HMM-Based Stressed Speech Modeling with Application to Improved Synthesis and Recognition of Isolated Speech Under Stress

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

Abstract—In this study, a novel approach is proposed for modeling speech parameter variations between neutral and stressed conditions and employed in a technique for stressed speech synthesis and recognition. The proposed method consists of modeling the variations in pitch contour, voiced speech duration, and average spectral structure using hidden Markov models (HMM’s). While HMM’s have traditionally been used for recognition applications, here they are employed to statistically model characteristics needed for generating pitch contour and spectral perturbation contour patterns to modify the speaking style of isolated neutral words. The proposed HMM models are both speaker and word-independent, but unique to each speaking style. While the modeling scheme is applicable to a variety of stress and emotional speaking styles, the evaluations presented in this study focus on angry speech, the Lombard effect, and loud spoken speech in three areas. First, formal subjective listener evaluations of the modified speech confirm the HMM’s ability to capture the parameter variations under stressed conditions. Second, an objective evaluation using a separately formulated stress classifier is employed to assess the presence of stress imparted on the synthetic speech. Finally, the stressed speech is also used for training and shown to measurably improve the performance of an HMM-based stressed speech recognizer.

Index Terms—Lombard effect, robust speech recognition, speech synthesis, speech under stress.

I. INTRODUCTION

In this study, we consider the problem of speech under stress with applications to stress modification for speech synthesis, and improved training for robust speech recognition. Stress in this context refers to environmental, emotional, or workload stress. Stress has been shown to alter the normal behavior of human speech production and the resulting speech feature characteristics. The variability introduced by a speaker under stress causes speech recognition systems trained with neutral speech tokens to fail [1]–[4]. Hence, available speech recognition systems are not robust in actual stressful environments such as fighter cockpits, where a pilot is subjected to a number of stress factors such as G-force (gravity), environmental stress due to background noise (Lombard effect [5]), workload stress resulting from task requirements of operating in a cockpit, and emotional stress such as fear. In such environments, a speaker may experience a mixture of emotions or stress conditions rather than a single emotion. Therefore, it is important from the standpoint of voice communication and speech algorithm development to characterize the effects of each condition in order to understand the combined stress effect on speech characteristics. In addition, the same speaker may be subjected to different levels of stress, from mild to extreme, which may affect the variability of speech characteristics. It should also be noted that each person responds differently to a given stressful condition, and therefore, it is necessary to account for speaker variability under stress.

In this paper, we study the effects of individual stressful conditions on speech characteristics as opposed to a mixture of conditions. While a variety of speech under stress conditions are possible, the stress conditions of interest in our study are angry, loud, and the Lombard effect. Although, it is equally feasible to model the speech variations introduced by a particular speaker under stress, here the variations across a number of speakers are modeled. Our modeling is intended to represent general characteristics of speech under stress and not variations particular to an individual speaker. This would allow us to develop a general method of stress perturbation, which could be applied to modify the speaking style of any new input synthesis speaker in a way that would convince a majority of listeners that the modified speech is under stress. Therefore, this study develops a novel technique for pitch contour, duration and spectral contour modeling using hidden Markov models (HMM’s) for the purpose of stressed speech synthesis with application to stressed speech recognition. The HMM perturbation models are word-independent assuming the word consists of any number of unvoiced regions and one voiced region. The advantages of modeling the parameter variations using HMM’s are as follows.

1) The models can characterize the stressed data and can also reproduce unlimited observation sequences with the same statistical properties as the training data (due to the regenerative property of HMM’s).

1 The Lombard effect results when speakers attempt to modify their speech production system in order to increase communication quality while speaking in a noisy environment.
2) Since HMM's can regenerate a large number of observation sequences, a single neutral word can be perturbed in an unlimited number of ways (allowing for a broad range of emotional/stress conditions to be simulated).

3) A larger database of stressed synthetic speech can be generated from an originally smaller neutral data set.

