A Comparative Study of Traditional and Newly Proposed Features for Recognition of Speech Under Stress

Sahar E. Bou-Ghazale, *Member, IEEE,* and John H. L. Hansen, *Senior Member, IEEE*

***Abstract*—It is well known that the performance of speech recognition algorithms degrade in the presence of adverse environments where a speaker is under stress, emotion, or Lombard effect. This study evaluates the effectiveness of traditional features in recognition of speech under stress and formulates new features which are shown to improve stressed speech recognition. The focus is on formulating robust features which are less dependent on the speaking conditions rather than applying compensation or adaptation techniques. The stressed speaking styles considered are simulated angry and loud, Lombard effect speech, and noisy actual stressed speech from the SUSAS database which is available on CD-ROM through the NATO IST/TG-01 research group and LDC[1] . In addition, this study investigates the immunity of linear prediction power spectrum and fast Fourier transform power spectrum to the presence of stress. Our results show that unlike fast Fourier transform's (FFT) immunity to noise, the linear prediction power spectrum is more immune than FFT to stress as well as to a combination of a noisy and stressful environment. Finally, the effect of various parameter processing such as fixed versus variable preemphasis, liftering, and fixed versus cepstral mean normalization are studied. Two alternative frequency partitioning methods are proposed and compared with traditional mel-frequency cepstral coefficients (MFCC) features for stressed speech recognition. It is shown that the alternate filterbank frequency partitions are more effective for recognition of speech under both simulated and actual stressed conditions.**

***Index Terms*—Linear prediction, Lombard effect, speech recognition, speech under stress.**

I. Introduction

IT is well known that the performance of speech recognition systems degrade under the presence of stress [2], [4]–[6], [8], [20]. Stress in this context refers to speech produced under environmental, emotional, or workload stress. The stress conditions considered in this study include simulated angry and loud, Lombard effect conditions, and actual stressed speech all obtained from the SUSAS (Speech Under Simulated and Actual Stress) [9] database. The stress condition referred to as Lombard effect results when a speaker attempts to modify his or her speech production system while speaking in a noisy environment [13], [20]. To improve the performance of speech recognition algorithms under stress, a number of methods have been considered. These fall into three general areas of 1) robust features, 2) stress equalization methods, and 3) model adjustment or training methods. Fig. 1 shows a general speech recognition scenario which considers a variety of speech/speaker distortions, and the three general approaches to robust speech recognition. For this scenario, we assume that a speaker is exposed to some adverse environment, where ambient noise is present and a stress induced task is required (or the speaker is experiencing emotional stress). The adverse environment could be a noisy automobile where cellular communication is used, high-stress noisy helicopter or aircraft cockpits, or other environments where hands-free operation is needed. Since the user task could be demanding, the speaker is required to divert a measured level of cognitive processing, leaving formulation of speech for recognition as a secondary task. Some speech recognition studies have adapted the recognizer to the input stressed speech during training [14], or compensated for the effect of stress during recognition testing phase (e.g., formant location and bandwidth stress equalization [6], [7], [21]; whole-word cepstral compensation [2]; slope-dependent weighting [15]; formant shifting [17]; source-generator based codebook stress compensation [16],

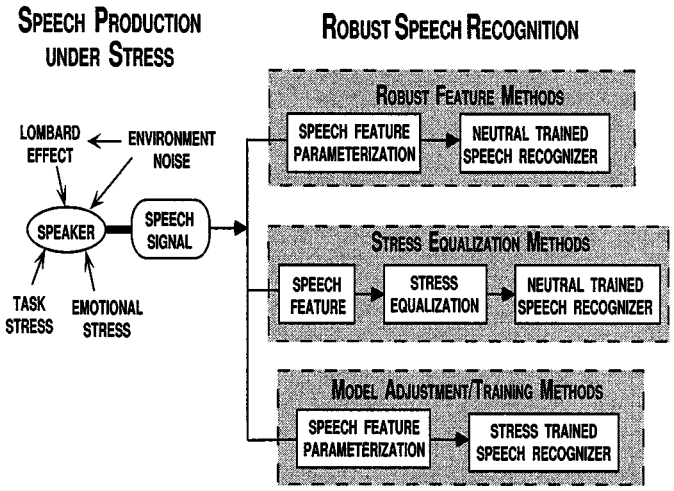


Fig. 1. Types of distortion which can be addressed for robust speech recognition.

Manuscript received November 3, 1997; revised June 21, 1999. This work was supported by a grant from the U.S. Air Force Research Laboratory, Rome, NY. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Wu Chou.

S. E. Bou-Ghazale was with Robust Speech Processing Laboratory, Center for Spoken Language Research, University of Colorado, Boulder, CO 80309-0258 USA. She is now with Network Access Division, Conexant Systems, Inc., Newport Beach, CA 92658-8902 USA.

J. H. L. Hansen is with Robust Speech Processing Laboratory, Center for Spoken Language Research, University of Colorado, Boulder, CO 80309-0258 USA (e-mail: jhlh@cslr.colorado.edu; http://cslr.colorado.edu/rspl/).

Publisher Item Identifier S 1063-6676(00)05331-1.

[1]http://morph.ldc.upenn.edu/Catalog/LDC99S78.html.

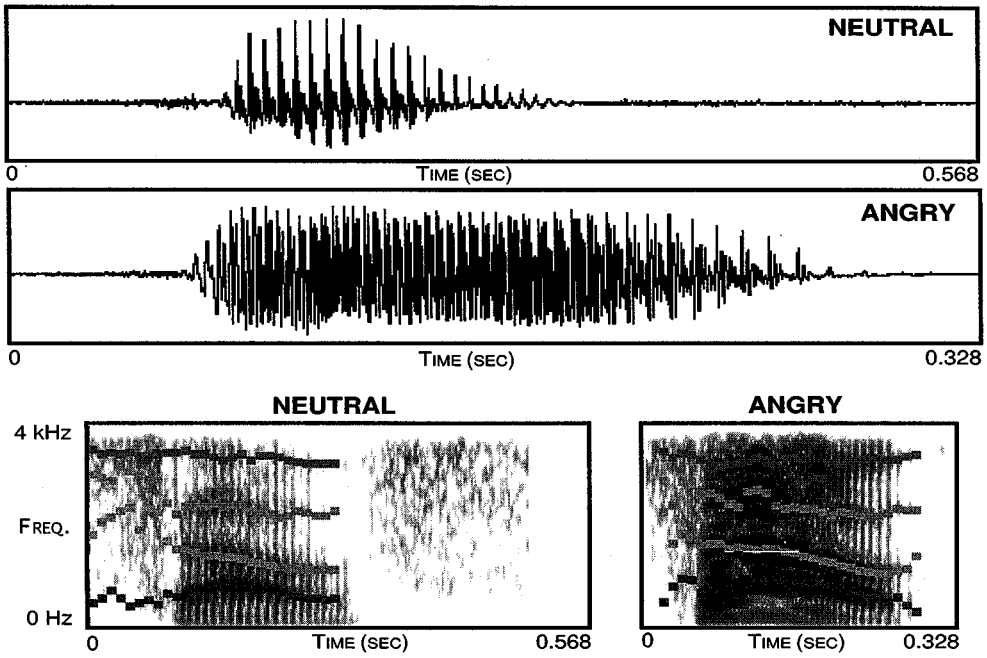


Fig. 2. Sample waveforms and speech spectrograms with formant overlays for the word *south* spoken by a male speaker in neutral and angry stressed speech conditions.

[21]; source-generator based adaptive cepstral compensation (MCE-ACC) [18], [21]). Others have also considered alternative training methods such as multi-style training [8]. However, multi-style training has been shown to be less effective than a neutral trained system when operating in an open speaker recognition task [19]. In general, however, these approaches normally require either a preprocessing stage or collection of stressed speech data for training in order to model the statistics of the input stressed speech and to incorporate this knowledge in the recognition system.

While these approaches have improved the recognition performance, they are restricted to the specific stressed condition or speaking style being addressed. A more general solution to the problem of speech recognition under stress would be to improve the signal modeling or parameterization stage as opposed to compensating the feature observation sequence for the effects of stress. The ultimate goal would be to achieve a signal modeling framework or robust features which are immune to the speech variations due to stress. The majority of previous studies on robust features, however, have focused on features which are robust to noise [22]–[25]. An extensive summary of speech recognition in noisy environments can be found in a recent survey [26]. Moreover, the majority of features employed in current speech recognition systems are based on the FFT power spectrum due to its reported immunity to noise [27]. There have been limited studies on robust features tailored to the effects of speech under *stress*. While research has been conducted on the analysis of speech under stress [4], [5], [34] a number of recent studies have begun to more extensively address the issue of changes in speech production under stress [10], [11], and their impact on speech systems [12], [1]. Depending on the type of stress, it is known that fundamental frequency, duration and intensity effects, glottal source, and vocal tract frequency structure are all effected in different ways. The significance of parameter variations with respect to neutral, as well as their relationship with other stressed conditions have also been addressed. For example, for speech under angry conditions, the distribution of fundamental frequency expands greatly (width of distribution more than doubles), the percentage of time spent in vowels and the corresponding amount of energy increases significantly at the expense of the percentage of time spent in consonants and consonant energy, the glottal spectral slope becomes more flat (more high frequency energy), and formant locations are almost always statistically different from neutral (and many formant bandwidths as well).

