GEMS signal clearly gives the presence of the low-frequency content of the nasal /n/ and voice bar in the voiced plosive /d/ in the word "dint". The resident mic, while not revealing the nasal and voice bar, more clearly shows the high-frequency burst energy in the /d/ and in the unvoiced plosive /t/. In a second example,

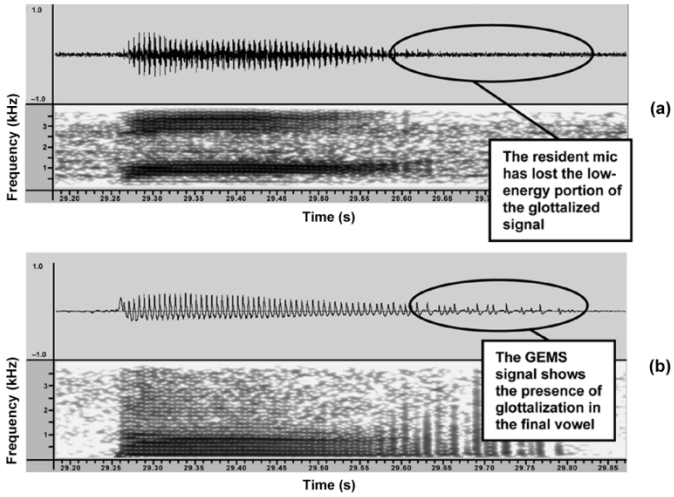

Fig. 2. Waveforms (from the M2H environment) and spectrograms of the (a) resident-mic and (b) GEMS outputs for the DRT word "bon." The GEMS signal shows the presence of the glottalization at the word termination, while the resident mic does not clearly reveal the low-SNR region of the irregular vocal-fold vibration.

Fig. 3. Frequency-based contributions of acoustic and nonacoustic sensors at low SNR.

glottalization is measured in the GEMS and is shown in Fig. 2, with waveforms (from the M2H environment) and spectrograms of the resident-mic and GEMS outputs. The word "bon" was again taken from a DRT word list of the ASE Pilot Corpus. Here we see that glottalization, typically occurring at the end of a word or phrase, is essentially lost in the acoustic signal, but clearly revealed in the GEMS output. Other forms of glottalized activity, such as secondary glottal pulses within a pitch period as in diplophonic speech, are also more clearly seen by GEMS in acoustic noise backgrounds. Generally, we found that voice bars and other precursors and voicing at the end of consonants are measurable by the GEMS, P-Mic, bone-conduction mic, and EGG, while glottalization is more clearly measured by the GEMS. As mentioned in Section II, the strength and duration of these vibrations appear to be speaker-dependent, as well as condition-dependent, being more present with harsher noise.

A summary of our findings for the sensors of interest is illustrated in Fig. 3. The focus of this matrix is on the sensors and the speech attributes that influenced our fusion-based enhancement for speech encoding at low SNR. The EGG, although important as a reference and for scientific study, is not considered

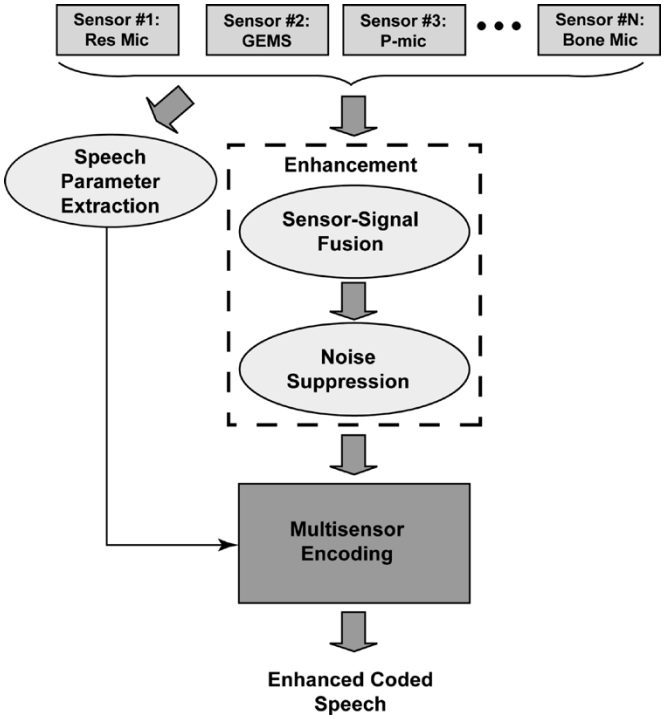

Fig. 4. General strategy for enhanced multisensor speech encoding.

practical for field use; and glottalization, although an important speech characteristic, has not yet been explicitly exploited. We should emphasize that the association of a speech attribute with a sensor is a loose one, based on observation, and was used as a guide in forming fusion and parameter substitution strategies described in the following sections.

IV. Multisensor Encoding

Since the various sensors capture different speech attributes and their output fidelity is frequency dependent and relies on the sensor and its placement, our general strategy is to construct a speech signal estimate from a fusion of components of nonacoustic and acoustic sensor signals. Each sensor output is filtered in a spectral region determined by our ad hoc attribute-sensor matrix of Fig. 3. The filtered outputs are then combined linearly and passed on to the encoding stage.

Our general strategy for enhanced encoding is shown in Fig. 4. We can think of our sensor fusion as one component of a preprocessing enhancement stage. Here we are adding sensor outputs from different frequency bands. The second enhancement component is noise suppression. Since we are working in the context of the government standard 2400-bps MELPe coder, we use, in this scenario, the MELPe preprocessor for noise suppression, although any noise suppression algorithm can be applied.

In addition to sensor fusion, we have included in our general strategy a speech parameter extraction module. In this component, parameters used by MELPe are estimated from the nonacoustic sensors, rather than the acoustic waveform. Pitch and voicing parameters, for example, may be estimated from the GEMS and P-mic signals and then substituted in the encoder; the remaining parameters are obtained from the enhanced fused signal.

In this section, we further describe each module in Fig. 4 and its DRT performance. In addition, we look at DRT performance of a fully integrated system. The DRT testing was conducted at ARCON Corporation and consists of word pairs that differ only in their leading consonant sound (e.g., "veal", "feel"). Listeners are presented with a choice from a word pair and select which of the two words they perceive to have been spoken [4], [21].

We begin with a brief review of the government standard 2400-bps MELPe coder.