Several areas in speech processing can benefit from establishing a model for parameter variations under stress. We focus here on the impact of our study on the areas of modeling, synthesis, and recognition of speech under stress. Since our study models variations in actual speech parameters, the resulting model should provide a better understanding of the effects of stress on actual speech characteristics. Consequently, these models can be applied directly to neutral speech or synthetic speech utterances to modify the speaking style. These models can also be applied to enhance the naturalness of synthetic speech and/or modify speaking style. Finally, the knowledge from these models can be integrated within a recognition system to improve performance under stress. Alternatively, the models can be used to generate synthetic stressed data from neutral speech. The synthetic stressed speech can then be used for training. This will eliminate the need for collecting stressed speech for training.

The general framework proposed in this paper can be divided into two goals (see Fig. 1). The first goal is that of speech parameter modeling via HMM’s, and the second consists of speaking style modification or perturbation using the HMM model. The modeling stage consists of identifying the speech parameters that are most sensitive to stress, and training an HMM model with the parameter variations that occur under stress. Note that the HMM models are trained with the variations that occur between neutral and stressed speech parameters rather than with actual parameter values, since this will vary from speaker-to-speaker. Therefore, the focus is not to develop a speaker-dependent stress modification scheme, but instead to develop a general method of stress perturbation that will convince a majority of listeners that the modified neutral speech is under stress. After training, and in the perturbation stage, the trained HMM perturbation models are used to statistically generate perturbation vectors that are used to modify the speaking style of input neutral speech.

The remainder of this paper is organized as follows. Section II summarizes the previous approaches to synthesis and recognition of speech under stress or emotions. The speech data base employed in the analysis and evaluations is discussed in Section III. Section IV presents the modeling and HMM training of speech parameter variations between neutral and stressed speech. The models characterizing the data are presented and discussed in Section V. In Section VI, the HMM perturbation models are employed to generate perturbation vectors for modifying neutral speech. Speech perturbation or, equivalently, speaking style modification, is discussed in detail in Section VII. A description of the generated stressed synthetic speech and its application to stressed speech recognition are presented in Section VIII. In Section VIII, we also present subjective listener evaluation results of the generated stressed synthetic speech, and objective evaluations using a stressed speech classifier. Finally, in Section IX, we summarize and draw conclusions from our study.

II. PREVIOUS APPROACHES TO STRESSED SPEECH SYNTHESIS AND RECOGNITION

A limited number of studies have integrated stressed speech variations in speech synthesis systems to improve the naturalness of synthetic speech [6]–[9]. Previous approaches directed...
at integrating emotion in text-to-speech synthesis systems have concentrated on formulating a set of fixed rules to represent each emotion. However, analysis studies on emotion and stress suggest that using a fixed set of rules would ultimately represent merely a single caricature of speech variations under a certain emotional condition rather than representing the range of variations for continuous speech that may exist under stress. A stressed speech parameter modeling and perturbation scheme based on a code-excited linear prediction (CELP) vocoder was previously employed for speaking style modification of neutral speech [10]. While the speech parameter perturbation within a CELP framework was effective and successful based on a formal listener assessment, the approach was text-dependent and restricted to the vocoder’s framework.

A number of studies have been suggested for improving recognition of speech under stress [1], [3], [4], [11]–[15], since the performance of a speech recognition system degrades if the recognizer is not trained and tested under similar speaking conditions. An approach referred to as multistyle training by Lippmann et al. [4] has been suggested for improving speaker-dependent recognition of stressed speech. This method required speakers to produce speech under simulated stressed speaking conditions and employed these multistyles within the training procedure. In addition to improving stressed speech recognition, this study showed that multistyle training also improved recognition performance under normal conditions by compensating for normal day-to-day speech variability. However, a later study by Womack and Hansen [16] showed that multistyle training actually degrades performance if employed in a limited but speaker-independent application. Hansen and Clements [3] proposed compensating for formant bandwidth and formant location in the recognition phase. Though the recognition performance improved, such compensation required knowledge of phoneme boundaries and is computationally expensive. Other front-end modifications have also been proposed that normalize the spectral characteristics of stress speech in the recognition phase so that stress speech parameters resemble neutral speech [1], [13], [14]. In Chen’s compensation [1], the impact of stress is assumed to remain constant across an entire word interval, resulting in a fixed whole word compensation stress vector. In the approach proposed by Hansen and Bria [14], three different maximum likelihood (ML) compensation vectors for voiced, transitional, and unvoiced speech sections were employed. In a subsequent study by Hansen [13], stress compensation was performed on an ML voiced/transitional/unvoiced source generator sequence. All of these methods modified the spectral characteristics of input stress speech tokens at the test phase such that the input stress word parameters resembled those from neutral speech. Each of these methods resulted in improved recognition performance; however it should be noted that as the level of compensation at the recognition phase becomes more complex, the computational requirements can become demanding. An alternative technique by Bou-Ghazale and Hansen, which also employs the source generator framework, turns the stress compensation around by generating simulated stressed tokens which are used for training a stressed speech recognizer [12], [17]. Generating simulated stress data in the training phase rather than compensating for the effect of stress in the recognition phase results in a computationally faster recognition algorithm. In the latter approach, both duration and spectral content (i.e., mel-cepstral parameters) were altered to statistically resemble a stressed speech token.