These factors clearly show that spectral structure is altered when speech is produced under stress. Therefore, the focus of this study is to evaluate the effectiveness of features which have been shown to be robust to frequency dependent noise for the purpose of recognition of stressed speech, and to propose alternative features that could improve stressed speech recognition. In addition, since the majority of previous speech recognition evaluations of power spectrum methods have targeted noisy environments, this study will also investigate the immunity of the FFT power spectrum to stress, and contrasts its performance to the linear prediction power spectrum. Finally, the effect of traditional parameter processing on the recognition performance is studied and new parameter processing techniques are proposed for further improving stressed speech recognition. A flow diagram of the stressed speech recognition framework is shown in Fig. 3. Sources of speech under high stress or adverse environments due to task demands, emotion, or Lombard effect are shown as effecting the input speech to the hidden Markov model (HMM) recognizer. A summary of the speech features considered in this study is shown on the left [(a) though (e)] of Fig. 3. Finally, we emphasize that neutral speech data is used for

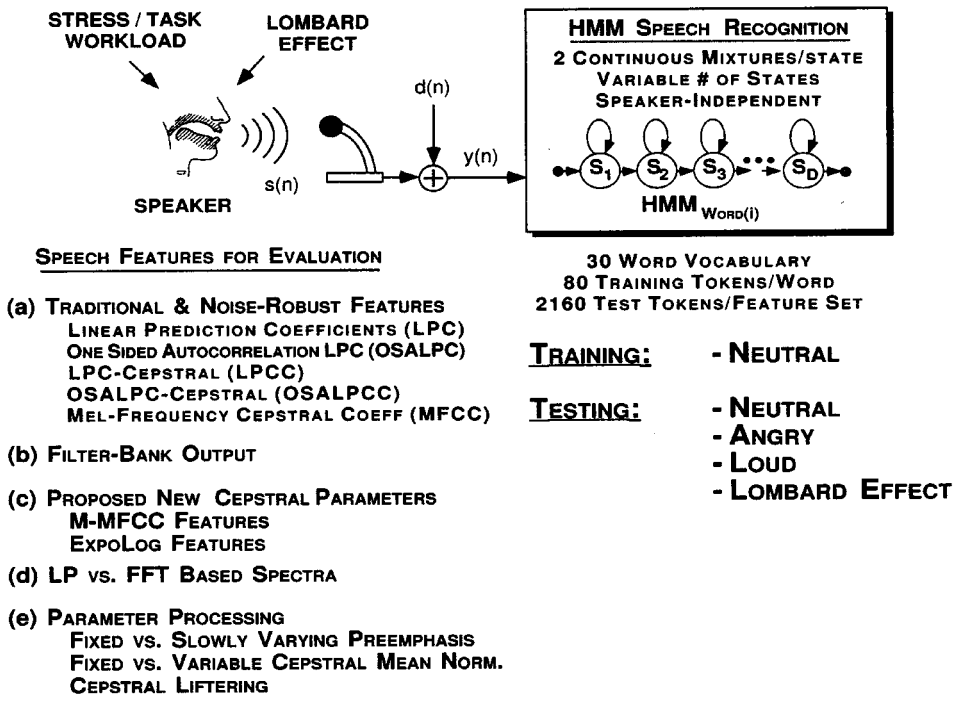


Fig. 3. A flow diagram of the speech recognition under stress framework. Sources of speech under high stress or adverse environments are shown as input to the HMM speech recognizer. A summary of the speech features considered in this study is shown on the left [(a)–(e)]. Finally, neutral speech data is used for all training evaluations (for each feature set). Round-robin open test evaluations are conducted for recognition of speech under angry, loud, Lombard effect, and actual stressed conditions.

all training evaluations (for each feature set). Round-robin open test evaluations are conducted for recognition of speech under various stressed conditions.

The remainder of this paper is organized as follows. The database employed in all of our evaluations is briefly summarized in Section II. In Section III, we study the effectiveness of traditional as well as noise robust features in recognition of speech under stress. In Section IV, we investigate the recognition performance across individual frequency bands to determine frequency regions less affected by stress. Based on these results, we propose in Section V two new frequency scales which are less sensitive to the effects of stress as compared to the traditional mel-frequency scale. In Section VI, we compare the performance of linear prediction and FFT power spectrum based features in the presence of stress. In Section VII, we summarize the effect of various parameter processing methods to recognition performance and propose new parameter processing approaches based on the variable impact of stress on speech classes for improving recognition performance. A final summary and a series of conclusions are drawn in Section VIII.

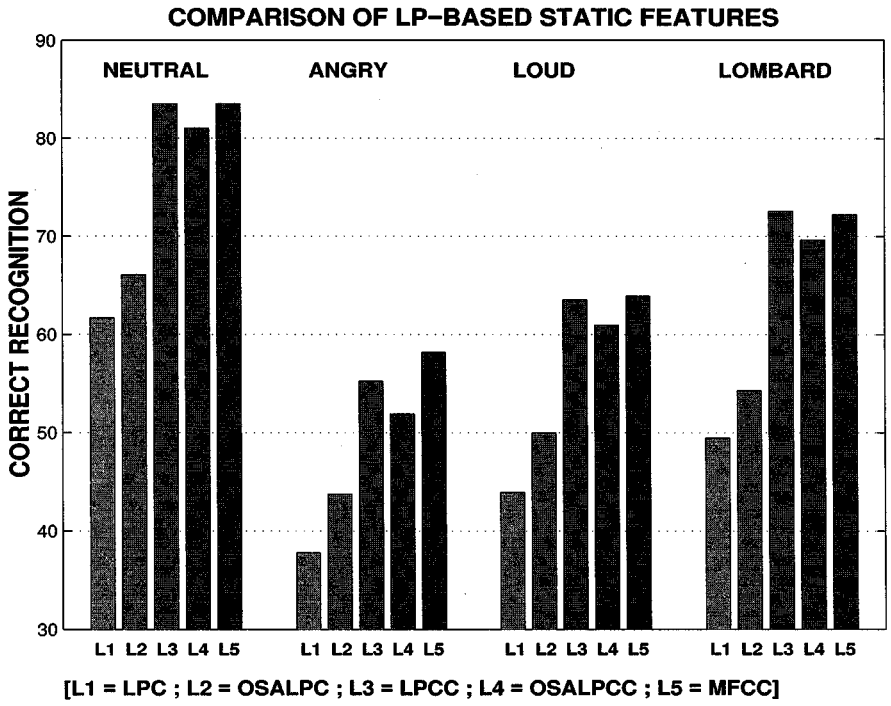
II. Recognizer and Database

The speech data employed in this study is a subset of the SUSAS[2] (Speech Under Simulated and Actual Stress) database [9]. Approximately half of the SUSAS database consists of styled data (such as *normal, angry, soft, loud, slow, fast, clear*) donated by the Lincoln Laboratory [8], and Lombard effect speech. Lombard effect speech was obtained by having speakers listen to 85 dB SPL pink noise through headphones while speaking (i.e., recordings are noise-free). A common vocabulary set of 35 aircraft communication words make up over 95% of the data base. These words consist of mono- and multisyllabic words which are highly confusable. Examples include /go-oh-no/, /wide-white/, and /six-fix/. Twelve tokens of each word in the vocabulary were spoken by nine native American speakers for the neutral conditions, and two tokens for each style condition. This database has been employed extensively in the study of how speech production and recognition varies when speaking under stressed speech conditions [21].

To illustrate an example of the effects of stress on speech, Fig. 2 shows time waveforms and speech spectrograms with formant overlays for the word “south” spoken by a male speaker in both neutral and angry stress conditions. For this example, overall word duration decreased by 42%, while the voiced duration is 218 ms for angry and 203 ms for neutral. The resulting spectral response is more intense for the voiced speech section, with a change in formant bandwidth (especially the second); while only slight changes in overall formant location tracks were noted. Finally, there is a clear change in overall spectral slope (note significantly higher energy levels in high frequency portion of spectrogram for the *angry* token).

In this study, all recognition evaluations are speaker-independent, and consider only male speakers. A 30-word HMM-based recognizer is formulated using a variable-state-size, left-to-right model, with two continuous mixtures per state. The HMM models are trained with neutral speech of eight speakers while a ninth speaker is left for open testing. A total of ten tokens per speaker are employed for training each neutral HMM word

[2]The SUSAS Stressed Speech Database from RSPL is available from the Linguistics Data Consortium at the following web location: http://morph.ldc.upenn.edu/Catalog/LDC99S78.html.



Recognition Performance of Neutral Trained Models Employing Static Features Derived from LP Power Spectrum

Training/Testing Features	*Speaking Styles Tested*				Overall Recognition
	NEUTRAL	ANGRY	LOUD	LOMBARD	
LPC	61.65%	37.78%	43.89%	49.44%	48.19%
OSALPC	66.06%	43.70%	50.00%	54.26%	53.51%
LPCC	83.53%	55.19%	63.52%	72.59%	68.71%
OSALPCC	81.06%	51.85%	60.93%	69.63%	65.87%
MFCC	83.52%	58.15%	63.89%	72.22%	69.45%

Fig. 4. Recognition performance of linear prediction power spectrum based static features in the presence of stress. The graph shows the recognition rates of neutral trained models tested with four speaking conditions for five different sets of features: LPC, OSALPC, LPCC, OSALPCC, and MFCC.