A. MELPe

The MELP coder [10] is the U.S. Federal Standard at 2400 bps (MilStandard MIL-STD-3005) [18]. MELPe, a variant of the government standard MELP, is the new draft NATO standardization agreement (STANAG) 4591 coder [22]. The enhancement over MELP consists primarily of an integrated noise preprocessor (for acoustic-noise reduction) and a harmonic synthesizer (for mixing the periodic- and noise-excitations using a two-band model). Another enhancement is the addition of a 1200-bps rate to the conventional 2400-bps MELP rate. The MELPe noise preprocessor is a single-microphone speech-enhancement preprocessor that has been developed for voice communication in nonstationary acoustic-noise environments. The preprocessor algorithm was developed in conjunction with the MELP coder which, by itself, is susceptible to environmental noise. This integrated preprocessor is based on short-time spectral-amplitude estimation, soft-decision gain modification, tracking of the *a priori* probability of speech absence, and minimum statistics noise power estimation. Special emphasis is placed on enhancing the performance of the preprocessor in nonstationary noise environments [8], [9].

In the 2400-bps version of the MELP/MELPe coder, 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 MELPe coder 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, 500–1000 Hz, 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 [22] in its standard configuration with the above noise preprocessor and a postprocessor used to reduce noise in formant nulls and to narrow broadened formant bandwidths introduced in the encoding process.

B. Low-Frequency Fusion

One particular fusion configuration is shown in Fig. 5(a) [15]. In this scheme, we use the P-mic signal for the low-band 0–300 Hz and the resident-mic signal for the band 300–4000 Hz. Motivation for this fusion scheme is that the P-mic signal is relatively noise-immune and contains good vocal source and vocal tract content in the low-frequency region. The MELPe preprocessor is then applied and its output passed on to the 2400-bps MELP encoder. The low- and high-pass filters together form an identity.

The intelligibility impact of fusing the resident microphone and P-mic signals has been determined for coded and uncoded speech through formal listening tests in the M2H environment.

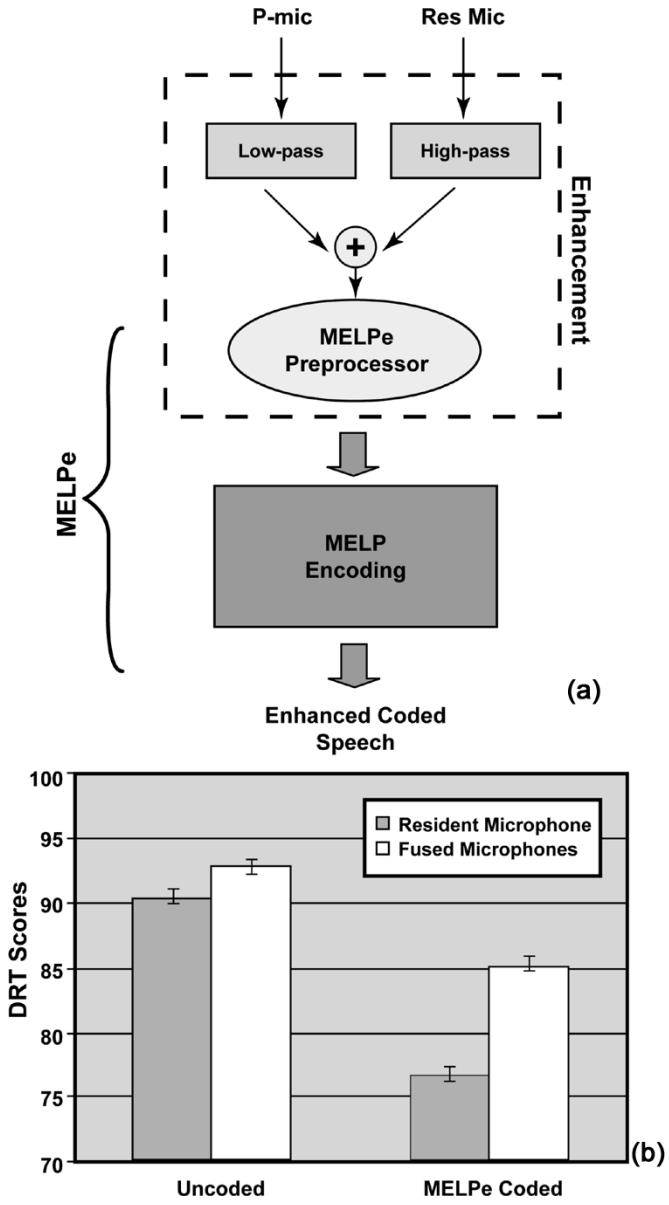

Fig. 5. Fusion with the low-frequency P-mic signal. (a) System configuration and (b) DRT performance with fusion only (no noise preprocessing and no encoding) and with fusion and MELPe preprocessing and encoding. For reference, DRT results are also provided for the resident microphone output with and without MELPe encoding. Results are for the M2 high-noise field.

Fig. 5(b) gives the DRT performance of the integrated system in Fig. 5(a) with and without fusion and with and without application of MELPe coding. In this test, three male and three female speakers from the ASE Pilot Speech Corpus, referred to as Talker Set #1, were used. The sensor fusion for the M2H noise environment results in significant DRT intelligibility improvements for uncoded (∼2.5 DRT points) and MELPe coded speech (∼8.1 DRT points). An example of fusion of the low-pass P-mic and high-pass resident-mic output is shown in Fig. 6, illustrating the improvement in spectral structure introduced by the fusion preprocessor. Specifically, low-frequency voicing including voice bars in the voiced fricative /z/ and voiced plosive /d/ in the word "zed", lost by the resident mic, have been approximately recovered through the low-pass P-mic signal. With coding, pitch, as well as voicing, improve overall due to the strong harmonicity in the low-pass P-mic, as we will see in a following example. Low-pass harmonicity is required by the MELP encoder for accurate pitch estimation which effects the coded signal over its full bandwidth.