In this paper, an approach similar to the token generation training method is proposed and shown to improve stressed speech recognition by using the generated stressed synthetic speech as training data [18]. The proposed method offers many advantages in that it is speaker and text-independent. This method is discussed in more detail in Section VIII-D.

### III. Speech Data Base

The speech data employed in this study are a subset of the Speech Under Simulated and Actual Stress (SUSAS) data base [2]. Approximately half of the SUSAS data base consists of styled data (such as normal, angry, soft, loud, slow, fast, clear) donated by the Lincoln Laboratories [4], and Lombard effect speech. Lombard effect speech was obtained by having speakers listen to 85 dB sound pressure level (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 that are highly confusable. Examples include /go-oh-nol/, /wide-white/, and /six-fix/. This data base has been employed extensively in the study of how speech production and recognition varies when speaking under stressed speech conditions. For this study, a vocabulary size of 29 words was used. Twelve tokens of each word in the vocabulary were spoken by nine native American English speakers for the neutral conditions, and two tokens for each style condition.

### IV. Speech Parameter Modeling and Training

As shown in Fig. 2, the following five separate perturbation models are obtained for each stress condition:

1) voiced duration variation;
2) pitch contour perturbation;
3) derivative of pitch contour perturbation;
4) explicit state occupancy for pitch-perturbation HMM;
5) average spectral mismatch.

All models are word-independent, given that an input word under test contains a single continuous voiced region (i.e., with no loss of generality, we limit our focus here to test utterances with a single voiced region, since further consideration of pitch contour modeling is needed to consider multiple voiced region speech utterances). Voiced duration variation, pitch-perturbation derivative, and state occupancy are modeled using probability mass functions (PMF) while pitch and spectral contour perturbations are modeled via HMM’s. The HMM can properly model the essential structure of the pitch-perturbation profile and its variations. However, when pitch-perturbation observations are regenerated from the pitch-perturbation HMM, these observations should be ordered to lead to an appropriate pitch-perturbation profile that reflects the time evolution of the training data. In order to properly order the pitch-perturbation observations, the derivative of
these observations is modeled as well. A detailed description of speech parameter training for pitch contour, spectral contour, and voiced duration follows.

A. HMM-Based Pitch Contour Training

Two studies have previously employed HMM-generated pitch contours for speech synthesis [19], [20]. The study presented here is the first HMM application to model variations in speech parameters for the purpose of stressed speech synthesis from neutral speech. The proposed work here, however, differs from these previous approaches in that

1) a single 3-state HMM is used for modeling the whole pitch perturbation, therefore typical subunit model concatenation is not necessary;
2) interpolation between observations or normalization of the generated contour are not required;
3) our approach makes no explicit use of the phonemic environment.

Also, note that in this approach the pitch perturbation HMM is trained with pitch perturbation contours as opposed to actual pitch contours. The advantage is that the pitch perturbation HMM can be applied to new speakers, since this scheme increases or decreases a speaker’s pitch according to the pitch perturbation vector as opposed to imposing a particular speaker’s pitch contour onto the input speaker.