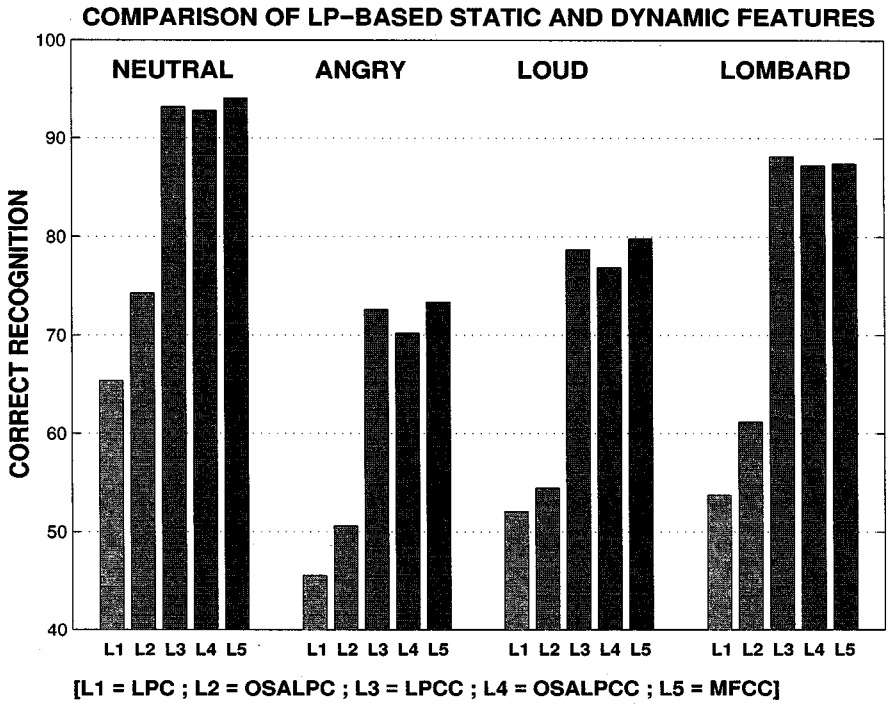
model resulting in 80 training tokens per word. The training and testing are done in a round robin scheme to allow all speakers and tokens to be tested in an open evaluation. In evaluating each of the neutral trained HMM models, a total of 2160 tokens are tested from the four speaking styles.

The last evaluation in this work employs actual stressed speech from the SUSAS database. The actual speech under stress data employed in this evaluation consists of speech produced during the completion of two types of subject motion-fear tasks. The speakers produced speech while participating in two amusement park rides (e.g., a traditional roller-coaster ride and a free-fall ride consisting of a 130 ft vertical drop machine). These two rides were chosen in an attempt to simulate the sudden change in altitude or direction which could be experienced in an aircraft cockpit under emergency conditions [9].

III. Performance of Traditional and Noise-Robust Features in Stress

The majority of previous research on robust features have addressed robustness in noisy or channel corrupted environments. The goal in this section is to investigate the effectiveness of previously proposed noise robust features in the recognition of stressed speech. The first set of features evaluated in this study are the one-sided autocorrelation linear prediction coefficients (OSALPC) which were shown to outperform linear prediction coefficients (LPC) as well as two other noise robust methods, the short-time modified coherence method (SMC) and the least-squares modified Yule Walker equations (LSMYWE), in severe noisy conditions [25]. The second set of features evaluated are the cepstral-based OSALPC, referred to as OSALPCC in the tables and figures, which are compared to the performance of traditional cepstral-based LPC and mel-frequency scale coefficients (MFCC). Therefore, the recognition performance of the following set of features will be compared 1) linear prediction coefficients (LPC), 2) linear prediction cepstral coefficients (LPCC), 3) one-sided autocorrelation linear prediction coefficients (OSALPC), 4) cepstral-based OSALPC (OSALPCC), and 5) mel-scale filter bank cepstral parameters (MFCC). We briefly discuss these features before presenting the evaluation results.

The OSALPC technique is based on the application of the windowed autocorrelation method of linear prediction to the one-sided autocorrelation sequence as discussed in [28], [25]. The one-sided autocorrelation (OSA) is obtained by computing $M = N/2$ autocorrelation lags from a speech frame of length N. A Hamming window from $m = 0$ to M is then applied to the



Recognition Performance of Neutral Trained Models Employing Static and Dynamic Features Derived from LP Power Spectrum

Training/Testing Features	*Speaking Styles Tested*				Overall Recognition
	NEUTRAL	ANGRY	LOUD	LOMBARD	
LPC + Δ	65.37%	45.56%	52.04%	53.70%	54.17%
OSALPC + Δ	74.26%	50.56%	54.44%	61.11%	60.09%
LPCC + Δ + CMN + Liftering	93.15%	72.59%	78.70%	88.15%	83.15%
OSALPCC + Δ + CMN + Liftering	92.78%	70.19%	76.85%	87.22%	81.76%
MFCC + Δ + CMN + Liftering	94.07 %	73.33%	79.81%	87.41%	83.66%

Fig. 5. Recognition performance of linear prediction power spectrum based static and dynamic features in the presence of stress. The graph shows the recognition rates of neutral trained models tested with four speaking conditions for five different sets of features: LPC, OSALPC, LPCC, cepstral-based OSALPC, and MFCC.

computed one-sided autocorrelation. This is followed by computing the first $p+1$ autocorrelation values of the one-sided autocorrelation sequence, where p is the prediction order. The autocorrelation values are used as entries to the Levinson-Durbin algorithm when estimating the AR parameters. Finally, cepstral coefficients are computed from the linear parameters using the recursive relation given below. Both LPC and OSALPC based cepstral parameters are derived from the linear spectral parameters according to the following equation:

$$c_k = a_k + (1/k)\sum_{i=1}^{k-1} i c_i a_{k-i} \quad 1 \le k \le p \tag{1}$$

where a_k are either the LPC or the OSALPC coefficients, p is the LPC analysis order, and c_k are the resulting cepstral coefficients. The final set of parameters considered in this section are the mel-frequency cepstral coefficients.

Next, we consider the performance of these features for recognition of speech under stress. Two sets of evaluations are presented. The first compares the performance of HMM models trained with static features with no parameter processing while the second compares the performance of models trained with static and dynamic features in addition to parameter processing such as cepstral liftering and cepstral mean normalization (CMN). In both evaluations, the models are speaker-independent *neutral* trained and are tested with speech from four speaking conditions: neutral, angry, loud, and Lombard effect.

The results plotted in Figs. 4 (static) and 5 (static, and delta with parameter processing) show that for both evaluations the one-sided autocorrelation linear prediction coefficients (OSALPC) performs better than traditional LPC for all three stress conditions. OSALPC, however, does not achieve the highest performance among the evaluated features. In fact, the three remaining cepstral features, cepstral-based LPC, cepstral-based OSALPC and MFCC, achieve higher recognition rates than OSALPC. In addition, our results show that cepstral-based OSALPC outperforms OSALPC by 12.36% across the four speaking conditions for static features (see Fig. 4), and by 21.67% for static and dynamic feature trained models (see Fig. 5).

The mel-frequency cepstral-based coefficients and the cepstral-based linear prediction coefficients achieve the highest recognition rates in both static and combined (static, dynamic, with parameter processing) evaluations. Their performance is very similar across all four speaking conditions as shown in Figs. 4 and 5. Both features achieve a higher level of recognition performance than cepstral-based OSALPC across

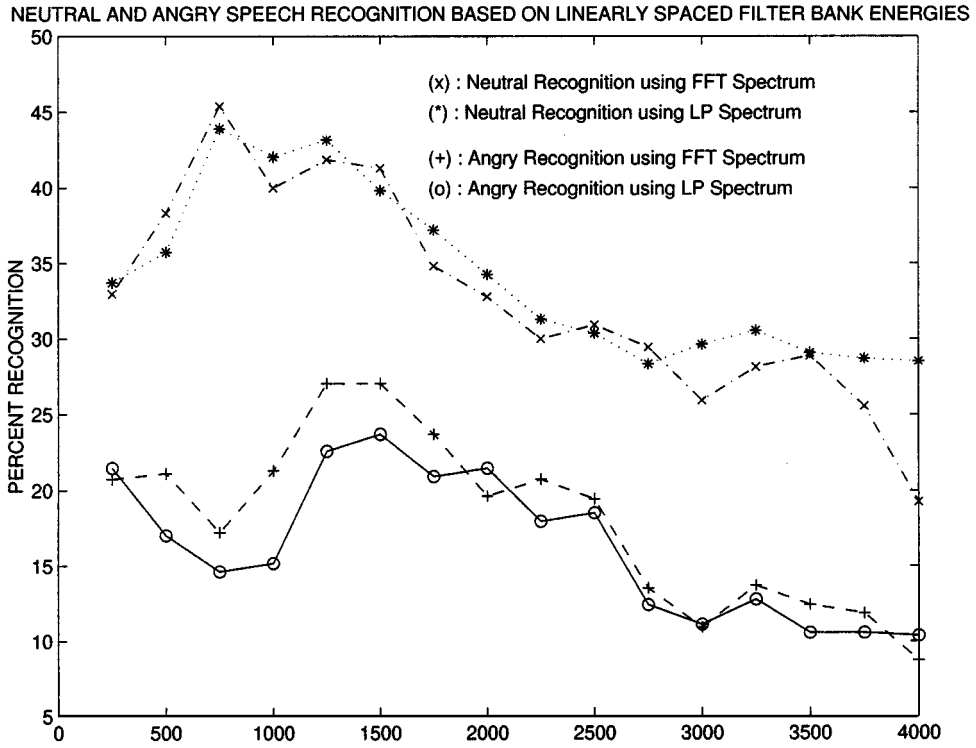


Fig. 6. Recognition based on individual linearly spaced filter bank output energies. Neutral trained models are tested with neutral and angry speech. The plot shows the results of both using a linear prediction and a FFT power spectrum.

all four speaking styles in both scenarios. In summary, these results show that while noise-robust features such as OSALPC may be robust in noise, they are not necessarily robust to the presence of speaker stress. These results also suggest that features based on cepstral analysis (LPCC, OSALPCC, MFCC) clearly outperform direct linear predictive based features (LPC, OSALPC), with overall recognition rates of 65%–69% versus 48%–53%. It is also recommended that due to the variability across the three stress conditions, new feature sets are needed for improved stressed speech recognition.