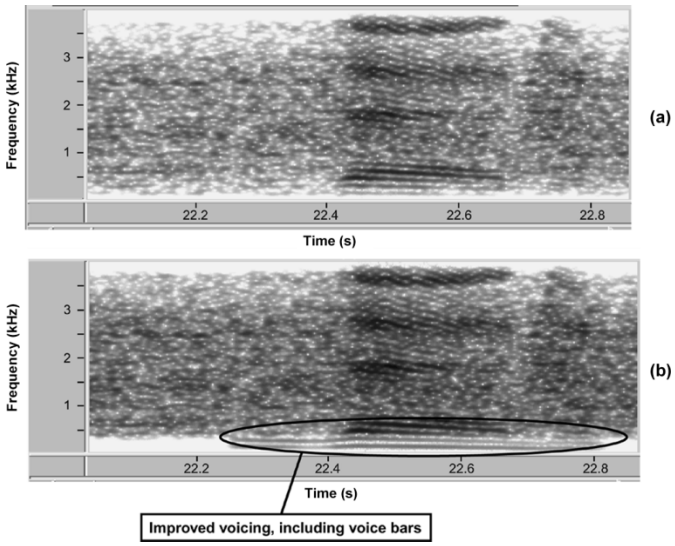

Fig. 6. Example spectrograms of fusion with the low-pass P-mic signal for the DRT word "zed". (a) Original noisy acoustic signal and (b) fusion with low-pass P-mic. Signals are from the M2 high-noise environment.

With regard to quality, through informal listening, we have found that inclusion of low-frequency content of the low-pass P-mic signal improves what listeners refer to as "fullness" in the uncoded and coded signals. Listeners also perceive a reduction in impulsive-like artifacts that arise due to abrupt pitch errors in the coded speech.

C. Pitch and Voicing Substitution

A multisensor MELPe architecture has been implemented with pitch and voicing substitution, as shown in Fig. 7(a) [1], explicitly addressing the loss of these speech features in severe acoustic noise environments. The architecture uses the resident acoustical microphone, GEMS, and P-mic sensors. The latter two sensors are used to estimate MELPe pitch and voicing parameters, respectively, that can be substituted into the MELPe encoding of the resident acoustical microphone channel. The P-mic is used for encoding the bandpass voicing in the two lowest bandpass regions, 0–500 Hz and 500–1000 Hz. Since the pitch is calculated multiple times in the MELPe architecture [22], it is necessary to substitute these intermediate pitch estimates into the encoding of the resident acoustical microphone channel. These intermediate pitches include the initial integer pitch estimate, the fractional pitch obtained during bandpass voicing calculations, and the final smoothed pitch.

The intelligibility impact of GEMS and P-mic pitch and voicing substitution was formally tested for coded speech. Fig. 7(b) gives the DRT performance for Talker Set #1 of the integrated system in the M2H and BHH environments. Parameter substitution results in DRT intelligibility improvements for the BHH (∼1.5 DRT points) and for the M2H (∼3.5 DRT points) environments. Example spectrograms of the MELPe and multisensor-enhanced MELPe outputs are shown in Fig. 8, illustrating the effect of improved pitch and voicing for a waveform in the M2 high-noise environment. With respect to quality, informal listening has determined that more robust

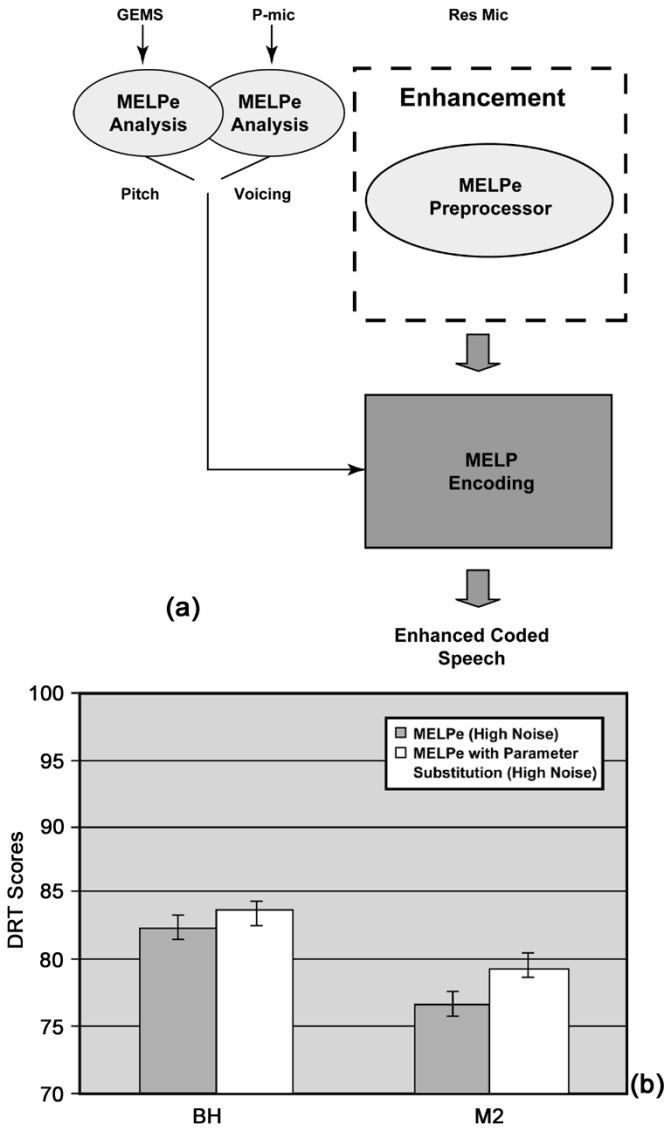

Fig. 7. Enhanced encoding with nonacoustic pitch and voicing substitution: (a) system configuration and (b) DRT performance with and without parameter substitution. Results are shown for the BH and M2 high-noise fields.

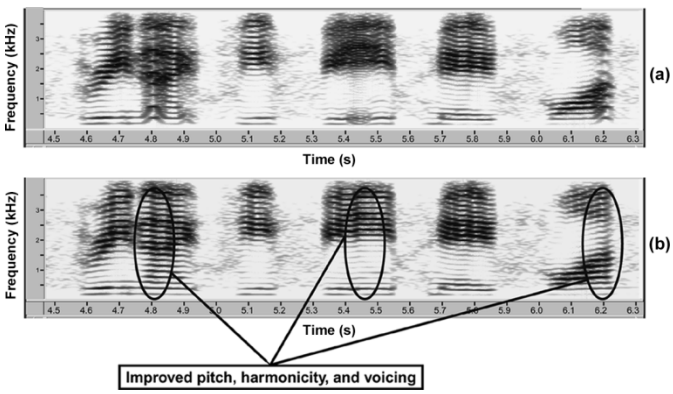

Fig. 8. Example spectrograms of MELPe output (a) without and (b) with GEMS pitch and P-mic voicing substitution for the phrase "three nineteen B page one."

pitch and voicing has reduced impulsive-like artifacts and excessive roughness due to abrupt pitch and voicing errors in the coded speech.