IV. Speech Recognition Based on Filter Bank Output

Next, we consider the impact of stress for speech recognition across frequency bands. Evaluations presented in Section III identified the need for a new set of features more robust to the presence of stress. Extensive studies have previously been conducted to understand how various types of stress affect human speech production. It is known, for example, that spectral slope varies significantly across stressed speaking styles. We therefore start by studying the impact of stress on individual frequency bands. This is achieved by evaluating the recognition performance based on the log-energy output of a 16 uniformly spaced filter bank. The ultimate goal is to formulate a new frequency scale which is less sensitive to variations caused by stress without degrading the performance of neutral speech recognition. We note that a similar study proved to be successful in the formulation of a set of accent sensitive frequency features [29]. A speaker-independent HMM model with variable state-duration is trained with neutral speech for each of the 16 frequency bands of a word. The neutral trained word models are tested with tokens of neutral and angry speech from SUSAS. The results shown in Fig. 6 are across 30 words spoken by nine speakers. The training and testing evaluations are done twice, once employing an FFT power spectrum, and a second employing a linear prediction spectrum. For completeness, we include below a short description of how the FFT and LP power spectra were computed.

A. FFT-Based Power Spectrum

The 8 kHz sampled signal, $s(n)$, is first processed by a simple preemphasis filter given by

$$H(z) = 1 - 0.97z^{-1} \tag{2}$$

to yield a roughly flat average speech spectrum. A 30-ms long Hamming window ($N = 240$ samples) with 50% overlap given by

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) \qquad n = 0, 1, \ldots, N-1 \tag{3}$$

is applied to the preemphasized waveform. A 512-point short-time Fourier transform is then computed for the successive overlapping windowed speech as given by

$$S(e^{jw}) = \sum_{n=m-N+1}^{m} s(n)w(m-n)e^{-jwn} \tag{4}$$

where $w(n)$ is a Hamming window of length N as given previously in (3). The magnitude squared power spectrum, $|S(e^{jw})|^2$, is then evaluated for frequencies ranging between 0 and 4 kHz. To compute the FFT-based log energy values, we apply 16 uniformly spaced digital filter banks on the resulting FFT power spectrum and compute the log energy in each band. The 16 log

energy values corresponding to each frame are employed for training as well as recognition.

B. LP-Based Power Spectrum

The speech signal, $s(n)$, is preemphasized and windowed as was previously described in (2) and (3). A 12th order linear prediction coefficients, a_i, are derived using the autocorrelation method which computes the autocorrelation coefficients and then applies the Levinson-Durbin recursion as described in [3]. The magnitude squared power spectrum is then directly computed from the LPC coefficients according to the following:

$$|H(e^{jw})|^2 = \frac{G^2}{\left|1 - \sum_{i=1}^{p} a_i e^{-jwi}\right|^2} \tag{5}$$

where p is the linear predictor filter order, and G^2 is the gain squared which is equivalent to the mean squared prediction error [35] as given by

$$G^2 = R_n(0) - \sum_{i=1}^{p} a_i R_n(i) \tag{6}$$

where $R_n(0)$ is the zeroth autocorrelation coefficient, p is the order of the linear prediction filter, a_i is the ith linear prediction coefficient, and $R_n(i)$ is the ith autocorrelation coefficient for frame n. The power spectrum is evaluated for frequencies ranging between 0 and 4 kHz. To compute LP-based log energy values, we apply 16 uniformly spaced digital filter banks on the resulting linear prediction power spectrum and compute the log energy in each band. These log energy values are employed for training and recognition.

C. Recognition Results Based on Filter Bank Log Energy

As shown in Fig. 6, for both power spectral methods the highest recognition performance for neutral speech (top two lines) occurs around the first formant location (200–1000 Hz) while the highest recognition of angry speech (lower two lines) occurs in the neighborhood of the second formant location in the range of 1250 to 1750 Hz. This may be attributed to the observation that the second formant location is more closely related to tongue movement, and that the variations in tongue movement from neutral to stress conditions are less dramatic than other changes such as excitation for example. Therefore, since the second formant location experiences less variability under stress, then it would be more reliable for stressed speech recognition. Recall that since a mel-scale is almost linear for frequencies below 1000 Hz and increases logarithmically above 1000 Hz, then the contribution of the second formant is de-emphasized compared to the first formant. This attribute makes the mel-scale ideal for neutral speech recognition but not equally effective for angry speech recognition. Therefore, a new frequency scaling is suggested which would emphasize frequencies around the first and second formant locations without degrading the recognition performance of neutral speech.

The results based on individual frequency band output energies do not conclusively determine which power spectral estimation method (linear prediction versus FFT) would be more robust to stress. In a later section, we investigate the performance of both spectral estimation methods in stress using whole-word based models. In the following section, we introduce two new frequency scale methods targeted at emphasizing the second formant in both neutral and stressed speech.

V. MFCC versus Newly Proposed Frequency Scales

From the previous filter bank analysis, it became evident that a new frequency scale is needed that would emphasize mid-frequencies while de-emphasizing lower and higher frequencies. Before proposing such a scale, we recall how traditional mel-frequency cepstral parameters are computed and then derive two new frequency scales that would achieve the desired frequency partitioning.

The mel-frequency cepstral coefficients can be computed from either the FFT or the LP power spectrum which were derived earlier in Sections IV-A and IV-B. The main difference between a mel-scale and a linear scale is that a linear scale places equal emphasis on every part of the frequency scale from zero up to the maximum representable frequency while a mel frequency scale places the bins on a nonlinear frequency scale. The bands in a mel-scale are made successively broader with increasing frequency above 1 kHz to reflect the frequency resolution of the human ear. To compute the mel-frequency cepstral coefficients, we place a set of 16 triangular bandpass filters on the desired power spectrum (either FFT or LP) according to a mel frequency scale [38] and compute the log energy in each band. Finally, a cosine transform is applied to convert the set of log energies to a set of *cepstral coefficients*. Only 12 of these cepstral coefficients are kept for training and recognition.

As was noted earlier, the mel-frequency scale is basically a mapping of the linear frequency scale to a scale that resembles the frequency resolution of the human ear. The general form of this logarithmic mapping can be written as

$$y = C \times \log\left(1 + \frac{f}{k}\right), \quad f = 0, \ldots, 4000 \text{ Hz} \tag{7}$$

where f represents the linear frequency and y represents the value of the transformed frequency. This equation can also be written as

$$C = \frac{y}{\log\left(1 + \frac{f}{k}\right)}. \tag{8}$$

Furthermore, if we require that the mapped values of the linear frequencies 0 and 4 kHz be equal to the mapped values of the mel frequency scale, i.e., 0 and $2595 \times \log(1 + 4000/700)$, we obtain the following equation relating C to k

$$C = \frac{2146.0645}{\log\left(1 + \frac{4000}{k}\right)}. \tag{9}$$

To compute the modified mel-scale cepstral coefficients, we place a set of 16 triangular bandpass filters on the desired power spectrum according to the modified mel-scale and compute the log energy in each band. Once again, we apply a cosine transform to convert the set of log energies to a set of cepstral coefficients. Therefore, the main difference between MFCC and modified MFCC is in the frequency placement of the bandpass filters.

The previously derived logarithmic mapping function did not yield the desired effect of heavily concentrating the filters at

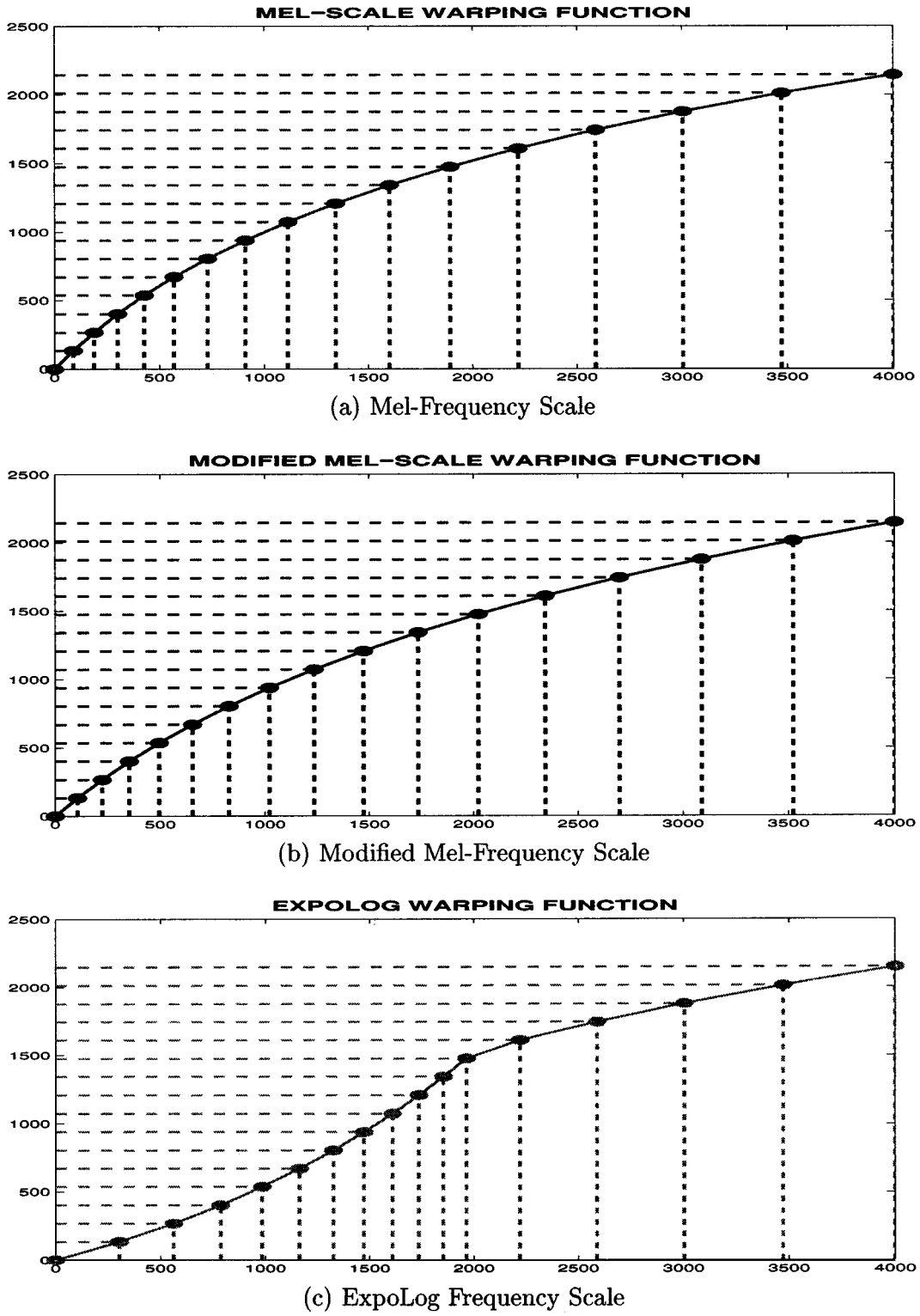


Fig. 7. This graph shows three mapping functions: (a) mel-frequency, (b) modified mel-frequency, and (c) ExpoLog, employed to warp a linear scale in the frequency domain.

mid-frequencies. To achieve this effect, we resort to an exponential mapping function for linear frequencies between 0 and 2000 Hz, and a logarithmic mapping function for linear frequencies between 2000 and 4000 Hz. The general form of these functions can be written as

$$y_1 = C_1 \times (10^{f/k_1} - 1) \quad 0 \le f \le 2000 \text{ Hz} \tag{10}$$

and

$$y_2 = C_2 \times \log\left(1 + \frac{f}{k_2}\right) \quad 2000 < f \le 4000 \text{ Hz} \tag{11}$$

We can solve for the values of $C_1, k_1, C_2,$ and k_2 by solving a set of equations. Once again, if we require that the mapped values of the linear frequencies 0 and 4000 Hz be equal to the

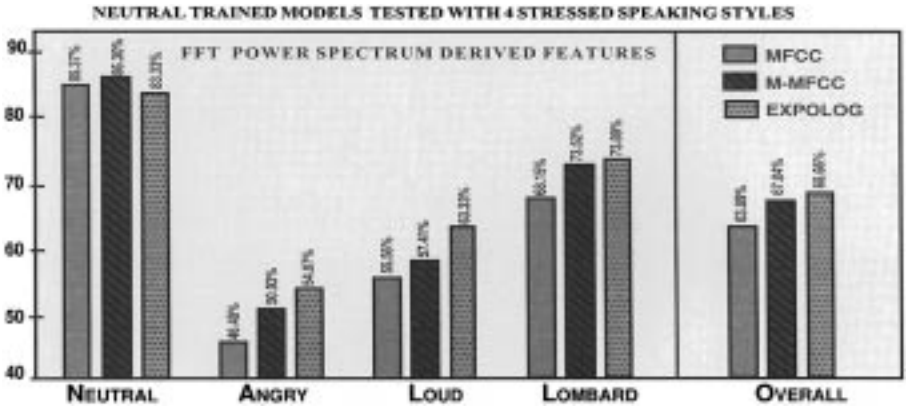


Fig. 8. Performance of the static features of three different frequency mapping functions: MFCC, M-MFCC, and ExpoLog, based on an FFT power spectrum.

mapped values of the mel scale, we obtain the following relation between C_2 and k_2

$$C_2 = \frac{2146.0645}{\log\left(1 + \frac{4000}{k_2}\right)}. \tag{12}$$

In addition, both the exponential and the logarithmic mapping functions should be equal at the boundary for the linear frequency $f = 2000$ Hz. This results in the following equation:

$$C_1 \times \left(10^{\frac{2000}{k_1}} - 1\right) = C_2 \times \log\left(1 + \frac{2000}{k_2}\right). \tag{13}$$

By using the same logarithmic mapping function derived earlier, we obtain the following relation between C_1 and K_1

$$C_1 \times \left(10^{\frac{2000}{k_1}} - 1\right) = 3070 \times \log\left(1 + \frac{2000}{1000}\right) \tag{14}$$

$$= 1464.76. \tag{15}$$

To compute the ExpoLog cepstral coefficients, we place a set of 16 triangular bandpass filters on the desired power spectrum according to the ExpoLog scale and compute the log energy in each band. We apply a cosine transform to convert the set of log energies to a set of cepstral coefficients. As noted before, the main difference among the three scales is in the frequency placement of the bandpass filters.

The modified mel-scale (M-MFCC), the ExpoLog scale, and the traditional mel-scale are as follows:

$$\text{mel-scale} = 2595 \times \log\left(1 + \frac{f}{700}\right) \tag{16}$$

$$\text{Modified mel-scale} = 3070 \times \log\left(1 + \frac{f}{1000}\right) \tag{17}$$

$$\text{ExpoLog} = \begin{cases} 700 \times \left(10^{\frac{f}{3988}} - 1\right) & 0 \le f \le 2000 \text{ Hz} \\ 2595 \times \log\left(1 + \frac{f}{700}\right) & 2000 < f \le 4000 \text{ Hz} \end{cases} \tag{18}$$

These three frequency warping functions are plotted in Fig. 7 for comparison. The y-axis represents the linear scale which is warped to the desired scale according to the mapping function. Note how for the ExpoLog mapping the filter banks are highly concentrated at mid frequencies while they are sparsely distributed at frequencies below 750 Hz and above 2000 Hz. In this section, we will contrast the performance of the three warping functions using an FFT power spectrum. A comparison of their performance using a linear prediction power spectrum will be discussed in the following section.

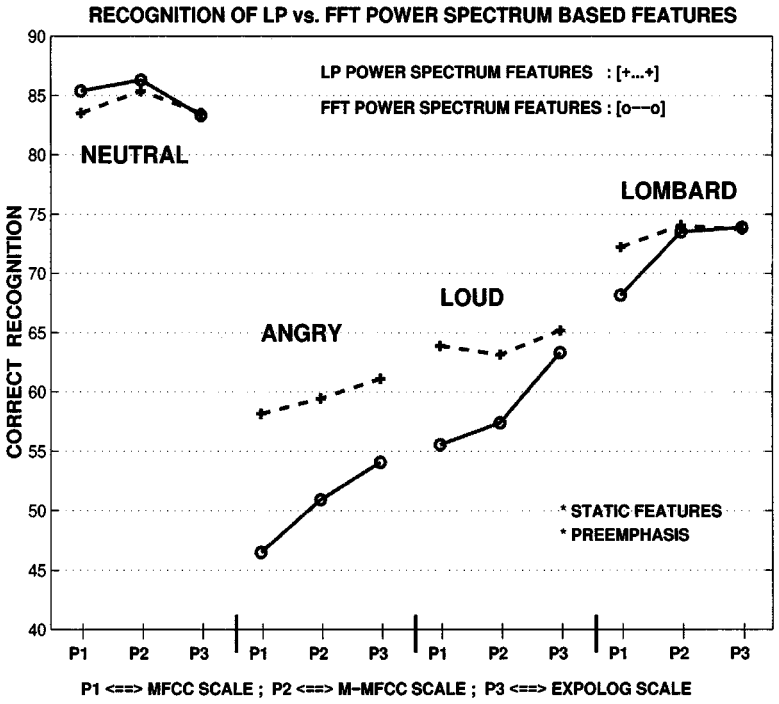
Fig. 8 shows results from an evaluation of the three frequency warping scales in obtaining cepstral parameters (MFCC, M-MFCC, ExpoLog). Recognition rates are shown for neutral models trained with static features and tested with speech from neutral and three stressed speaking conditions. Each scaling method was evaluated with a total of 2160 open test tokens. When static features are employed for recognition, M-MFCC outperforms traditional MFCC by 4.45% for angry, 1.85% for loud, and 5.37% for Lombard effect. The performance of ExpoLog static features also outperforms the mel-scale, for all stress styles, with an average performance improvement of 4.77%. Note that for angry and loud speech recognition, ExpoLog exceeds MFCC by as much as 7.59% and 7.77%. These results clearly show that with a slight modification in the manner in which cepstral parameters are obtained, we can improve recognition performance in stressed speech conditions.