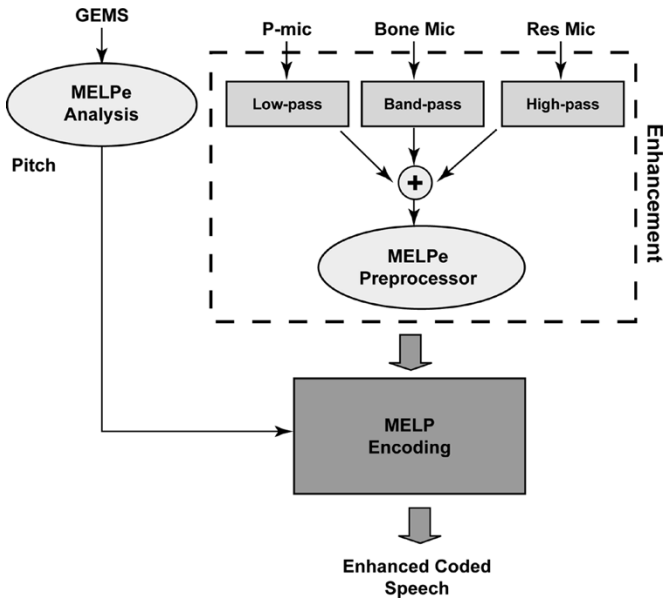

Fig. 9. Integrated multisensor MELPe with pitch substitution and fusion with the low-pass P-mic, bandpass bone-conduction mic, and high-pass resident mic. The bone-conduction mic output is available only in the MOUT noise condition.

D. Composite System

We have developed a system that merges both the fusion and pitch and voicing substitution methodologies. The integration is based on a two-step process illustrated in Fig. 9. In the first step, for the M2 and BH environments, a low-frequencyband 0–300 Hz and high-frequency band 300–4000 Hz fusion occurs between the P-mic and resident mic, respectively. For the MOUT environment, additional fusion with the bone-conduction mic output is performed in the 300–700 Hz interval due to robust vocal tract spectral content in this region. In the second step, the GEMS pitch is substituted for the fused signal's pitch during the encoding process. Because we are introducing a high-quality low-frequency signal, we found voicing substitution to be redundant, providing no DRT gain. The DRT results of this system are shown in Fig. 10(a) for the above Talker Set #1. This same system was also DRT-tested using a different set of three male and three female talkers from the ASE Pilot Corpus, referred to as Talker Set #2, and is shown in Fig. 10(b).

Fig. 10 shows that we have made significant DRT gains in all three environments over the MELPe coder for the average across male and female talkers. We also observe a large talker-set dependence in our results for the M2H and BHH environments relative to the baseline performance in the low-noise condition that is also shown in Fig. 10. Although our absolute gains in the M2H and BHH environments are comparable across talker sets, the gap from the low-noise condition is much larger in Talker Set #2. Example spectrograms of the MELPe and multisensor MELPe outputs are shown in Fig. 11, illustrating the effect of pitch substitution and fusion for a waveform in the M2 high-noise environment. Informal listening for quality gains give a composite of perceived improvements obtained through the system components introduced in Sections IV-A and IV-B: increased low-frequency fullness and reduced pitch and voicing artifacts. The DRT test results we have seen thus far give a composite score representing a broad view of a variety of speech characteristics. The DRT test also provides subscores for six

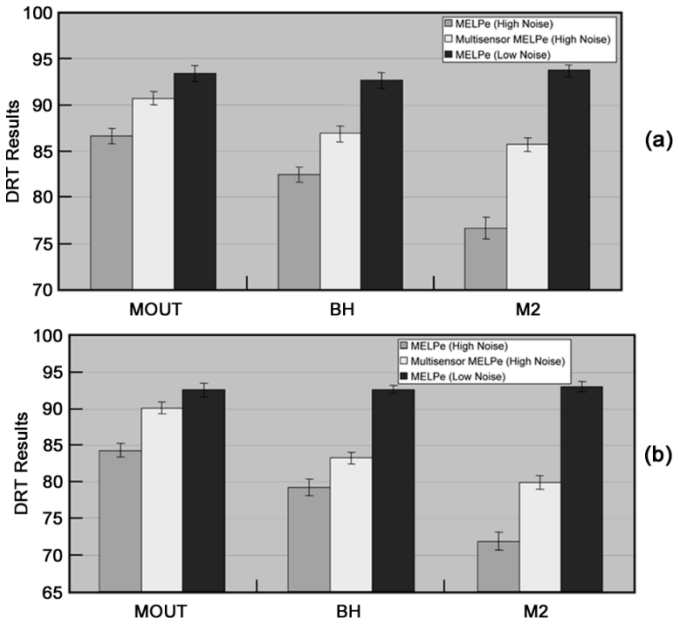

Fig. 10. DRT results for (a) Talker Set #1 and (b) Talker Set #2 using the multisensor MELPe of Fig. 9. Note the vertical scale difference in the two bar charts. DRT results for the multisensor MELPe are for the MOUT, BH, and M2 high-noise fields. MELPe in high- and low-noise conditions are also provided for reference.

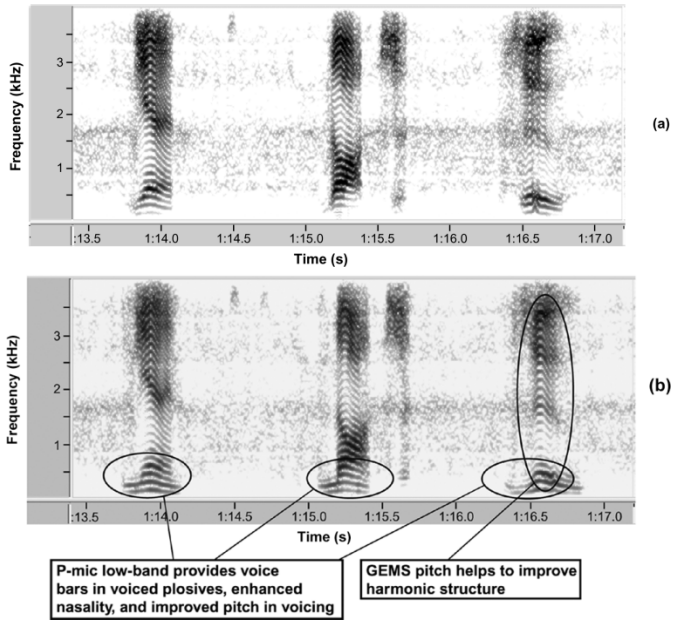

Fig. 11. Example spectrograms of MELPe output (a) without and (b) with GEMS pitch substitution and P-mic fusion, i.e., the multisensor MELPe of Fig. 9, for the DRT word sequence "yen, not, zoo". Example is for the M2 high-noise environment.