VI. Fast Fourier Transform versus Linear Prediction Power Spectrum

In a recent survey of contemporary recognition systems by Picone [27], it was established that fast Fourier transform based spectral parameters are preferred to linear prediction based parameters since they are believed to be more immune to the presence of noise. Only a third of all the surveyed systems employed linear prediction derived parameters, the remainder used FFT based processing. For this reason, a number of systems rely on the FFT-based filter bank analysis. In order to compare the immunity of each power spectral estimation to stress, we conducted two recognition evaluations using parameters derived from FFT and linear prediction power spectral estimation methods. In addition, we performed an additional recognition evaluation employing actual stressed speech produced in a noisy environment in order to determine which power spectral estimation method would be more robust to the presence of both noise and stress. The noise in this case represents time varying mechanical and wind noise obtained from speech recorded during amusement park roller coaster rides.

A. Performance in Noise-Free Simulated Stress Conditions

Our results show that contrary to their noise immunity, FFT-based spectral parameters are not equally robust to the presence of stress. Fig. 9 compares the performance of linear



Recognition Performance of Neutral Trained Models Employing Static MFCC, M-MFCC, and ExpoLog Features Derived from FFT and LP Power Spectrum

Training/Testing Features		*Speaking Styles Tested*				Average Recognition
		NEUTRAL	ANGRY	LOUD	LOMBARD	
MFCC	[FFT]	85.37%	46.48%	55.56%	68.15%	63.89%
M-MFCC	[FFT]	86.30%	50.93%	57.41%	73.52%	67.04%
ExpoLog	[FFT]	83.33%	54.07%	63.33%	73.89%	68.66%
MFCC	[LP]	83.52%	58.15%	63.89%	72.22%	69.45%
M-MFCC	[LP]	85.37%	59.44%	63.15%	74.07%	70.51%
ExpoLog	[LP]	83.52%	61.11%	65.19%	73.70%	70.88%

Fig. 9. This graph compares the performance of FFT versus linear prediction power spectrum derived features of neutral trained models using static features. The performance of MFCC is compared to two different frequency scales.

prediction and FFT power spectrum based features. The dotted line represents linear prediction based recognition and the solid line represents FFT based recognition rates. For neutral training and testing, FFT based parameters perform slightly better than cepstral parameters derived from a linear prediction spectrum. However, the linear prediction power spectrum performs significantly better than the FFT power spectrum when neutral trained models are tested with angry, loud, and Lombard effect speech. We also point out that modified MFCC (M-MFCC) and ExpoLog based features consistently outperformed MFCC parameters using both FFT and linear prediction based spectra, but that linear prediction derived ExpoLog produced the highest recognition rates across stressed styles using static features. Next, we consider extending the static features to include time derivatives and feature processing. Time derivatives or delta parameters were shown to greatly enhance the performance of stressed speech recognition [30]. Having established the ExpoLog frequency scale as being superior to mel and modified-mel scales, we now consider time derivatives and parameter processing. Fig. 10 also compares both spectral methods. It shows the performance of ExpoLog static and dynamic features with parameter processing. Once

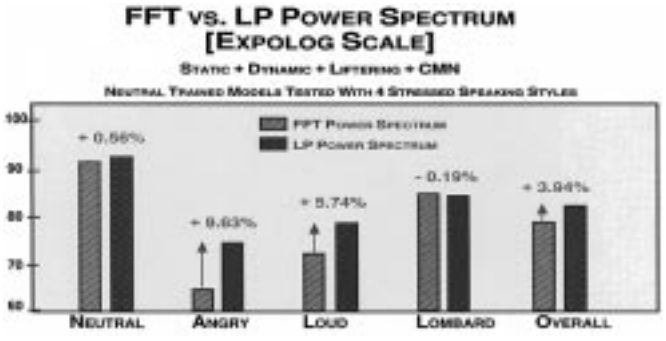


Fig. 10. Performance of FFT and linear prediction power spectrum based ExpoLog static and dynamic features.

again, the linear prediction based features outperform FFT by an overall 3.94%. For angry speech recognition, the difference in recognition is as high as 9.63%.

An error analysis was also conducted for linear prediction and FFT derived ExpoLog cepstral parameters. For this evaluation, the database was divided into confusable and nonconfusable word sets as shown in Table I. For both linear prediction and FFT power spectrum, the nonconfusable errors account for the

TABLE I
A LIST OF THE 30 WORD VOCABULARY EMPLOYED IN ALL EVALUATIONS. THIS TABLE SHOWS THE GROUPING OF CONFUSABLE AND NONCONFUSABLE WORDS

Confusable Words			Non-Confusable Words
break/strafe	change/gain	east/eight	enter, hot, mark, on,
fix/six	go/no/oh	hello/help	three, point, stand
out/south	white/wide	degree/freeze/three	nav, steer, ten, zero

TABLE II
PERFORMANCE OF FFT AND LINEAR PREDICTION POWER SPECTRUM BASED FEATURES IN ACTUAL STRESSED NOISY SPEECH

Recognition Performance of Neutral Trained Models Employing Static and Dynamic MFCC, M-MFCC, and ExpoLog Features Derived from FFT and LP Power Spectrum in noisy actual stressed conditions		
Feature Set	LP Power Spectrum	FFT Power Spectrum
MFCC	36.72%	28.81%
M-MFCC	36.16%	25.42%
ExpoLog	37.29%	22.60%

majority of errors in angry and loud speech recognition. For neutral and Lombard conditions, both methods produced comparable word-set error rates. Our findings suggest that the linear prediction power spectrum mainly resolves nonconfusable errors. The error rates for confusable words are comparable for both FFT and linear prediction spectral methods.

B. Performance in Actual Noisy Stressful Conditions

Having established that the linear prediction power spectra outperforms an FFT based power spectra in noise-free simulated stress conditions, we now consider a second evaluation using actual noisy stressful speech from the SUSAS database. This evaluation is intended to determine which power spectral estimation method is most effective when speech is subjected to a combination of noise and stress. The results, as summarized in Table II, indicate that the linear prediction based features outperform the FFT-based features not only for noise-free simulated stress conditions but also for noisy actual stressed speech. We believe that the spectral smoothing inherent in the linear prediction model provides a more overall smooth set of parameters capable of not representing the fine variations caused by excitation changes (i.e., pitch structure) that exist under stressful conditions.

VII. PARAMETER PROCESSING

While it is desirable to formulate speech features which are inherently robust to the variability of speech under stress, there are a number of possible subsequent parameter processing methods that could be used which have been shown to be effective for noise and communication channel effects. We note that other methods such as stress equalization feature processing (MCE-ACC [18], and others in [21]) have been shown to be effective in reducing the impact of stress. However, such stress equalization processing requires stress and/or word dependent compensation terms. The goal in this section is to consider only feature processing methods which do not require knowledge of either word or phoneme class sequence content, or the type of speaker stress. In this section, we consider the following three parameter processing methods: preemphasis, liftering, and cepstral mean normalization which have been widely used for improved speech recognition and speaker identification. Here, we evaluate their contribution to stressed speech recognition.

A. Fixed and Slowly-Varying Preemphasis

Previous analysis studies on stressed speech have shown that the spectral structure and overall average spectral slope varies for different speaking conditions [5], [21], [31]–[34]. In speech recognition, a preemphasis filter is normally used to raise the high frequency content by approximately 20 dB per decade. The preemphasis filter has the effect of compressing the dynamic range in the frequency domain of the speech signal by flattening the spectral tilt so as to improve linear modeling of the formant structure. This is especially useful for voiced sections since they naturally have a negative spectral slope due to physiological characteristics of the speech production. Since the average spectral slope of the input speech is different for various stressed speaking styles, then in order to flatten the spectral tilt, it is necessary to vary the filter parameters according to the input speech. Therefore, we propose using an adaptive preemphasizer where only the spectral slope of voiced speech is adaptively flattened while the unvoiced speech sections are not preemphasized. The adaptive preemphasizer is a slowly-varying first order filter [36] given by

$$H(z) = 1 - \hat{a}_n z^{-1}$$

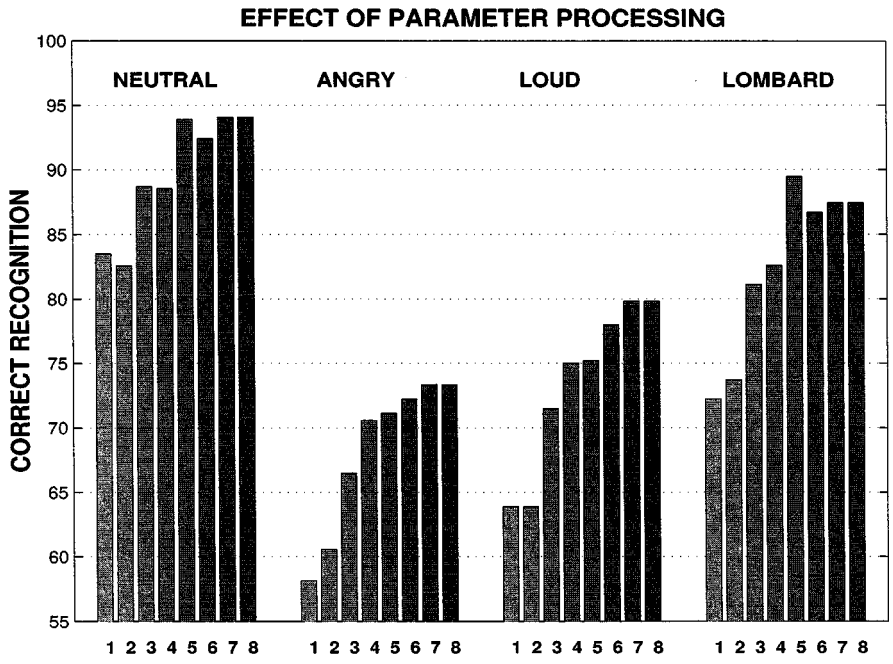
where $\hat{a}_n = r_n(1)/r_n(0)$. The variable filter coefficient is represented as a ratio of the first to the zeroth order lag autocorrelation parameters. The filter is applied to utterances both during training and testing.