binary attributes of speech in their present and absent states: voicing, nasality, sustention, sibilation, graveness, and compactness [21]. The scores can be considered as measures of qualitative differences in intelligibility that can be associated with acoustic correlates. In analyzing specific DRT attributes, we have found that strong DRT attribute improvements are evident for voicing and nasality. This corresponds to the low-frequency and harmonic content of the P-mic and GEMS, as expected. The

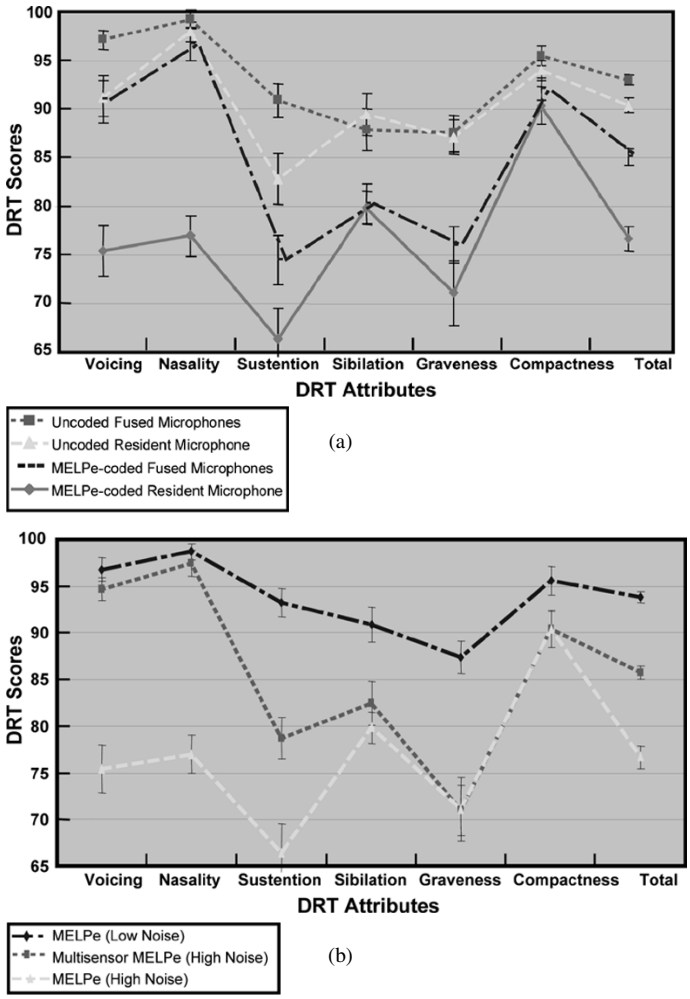

Fig. 12. DRT attribute results for (a) the fusion-only system of Section IV-B and (b) the fully composite system of Section IV-D. Results are for the M2 high-noise environment. In panel (b), the MELPe low-noise results are also provided for reference.

DRT attribute analysis also reveals where weakness and potential improvements lie. The DRT attribute analysis is shown in Fig. 12 for our fusion-only experiments (Section IV-B) and our fully composite system (Section IV-D) in the M2 environment.

V. High-Frequency Fusion

In another fusion approach, we use the wide-band signal 0–8000 Hz from the resident mic in the ASE Pilot Speech Corpus. Our experiments indicate that high frequency (>4 kHz) speech content provides significant intelligibility content. As alluded to this paper's introduction, this finding is consistent with the work of Kang and Everett [7] who showed that spectral mirror-imaging near the 4000-Hz bandwidth cutoff can give added intelligibility in an early LPC coder. Likewise, McCree has demonstrated quality qains with high-band parametric coding at 14 kbps [11].

In our first experiment with high-frequency content, MELPe-coded speech was augmented with high-frequency (4–8 kHz) uncoded speech obtained by attenuating the original noisy speech by 100 dB in the 0–4 kHz band by a high-pass filter. Our DRT attribute results for this experiment are shown in Fig. 13 for Talker Set #1 in the M2H environment and illustrate

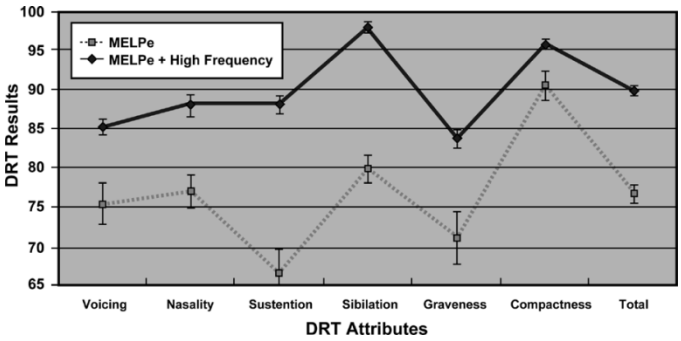

Fig. 13. DRT attribute results for high-frequency fusion-only.

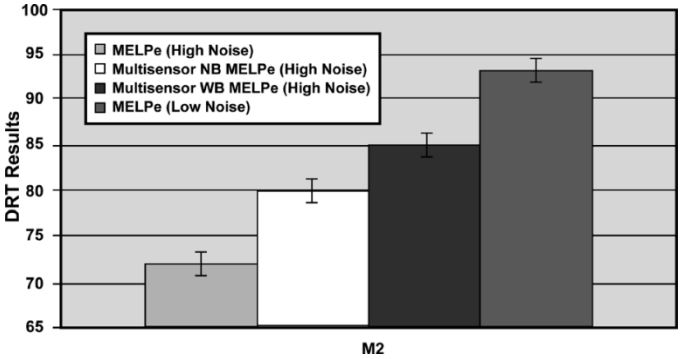

Fig. 14. DRT results for Talker Set #2 with high-frequency fusion (WB=wideband) compared against our previous system with only low-frequency fusion (NB=narrowband). Results are for the M2 high-noise field.

a significant gain in the intelligibility of the MELPe-coded speech when it is augmented with the uncoded high-frequency information. Intelligibility improvements are seen across all DRT attributes, particularly in sustension and sibilation.

We have developed a high-frequency coder based on this approach that has been DRT-intelligibility tested. This high-frequency approach takes advantage of two observations:

- Some ASE Pilot Corpus noise fields roll off significantly by about 4500 Hz.
- The human ear has less sensitivity at higher frequencies, allowing for low resolution (low bit-rate) coding.