An evaluation of the MFCC features with preemphasis was conducted across neutral and three stressed speaking conditions. The slowly-varying preemphasis filter improved recognition of angry speech by 2.41% and that of Lombard effect speech by 1.48%. The recognition of neutral speech dropped by 1% while the same performance was achieved for loud speech. The overall recognition performance was slightly improved by applying a slowly-varying preemphasis.

B. Fixed and Variable Cepstral Mean Normalization

In a study of channel compensation techniques for speaker identification, simple cepstral mean removal was the best channel compensation method versus RASTA processing and quadratic trend removal [37]. Cepstral mean normalization or removal is a simple yet effective method which assumes no knowledge about the environment and is employed for reducing long term differences in channel characteristics. The channel distortion is computed as a long-term cepstral average by estimating the mean of each cepstral value over the entire utterance. It is assumed that the speech signal is rich in phonemic content, so that the estimated mean will reflect only that spectral structure which is common to all observation frames (i.e., frequency structure due to microphone or channel effects). The channel effects are then removed by subtracting this mean from the cepstral value in each frame.

A number of variations on CMN have also been used to compensate for the variation of cepstral coefficients when speech is produced with different speaking styles [2], [16].



	Recognition Performance of MFCC Neutral Trained Models Tested with Neutral and Three Stressed Speaking Styles					Average Recognition
	Training/Testing Features	NEUTRAL	ANGRY	LOUD	LOMBARD	Performance
1	Fixed Preemphasis (FxdP)	83.52%	58.15%	63.89%	72.22%	69.45%
2	Varying Preemphasis	82.59%	60.56%	63.89%	73.70%	70.19%
3	FxdP + CMN	88.70%	66.48%	71.48%	81.11%	76.94%
4	FxdP + Variable CMN	88.52%	70.56%	75.00%	82.59%	79.17%
5	FxdP + Δ	93.89%	71.11%	75.19%	89.44%	82.41%
6	FxdP + Δ + Variable CMN	92.41%	72.22%	77.96%	86.67%	82.32%
7	FxdP + Δ + CMN	94.07%	73.33%	79.81%	87.41%	83.66%
8	FxdP + Δ + CMN + Liftering	94.07%	73.33%	79.81%	87.41%	83.66%

Fig. 11. Effect of preemphasis (fixed and variable), cepstral mean normalization (fixed and variable), time-derivative (Δ coefficients), and cepstral liftering on the recognition performance of linear prediction based MFCC's. Note, "FxdP" refers to "Fixed Preemphasis," and Δ refers to time-derivative delta coefficients.

Since, the presence of stress impacts voiced and unvoiced speech phonemes differently [5], [6], we propose computing an overall separate cepstral mean for voiced and unvoiced sections for each token under test instead of computing a single mean across the entire utterance. This was achieved by first using a simple voiced/unvoiced speech detection approach. The average voiced cepstral mean is then subtracted from voiced sections, and the unvoiced cepstral mean is subtracted from unvoiced sections. Unlike applying a compensation vector, this method does not require prior analysis of neutral and stress speech data since it computes the mean during the parameterization stage.

A series of the recognition evaluations was performed using MFCC static parameters, with various configuration of delta parameters, fixed or variable preemphasis, and fixed or variable cepstral mean normalization (CMN). The results in Fig. 11 show that variable cepstral mean normalization performs better than traditional CMN when no delta parameters are employed. The recognition of angry, loud, and Lombard effect speech is improved respectively by 4.08%, 3.52% and 1.48%. Variable CMN is most effective with static features and fixed preemphasis. If time derivatives are included in the feature set, variable CMN is not as effective.

C. Cepstral Liftering

The last feature processing method considered is cepstral liftering. Cepstral liftering is a weighting technique applied to cepstral coefficients in order to reduce the spectral slope or the undesirable broadband noise components of the spectrum, which affect low order cepstral parameters, while retaining the essential characteristics of the formant structure. The low-order cepstral coefficients are believed to be primarily sensitive to overall spectral slope, variations in transmission, speaker characteristics, or vocal efforts. The higher order cepstral coefficients represent fine spectral structure and are therefore more sensitive to noise and the artifacts of the LPC analysis. The cepstral liftering applied here is given by

$$c'_n = \left(1 + \frac{L}{2}\sin\frac{\pi n}{L}\right) c_n$$

where L was set to 12.

Cepstral liftering was also evaluated using MFCC parameters for neutral and three stressed speech conditions. A summary of

the results is shown in Fig. 11. Our evaluations show that when cepstral liftering is employed with time derivative parameters and cepstral mean normalization, it has no effect on the recognition performance of neutral trained models tested with neutral, angry, loud, and Lombard conditions.

VIII. Conclusion

In this study, we have considered the effectiveness of noise robust features for recognition of speech under stress, and have proposed alternative feature extraction methods for improved stressed speech recognition. Noise robust features, such as the one-sided autocorrelation linear prediction (OSALPC) and cepstral-based OSALPC features, were compared to traditional features such as linear prediction coefficients (LPC), LPC-based cepstral (LPCC) parameters, and mel-frequency cepstral (MFCC) parameters for stressed speech recognition. Our investigation showed that noise robust features are not necessarily robust to the presence of stress since the effect of noise on the acoustic speech signal is different than the effect of stress.

In order to formulate features less sensitive to the effects of stress, we studied the recognition performance of stressed speech based on individual frequency bands. This evaluation showed that frequencies near the second formant location achieve the highest recognition rates for angry speech, while frequencies near the first formant location achieve the highest recognition rates for neutral speech. From these observations, we formulated two new feature extraction methods, a modified mel-frequency scale (M-MFCC), and an exponential-logarithmic scale (ExpoLog) for improved stressed speech recognition. Both methods were shown to outperform the traditional mel-scale for recognition of speech under a variety of stressed styles.

Since the majority of current commercial speech recognition systems are based on an FFT power spectrum due to its reported immunity to noise, this study compared the performance of FFT-based features to linear prediction based power spectrum features in the presence of stress. Contrary to the FFT's immunity to noise, the FFT power spectrum was less robust to stress. Features based on the linear prediction power spectrum outperformed the FFT-based features not only for noise-free simulated stress conditions but also for speech under actual noisy stressful conditions.

Finally, the effect of parameter processing on stressed speech recognition was evaluated. The evaluations considered the effect of liftering, fixed versus variable preemphasis, and fixed versus variable cepstral mean normalization (CMN) on recognition performance. It was determined that cepstral liftering had no effect on stressed speech recognition performance. Variable preemphasis slightly improved recognition over fixed preemphasis. Finally, variable CMN improved recognition over fixed CMN by 3% when static parameters are employed for recognition. Their performance is not equally effective when time derivative parameters are included. It may also be of interest to comment on how these results could be extended to large vocabulary continuous speech recognition systems. As Table III shows, the phoneme coverage for the vocabulary of SUSAS covers many but not all phonemes in English. Until more extensive stress speech databases are available, we can only suggest that at least at the phone level, recognition performance should be comparable. Further investigations are needed to determine changes in language models and sentence structure when recognizing continuous speech under stress. The final recommendation from this study is that for effective speech recognition performance in both neutral and stressed conditions, speech recognizers should 1) employ features derived from a linear prediction as opposed to an FFT based power spectrum and 2) use a modified frequency partition such as M-MFCC or ExpoLog if possible. In addition, variable preemphasis and variable CMN both improve stressed speech recognition performance, but that their impact is reduced if time derivative parameters are also included.

TABLE III
A List of the Phonemes in the SUSAS Database Using the *Advanced Research Projects Agency uppercase alphabet notation (ARPAbet)*

A List of the Phonemes in the SUSAS Database							
/m/	/n/	/s/	/th/	/f/	/sh/	/z/	/v/
/jh/	/ch/	/p/	/t/	/k/	/b/	/d/	/g/
/hh/	/ih/	/eh/	/ae/	/aa/	/axr/	/ah/	/ay/
/oy/	/ey/	/ow/	/aw/	/iy/	/w/	/l/	/r/

References

[1] J. H. L. Hansen, C. Swail, A. J. South, R. K. Moore, H. Steeneken, E. J. Cupples, T. Anderson, C. R. A. Vloeberghs, I. Trancoso, and P. Verlinde, "The impact of speech under 'stress' on military speech technology," NATO Res. Technol. Org. RTO-TR-10, AC/323 (IST) TP/5 IST/TG-01, (ISBN: 92–837–1027–4), Mar. 2000.

[2] Y. Chen, "Cepstral domain stress compensation for robust speech recognition," in *Proc. IEEE Inter. Conf. on Acoustics, Speech, and Signal Processing*, Dallas, TX, Apr. 1987, pp. 717–720.

[3] J. Deller, J.H.L. Hansen, and J. Proakis, *Discrete-Time Processing of Speech Signals*, 2nd ed. New York: IEEE Press, 2000.

[4] J. H. L. Hansen and M. A. Clements, "Evaluation of speech under stress and emotional conditions," *Proc. Acoust. Soc. Amer.*, vol. 82, Nov. 1987.

[5] J. H. L. Hansen, "Analysis and compensation of stressed and noisy speech with application to robust automatic recognition," Ph.D. dissertation, Georgia Inst. Technol., Atlanta, July 1988.

[6] J. H. L. Hansen and M. A. Clements, "Stress compensation and noise reduction algorithms for robust speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Glasgow, U.K., May 1989, pp. 266–269.