The new coded high-frequency system uses the composite system of Section IV-D for the low band 0–4000 kHz, to which is added a separately coded high-frequency signal (roughly 4–8 kHz). The high-frequency channel uses a linear predictive all-pole model with a *fixed* set of four autoregressive coefficients, providing two poles at about 5500 and 6500 Hz. This eliminates the need for coding the spectral content. The high-frequency speech is assumed to be unvoiced, corresponding to a white-noise input to the all-pole filter. The only information that is coded is a six-level (2.6 bits) gain for the noise, resulting in only 116 bits/frame addition for a 2516-bps coder. The gain is selected to force the energy in the filter output to equal the energy in the original high-frequency signal over a MELP-analysis frame duration.

Significant intelligibility performance has been achieved for this high-frequency system (85.09) versus the baseline system (79.93) for Talker Set #2, shown in Fig. 14 for the M2H environment. This 85.09 score is comparable to the intelligibility of the original uncoded speech (85.92). This performance improvement is across all attributes of DRT intelligibility, though a remarkably impressive improvement has been made in the sibilation attribute (high-frequency content) where all tested systems have fallen short, to date. With respect to quality, informal listening shows a marked increase in what listeners describe as "crispness" and "articulation" of the uncoded and coded synthesized speech with the addition of high-frequency content.

A caveat to this approach is that for high-frequency fusion to be effective, the background acoustic noise field should roll off in the high-frequency region, unless noise in this frequency region can be effectively reduced. For example, the Blackhawk high-noise environment may not be a good candidate for this algorithmic approach because it does not roll off significantly above 4500 Hz.

VI. Summary and Future Directions

In this paper, we presented an approach to speech encoding that capitalizes on recent developments in nonacoustic sensors that are relatively immune to acoustic background noise. The GEMS, P-mic, and bone-conduction nonacoustic sensors were considered. These sensors can directly measure an approximate speech glottal excitation and reveal source activity, such as voice bars and glottalization, but also approximately measure some aspects of vocal tract activity, including low-energy, low-frequency events such as nasality.

Different aspects of acoustic and nonacoustic signals were used according to their capability in representing specific speech characteristics in different frequency bands. Preliminary testing involved the time-synchronous multisensor DARPA ASE Pilot Speech Corpus collected in harsh acoustic noise environments. A variety of enhanced encoding strategies were introduced and evaluated with DRT measures that show significant intelligibility gains.

The multisensor strategies and results of this paper point to numerous refinements and new directions, including the following:

High-Frequency Fusion: One of the previously mentioned challenges of the high-frequency coding approach is the requirement for extra bits for coding. We are currently considering approaches for reallocating bits from elsewhere in the codebook, as well as improvements to this high-frequency model, including a voiced model for the high-frequency content and adding subframe onsets. Also under investigation is a plosive model suggested by Unno *et al.* [19]. Plosive modeling of this type provides more accurate temporal structure in the coded high-frequency speech component, and thus should increase intelligibility scores in the DRT attributes of sustention and graveness.
Sensor Phase: Frequency-domain sensor phase, as well as magnitude, was found to contribute to improved signal enhancement and thus improved encoding within different frequency bands [15]. This is a promising area of research.
Extension to Ultra-Low Rate Coding: We are considering new low-rate coding paradigms involving extensions of our multiband fusion approach and speaker-dependent source characterization that exploit nonacoustic sensor outputs in high-noise environments [3].

ACKNOWLEDGMENT

The authors acknowledge discussions with researchers at Rutgers, Georgia Tech, Aliph, Sarnoff, and BBN who have participated in the DARPA ASE program. Thanks also goes to John Collura for providing the draft NATO standardization agreement (STANAG) 4591 coder.

REFERENCES

[1] K. Brady, T. F. Quatieri, W. B. Campbell, J. P. Campbell, M. Brandstein, and C. J. Weinstein, "Multisensor MELPe using parameter substitution," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, Montreal, Canada, May 2004, pp. 477–480.

[2] G. C. Burnett, J. F. Holzrichter, T. J. Gable, and L. C. Ng, "The use of glottal electromagnetic micropower sensors (GEMS) in determining a voiced excitation function," in *Proc. 138th Meeting of the Acoustical Society of America*, Columbus, OH, Nov. 1999.

[3] W. M. Campbell, T. F. Quatieri, J. P. Campbell, and C. J. Weinstein, "Multimodal speaker authentication using nonacoustic sensors," in *Proc. Workshop Multimodal User Authentication*, Santa Barbara, CA, Dec. 2003, pp. 215–222.

[4] M. F. Cohen, "Effects of stimulus presentation rate upon intelligibility-test scores," *J. Acoust. Soc. Amer.*, vol. 37, 1965.

[5] J. F. Holzrichter, G. C. Burnett, L. C. Ng, and W. A. Lea, "Speech articulator measurements using low-power EM-wave sensor," *J. Acoust. Soc. Amer.*, vol. 103, no. 1, p. 622, 1998.

[6] J. F. Holzrichter, L. C. Ng, G. J. Burke, N. J. Champagne II, J. S. Kallman, R. M. Sharpe, J. B. Kobler, and R. E. Hillman, "Measurements of glottal structure dynamics," Lawrence Livermore National Laboratory, Univ. California Rep. UCRL-JRNL-14 775, 2003.

[7] G. S. Kang and S. S. Everett, "Improving LPC analysis," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, Boston, MA, Apr. 1983, pp. 89–92.

[8] R. Martin, I. Wittke, and P. Jax, "Optimized estimation of spectral parameters for the coding of noisy speech," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, Istanbul, Turkey, Jun. 2000, pp. 1479–1482.

[9] R. Martin, D. Malah, R. V. Cox, and A. J. Accardi, "A noise reduction preprocessor for mobile voice communication," *EURASIP J. Appl. Signal Process.*, pp. 1046–1058, 2004.

[10] A. McCree, K. Truong, E. B. George, T. P. Barnwell, and V. Viswanathan, "A 2.4 kbit/s MELP coder candidate for the new US Federal standard," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, Atlanta, GA, May 1996, pp. 200–203.

[11] A. McCree, "14 kbps wide-band speech coder with a parametric high-band model," in *Proc. Int. Conf. Acoustics, Speech, and Signal, Processing*, Istanbul, Turkey, Jun. 2000, pp. 761–764.

[12] D. Messing, "Noise suppression using spectral magnitude phase from nonair-acoustic sensors," M.S. thesis, MIT, Aug. 2003.

[13] L. C. Ng, G. C. Burnett, J. F. Holzrichter, and T. J. Gable, "Denoising of human speech using combined acoustic and EM sensor signal processing," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, Istanbul, Turkey, June 2000, pp. 1914–1917.

[14] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*. Englewood Cliffs, NJ: Prentice-Hall, 2002.

[15] 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," in *Proc. Workshop on Multimodal User Authentication*, Santa Barbara, CA, Dec. 2003, pp. 66–73.

[16] M. Rothenberg, "A multichannel electroglottograph," *J. Voice*, vol. 6, no. 1, pp. 36–43, 1992.

[17] M. V. Scanlon, "Acoustic sensor for health status monitoring," in *Proc. IRIS Acoustic Seismic Sensing*, vol. 2, 1998, pp. 205–222.

[18] L. M. Supplee, R. P. Cohn, J. S. Collura, and A. V. McCree, "MELP: the new federal standard at 2400 bps," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, Munich, Germany, Apr. 1997, pp. 1591–1594.

[19] T. Unno, T. P. Barnwell III, and K. Truong, "An improved mixed excitation linear prediction (MELP) coder," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, Phoenix, AZ, May 1999, pp. 1419–1422.

[20] V. R. Viswanathan, K. F. Karnofsky, K. N. Stevens, and M. N. Alakel, "Multisensor speech input for enhanced immunity to acoustic background noise," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, Mar. 1984, pp. 57–60.

[21] W. D. Voiers, "Diagnostic evaluation of speech intelligibility," in *Benchmark Papers in Acoustics*, M. Hawley, Ed. Stroudsburg: Dowden, Hutchinson and Ross, 1977, vol. 11, Speech Intelligibility and Speaker Recognition.

[22] T. Wang, K. Koishida, V. Cuperman, A. Gersho, and J. S. Collura, "A 1200/2400 BPS coding suite based on MELP," in *NATO AC/322(SC/6-AHWG/3), AD HOC Working Group on Narrow Band Voice Coding, Proc. 2002 IEEE Workshop on Speech Coding, Special Session 1: Topics on NATO Standardization*, Tsukuba, Japan, Oct. 2002.

[23] T. Yanagisawa and K. Furihata, "Pickup of speech signal utilization of vibration transducer under high ambient noise," *J. Acoust. Soc. Jpn.*, vol. 31, no. 3, pp. 213–220, 1975.

Thomas F. Quatieri (S'78–M'79–SM'87–F'99) received the B.S. degree (summa cum laude) from Tufts University, Medford, MA, in 1973, and the S.M., E.E., and Sc.D. degrees from the Massachusetts Institute of Technology (MIT), Cambridge, in 1975, 1977, and 1979, respectively.

He is currently a Senior Member of the Technical Staff at MIT Lincoln Laboratory, Lexington, and a member of the faculty of the Speech and Hearing Bioscience and Technology Program at MIT. In 1980, he joined the Sensor Processing Technology Group of MIT, Lincoln Laboratory where he worked on problems in multidimensional digital signal processing and image processing. Since 1983, he has been a member of the Information Systems Technology Group at MIT Lincoln Laboratory where he has been involved in digital signal processing for speech and audio applications, underwater sound enhancement, and data communications. His current focus is in the areas of speech and audio enhancement and modification, modeling of signal processing in the auditory system, and robust speaker recognition. He has contributed many publications to journals and conference proceedings, written several patents, and co-authored chapters in numerous edited books. At MIT he developed the graduate course Digital Speech Processing, and is active in advising graduate students on the MIT campus. Based on his MIT course, he has written the text book *Discrete-Time Speech Signal Processing: Principles and Practice* (Englewood Cliffs, NJ: Prentice-Hall, 2001).

Dr. Quatieri is the recipient of four IEEE paper awards: The 1982 Paper Award of the IEEE Acoustics, Speech and Signal Processing Society, both the 1990 and 1994 Senior Award of the IEEE Signal Processing Society, and the 1995 IEEE W.R.G. Baker Prize Award. He was a member of the IEEE Digital Signal Processing Technical Committee and the IEEE Speech Technical Committee, from 1983 to 1992 served on the steering committee for the bi-annual Digital Signal Processing Workshop, and was Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING in the area of nonlinear systems. He is a member of Tau Beta Pi, Eta Kappa Nu, Sigma Xi, and the Acoustical Society of America.

Kevin Brady received the B.S. and M.S. degrees in computer engineering from Clarkson University, Potsdam, NY, and the M.S. and D.Sc. degrees in systems engineering from Washington University, St. Louis, MO, in 1990, 1992, 1994, and 1997, respectively.

He is currently a member of the Technical Staff in the Information Systems Technology Group, Lincoln Laboratory, Massachusetts Institute of Technology (MIT), Lexington, working on speech coding and recognition problems. From 1997 to 2002, he was a member of the System Engineering and Analysis Group at MIT Lincoln Laboratory working on estimation and discrimination problems.

Dave Messing was born in Redwood City, CA. He attended undergrade studies at the University of California, Berkeley, majoring in EECS, graduating with honors and a B.S. degree in 2001. He began graduate school at MIT in 2001 and received his M.S. degree in EECS in 2003 working under Tom Quatieri of MIT Lincoln Laboratory. He is currently pursuing the Ph.D in EECS at the RLE Sensory Communications Group and working under Louis Braida.

His main interests are signal processing, communications, and speech., and his current work delves into human auditory modeling, focusing on nonlinear signal processing issues related to noise.

Joseph P. Campbell (S'90–M'92–SM'97–F'05) received the B.S., M.S., and Ph.D. degrees in electrical engineering from Rensselaer Polytechnic Institute, Troy, NY, The Johns Hopkins University, Baltimore, MD, and Oklahoma State University, Stillwater, in 1979, 1986, and 1992, respectively.

He is currently a Senior Member of the Technical Staff at MIT Lincoln Laboratory in the Information Systems Technology Group, where he conducts speech-processing research and specializes in advanced speaker recognition methods. His current foci are high-level features for and forensic-style applications of speaker recognition, creating corpora to support speech processing research and evaluation, robust speech coding, biometrics, and cognitive radio. Before joining Lincoln, he served 22 years at the National Security Agency (NSA). From 1979 to 1990, he was a member of NSA's Narrow-band Secure Voice Technology research group. Joe and his teammates developed the first DSP-chip software modem and LPC-10e, which enhanced the Federal Standard 1015 voice coder and improved US and NATO secure voice systems. He was the Principal Investigator and led the U.S. Government's speech coding team in developing the CELP voice coder, which became Federal Standard 1016 and is the foundation of digital cellular and voice over the Internet telephony systems. From 1991 to 1998, he was a Senior Scientist in NSA's Biometric Technology research group, where he led voice verification research. From 1994 to 1998, Joe chaired the Biometric Consortium, the US Government's focal point for research, development, test, evaluation, and application of biometric-based personal identification and verification technology. From 1998 to 2001, he led the Acoustics Section of NSA's Speech Research branch, conducting and coordinating research on and evaluation of speaker recognition, language identification, gender identification, and speech activity detection methods.