[7] ——, "Source generator equalization and enhancement of spectral properties for robust speech recognition in noise and stress," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 407–415, Sept. 1995.

[8] R. P. Lippmann, E. A. Martin, and D. B. Paul, "Multi-style training for robust isolated-word speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Dallas, TX, Apr. 1987, pp. 705–708.

[9] J. H. L. Hansen and S. E. Bou-Ghazale, "Getting started with SUSAS: A speech under simulated and actual stress database," in *EUROSPEECH-97: Eur. Conf. Speech Communication Technology*, vol. 4, Rhodes, Greece, Sept. 1997, pp. 1743–1746.

[10] J. H. L. Hansen, "Analysis of acoustic correlates of speech under stress. Part 1: Fundamental frequency, duration, and intensity effects," *J. Acoust. Soc. Amer.*, p. 34, Oct. 1998.

[11] ——, "Analysis of acoustic correlates of speech under stress. Part 2: Glottal source, and vocal tract spectral effects," *J. Acoust. Soc. Amer.*, p. 27, Oct. 1998.

[12] J. H. L. Hansen, G. Zhou, and R. Sarikaya, "Analysis of acoustic correlates of speech under stress. Part 3: Applications to speech recognition, speaker identification, and stress classification," *J. Acoust. Soc. Amer.*, p. 30, May 1999.

[13] E. Lombard, "Le signe de l'elevation de la voix," *Ann. Maladies Oreille, Larynx, Nez, Pharynx*, vol. 37, pp. 101–119, 1911.

[14] J. H. L. Hansen and S. E. Bou-Ghazale, "Robust speech recognition training via duration and spectral-based stress token generation," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 415–421, Sept. 1995.
[15] B. Stanton, L. Jamieson, and G. Allen, "Robust recognition of loud and Lombard speech in the fighter cockpit environment," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Glasgow, U.K., May 1989, pp. 675–678.
[16] J. H. L. Hansen and O. N. Bria, "Lombard effect compensation for robust automatic speech recognition in noise," in *Inter. Conf. Spoken Language Processing*, Kobe, Japan, Nov. 1990, pp. 1125–1128.
[17] Y. Takizawa and M. Hamada, "Lombard speech recognition by formant-frequency-shifter LPC cepstrum," in *Inter. Conf. Spoken Language Processing*, Kobe, Japan, Nov. 1990, pp. 293–296.
[18] J. H. L. Hansen, "Morphological constrained enhancement with adaptive cepstral compensation (MCE-ACC) for speech recognition in noise and Lombard effect," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 598–614, Oct. 1994.
[19] B. D. Womack and J. H. L. Hansen, "Classification of speech under stress using target driven features," *Speech Commun.*, vol. 20, pp. 131–150, Nov. 1996.
[20] J. C. Junqua, "The Lombard reflex and its role on human listeners and automatic speech recognizers," *J. Acoust. Soc. Amer.*, vol. 93, no. 1, pp. 510–524, 1993.
[21] J. H. L. Hansen, "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition,," *Speech Commun.*, vol. 20, pp. 151–173, Nov. 1996.
[22] S. V. Vaseghi, P. N. Conner, and B. P. Milner, "Speech modeling using cepstral-time feature matrices in hidden markov models," *Proc. Inst. Elect. Eng.*, vol. 140, pp. 317–320, Oct. 1993.
[23] K. Assaleh, R. Mammone, M. Rahim, and J. Flanagan, "Speech recognition using the modulation model," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Adelaide, Australia, Apr. 1994, pp. 664–667.
[24] T. Usagawa, M. Iwata, and M. Ebata, "Speech parameter extraction in noisy environment using a masking model," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Adelaide, Australia, Apr. 1994, pp. 81–84.
[25] J. Hernando and C. Nadeu, "Linear prediction of the one-sided autocorrelation sequence for noisy speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 5, pp. 80–84, Jan. 1997.
[26] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech Commun.*, no. 16, pp. 261–291, 1995.
[27] J. W. Picone, "Signal modeling techniques in speech recognition," *Proc. IEEE*, vol. 81, pp. 1215–1247, Sept. 1993.
[28] J. Hernando and C. Nadeu, "Speech recognition in noisy car environment based on OSALPC representation and robust similarity measuring techniques," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Adelaide, Australia, 1994, pp. 69–72.
[29] L. M. Arslan and J. H. L. Hansen, "A study of temporal features and frequency characteristics in American English foreign accent," *J. Acoust. Soc. Amer.*, vol. 102, no. 1, pp. 28–40, July 1997.
[30] B. A. Hanson and T. H. Applebaum, "Robust speaker-independent word recognition using static, dynamic and acceleration features: Experiments with Lombard and noisy speech," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 1990, pp. 857–860.
[31] C. E. Williams and K. N. Stevens, "Emotions and speech: Some acoustical correlates," *J. Acoust. Soc. Amer.*, vol. 52, pp. 1238–1250, 1972.
[32] R. Roessler and J. W. Lester, "Vocal patterns in anxiety," in *Phenomenology and Treatment of Anxiety*, W. E. Fann, A. D. Pokorny, and R. L. Williams, Eds. New York: Spectrum, 1979.
[33] K. R. Scherer, "Nonlinguistic vocal indicators of emotion and psychopathology," in *Emotions in Personality and Psychopathology*, C. E. Izard, Ed. New York: Plenum, 1979, pp. 493–529.
[34] B. J. Stanton, L. H. Jamieson, and G. D. Allen, "Acoustic-phonetic analysis of loud and Lombard speech in simulated cockpit conditions," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, New York, 1988, pp. 331–334.
[35] L. Rabiner and R. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978.
[36] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
[37] D. A. Reynolds, "Experimental evaluation of features for robust speaker identification," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 639–643, Oct. 1994.
[38] S.B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 357–366, Aug. 1980.

Sahar E. Bou-Ghazale was born in Baabda, Lebanon. She received the B.S.E.E. degree with honors from the University of North Carolina, Charlotte, in 1991, and the M.S. and Ph.D. degrees in electrical engineering from the Robust Speech Processing Laboratory, Department of Electrical Engineering, Duke University, Durham, NC, in 1993 and 1996, respectively.

In January 1997, she joined the Electrical and Computer Engineering Department, Duke University, as an Assistant Research Professor. Since November 1997, she has been with the Speech Technology Development Group, Network Access Division at Conexant Systems (formerly Rockwell Semiconductor Systems), Newport Beach, CA. Her research interests are focused in the field of speech processing, including analysis, modeling, recognition, and synthesis.

Dr. Bou-Ghazale was the recipient of a Jefferson-Pilot scholarship in 1990, and a Phi Kappa Phi Graduate fellowship award in 1991. She is a member of Phi Eta Sigma, Tau Beta Pi, and Phi Kappa Phi. She served as President of the North Carolina Delta Chapter of Tau Beta Pi in 1990.

John H. L. Hansen (S'81–M'82–SM'93) was born in Plainfield, NJ. He received the the B.S.E.E. degree with highest honors from Rutgers University, New Brunswick, NJ, in 1982 and M.S. and Ph.D. degrees in electrical engineering from Georgia Institute of Technology, Atlanta, in 1983 and 1988, respectively.

He is presently Associate Professor with the Departments of Speech, Language, and Hearing Sciences and Electrical and Computer Engineering, University of Colorado, Boulder. In 1988, he established and has since directed the Robust Speech Processing Laboratory (RSPL). He serves as Associate Director for the Center for Spoken Language Research (CSLR), and directs the research activities of RSPL at CSLU. He was a Faculty Member with the Departments of Electrical and Biomedical Engineering, Duke University, for ten years before joining the University of Colorado in 1999. His research interests span the areas of digital speech processing, analysis and modeling of speech and speaker traits, speech pathology, speech enhancement and feature estimation in noise, robust speech recognition with current emphasis on robust recognition and training methods for topic spotting in accent, noise, stress, and Lombard effect, and speech feature enhancement in hands-free environments for human-computer interaction. He has served as a technical consultant to the U.S. Government and industry including AT&T Bell Laboratories, IBM, Sparta, Signalscape, HRL Laboratories, ASEC, VeriVoice, and DOD in the areas of voice communications, wireless telephony, robust speech recognition, and forensic speech/speaker analysis. He is the author of more than 100 journal and conference papers in the field of speech processing and communications, and is coauthor of the textbook *Discrete-Time Processing of Speech Signals* (New York: IEEE Press, 2000).

Dr. Hansen was an invited tutorial speaker for IEEE ICASSP'95 and the 1995 ESCA-NATO Speech Under Stress Research Workshop, Lisbon, Portugal. He has served as Technical Advisor to U.S. Delegate for NATO (IST/TG-01: Research Study Group on Speech Processing, 1996–1998), Chairman for the IEEE Communications and Signal Processing Society of North Carolina (1992–1994), Advisor for the Duke University IEEE Student Branch (1990–1997), Tutorials Chair for IEEE ICASSP'96, Associate Editor for the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING (1992–1998), and is presently serving as Associate Editor for IEEE SIGNAL PROCESSING LETTERS. He has also served as Guest Editor of the October 1994 Special Issue on Robust Speech Recognition for the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING. He was the recipient of a Whitaker Foundation Biomedical Research Award, a National Science Foundation's Research Initiation Award, and has been named a Lilly Foundation Teaching Fellow for "contributions to the advancement of engineering education." He will serve as General Chair for the International Conference on Spoken Language Processing (ICSLP-2002), which will be held in Denver, CO.