Dr. Campbell was an Associate Editor of the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING from 1991 to 1999. He was an IEEE Signal Processing Society Distinguished Lecturer in 2001. He is currently a member of the IEEE Signal Processing Society's Board of Governors; an Editor of Digital Signal Processing journal; a Chair of the International Speech Communication Association's Speaker and Language Characterization Special Interest Group (ISCA SpLC SIG); a member of ISCA, Sigma Xi, and the Acoustical Society of America.

William M. Campbell received the B.S. degrees in electrical engineering, computer science, and mathematics from the South Dakota School of Mines and Technology in 1990, the M.S. degree in applied mathematics from Cornell University in 1993, and the Ph.D. degree in applied mathematics in 1995 from Cornell University (under a NSF fellowship).

From 1995 to 1999, he was a Senior Research Scientist at the Motorola Space and Systems Technology Group (SSTG) in the Speech and Signal Processing Lab. While at Motorola SSTG, he did research in Biometrics, speech interfaces for wearable computing, and communications for the battlefield. He participated in the creation of numerous products including the Tenor Pager, the CipherVox speaker verification SDK, the Force XXI wearable computer voice control, and an adaptive-rate voice communication system. From 1999 to 2002, he was a principal research scientist in Motorola Labs where he worked on biometrics, machine learning, and telematics. Since 2002, he has been a Staff Member of the Information Systems Technology Group at Lincoln Laboratory, where he is involved in speaker recognition, digital signal processing, and machine learning for speech applications. He has contributed numerous publications to journals and conferences.

Dr. Campbell received the Motorola Distinguished Innovator award and holds 12 patents. He is a member of the American Mathematical Society, Tau Beta Pi, and Eta Kappa Nu.

Michael S. Brandstein (M'96–SM'03) was born in Willimantic, CT in 1967. He received the Sc.B. degree from Brown University, Providence, RI, in 1988, the S.M. degree from Massachusetts Institute of Technology (MIT), Cambridge, MA, in 1990, and the Ph.D. degree from Brown University in 1995, all in electrical engineering.

He is currently a Member of the Research Staff at MIT Lincoln Laboratory, Lexington, MA, involved in signal processing for speech and multimedia applications. He was an Assistant and Associate Professor of electrical engineering at Harvard University, Cambridge, from 1996 to 2003. His work featured developments in source localization, microphone array design, time-delay estimation, multichannel speech enhancement, and acoustic and visual-based tracking.

Dr. Brandstein is a former Chairman of the IEEE Signal Processing Society's Audio and Electroacoustics Technical Committee and a past member of the Multimedia Signal Processing Technical Committee and the Technical Directions Committee. He has been an Associate Editor for the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING.

Clifford J. Weinstein (S'66–M'69–SM'84–F'94) received the S.B., S.M., and Ph.D. degrees in electrical engineering from the Massachusetts Institute of Technology, Cambridge.

He leads the Information Systems Technology Group at MIT Lincoln Laboratory and is responsible for initiating and managing research programs in speech technology, machine translation, and information assurance. He joined Lincoln Laboratory as an MIT graduate student in 1967, and became Leader of the Speech Systems Technology group (now Information Systems Technology group) in 1979. He has made technical contributions and carried out leadership roles in research programs in speech recognition, speech coding, machine translation, speech enhancement, packet speech communications, information system assurance and survivability, integrated voice/data communication networks, digital signal processing, and radar signal processing. He has published numerous papers in these areas, six of which were selected for reprint in IEEE Press books.

Dr. Weinstein was elected to the Board of Governors of the IEEE Signal Processing Society in 1993. From 1991 to 1993, he was chairman of the IEEE Signal Processing Society's Technical Committee on Speech Processing. In 1976–1978, he was chairman of that Society's Technical Committee on Digital Signal Processing. In 1993, he was elected as a Fellow of the IEEE for technical leadership in speech recognition, packet speech, and integrated voice/data networks. From 1986 to 1998, he was U.S. technical specialist on the NATO RSG10 (now IST-01) Speech Research Group, in which capacity he authored a comprehensive NATO report and journal article on opportunities for applications of advanced speech technology in military systems. From 1989 to 1994, he was chairman of the coordinating committee for the DARPA Spoken Language Systems Program, which was the major U.S. research program in speech recognition and understanding, and which involved coordinated efforts of a number of leading U.S. research groups. From 1999 to 2003, he was a member of the DARPA Information Sciences and Technology (ISAT) Panel, which provides DARPA with continuing assessments of the state of advanced information science and technology; as an ISAT member, he co-chaired ISAT studies in 2001 and in 2003.

John D. Tardelli received the B.S and M.S. degrees in physics from Lowell Technological Institute, Lowell, MA, in 1964 and 1970, respectively.

He joined ARCON Corporation in 1972 and is currently the Director of Digital Speech Processing. He has over 40 years of research experience in applied physics and mathematics. His efforts include the application of digital signal processing techniques to bandwidth compression and synthesis of speech, the test and evaluation of digital voice communication systems using subjective listener procedures, analysis of additive acoustic noise and its effects on speech and speech processing algorithms; and the development of T&E procedures for multispeaker conferencing systems. In addition, he works closely with government and industry voice coder standardization groups.

Paul D. Gatewood received the B.M. degree with a specialization in audio recording from the Berklee College of Music in 1983 and the M.S. degree in computer science from Boston University, Boston, MA, in 1998.

Since 1983, he has been with the ARCON Corporation Digital Speech Processing Division as an Audio Technician and currently as a Senior Scientist working in the fields of voice communications, audio recording and software development. He has supported all major test and evaluation efforts including the testing that led to the selection of MELP as the military standard voice coder at 2.4 kbps in 1995. His efforts have included the generation of input source material for subjective evaluation of narrow-band voice coders, the recording, characterization and simulation of acoustic noise environments, the implementation and interfacing of real time channel simulations and digital voice coders, and the design and implementation of subjective test procedures for speech intelligibility, quality, recognizability, and communicability.