A three-state single-mixture pitch-perturbation HMM is trained for each stressed condition. Each model is trained with 6264 pitch perturbation profiles. Pitch perturbation training contours are generated as follows. As shown in the modeling flow diagram of Fig. 2, two pitch contours are computed simultaneously for a neutral and stressed word (same speaker, same text). The duration of the stressed pitch profile is then time-scaled to match the neutral pitch contour. A pitch perturbation profile, $P_r$, is then computed as the ratio of the time-scaled stressed pitch contour to the neutral pitch contour. Next, the derivative of the pitch perturbation profile is computed over five frames. The set of pitch perturbation contours are modeled via an HMM, while the pitch perturbation derivative is modeled using a PMF distribution. The PMF distributions of the pitch perturbation derivative indicate that the initial slope of the pitch perturbation profile is always positive. The pitch perturbation derivative PMF is later used
in Section VI-A for ordering the pitch perturbation values generated by the first HMM state. In order to employ the regenerative feature of HMM’s, it is necessary to model the state duration. The explicit state duration modeling of the pitch-perturbation HMM is presented next.

1) Explicit State Occupancy Modeling for Pitch-Perturbation HMM: An extensive treatment of state duration modeling can be found in the work of Ferguson [21]. The inherent state occupancy probability density \( p_i(d) \) of \( d \) repetitions of state \( i \), with self-transition coefficient \( a_{ii} \) is of the form

\[
p_i(d) = a_{ii}^{d-1}(1 - a_{ii}).
\]

However, this implicit geometric, exponentially decaying, state duration density is almost always inappropriate for speech signal representation. As a result, other parametric representations of state duration occupancy have been proposed [22], [23]. However, the cost of incorporating a state duration density into the HMM framework is rather high. Thus, we have formulated an alternative procedure for modeling state duration in HMM’s using nonparametric distributions. A similar approach for modeling the state duration distributions is described in [24]. Here, the state duration probability is measured directly from the training sequences as follows. The set of training observation sequences, \( O = (O_1, O_2, \cdots, O_T) \), is segmented into states, \( Q = (q_1, q_2, \cdots, q_L) \), based on the HMM model \( \lambda_p \). This segmentation is achieved by finding the optimum state sequence via the Viterbi algorithm that maximizes \( P(O | \lambda_p) \). Normalized duration \( (100 \times d/T) \) is used for modeling state occupancy rather than absolute duration. Using normalized duration rather than absolute durations addresses the issue of a vocabulary with different word lengths.

To illustrate state occupancy modeling of the pitch perturbation HMM, let the random variable \( X \) represent the time duration percentage spent in state 1. The PMF of the random variable \( X \) is represented here as three separate events. The first event, denoted as \( [X = 0] \) with probability \( p_0 = P(X = 0) \), represents the probability of spending 0% time in state 1 (i.e., skipping state 1 altogether). The second event, denoted as \( [0 < X < x] \), represents the PMF of \( X \), \( \forall X \in (0, 100) \). In this case, the percent time duration \( X \) can take any value greater than zero and less than 100. The third event represents the case where 100% of the time is spent in state 1 and is denoted by \( [X = 100] \) with probability \( p_{100} = P(X = 100) \). The PMF of state 1 is summarized as follows:

\[
p_X = \begin{cases} 
p_X(0) & = P(X = 0), \\
p_X(x) & = P(X = x), \\
p_X(100) & = P(X = 100) \
\end{cases}
\]

with the condition that \( \sum_i p_X(x_i) = 1 \).
Let the random variable $Y$ represent the percentage of time spent in state 2. This is represented as two separate conditional PMF’s that depend on the values spanned by the random variable $X$ of state 1. These PMF’s are given by

$$p_Y|X=0 = p_Y[y|p_X = p_X(0)]$$
$$p_Y|0<X<100 = p_Y[y|p_X = p_X(x)]$$

where $\sum_{i} p_Y|X=i(y_i) = 1$. The percent of time spent in the second state depends on the percent of time spent in the first state. Also, the time percentage duration spent in state 3 is dependent on the percentage duration spent in the previous two states. The random variable $Z$ represents the percentage duration spent in state 3 and is determined by $Z = 100 - X - Y$. Therefore, it is not necessary to explicitly derive the PMF equations for state 3. Having trained the models associated with pitch contour, the next step is to train HMM’s for characterizing spectral contour variations.

**B. HMM-Based Spectral Contour Training**

Three spectral mismatch HMM models are trained for each of the stressed conditions. Each model represents a different voiced-to-unvoiced concentration in an utterance. Since voiced and unvoiced phonemes are affected differently under stress, it is more accurate to devise different frequency modification models which depend on the concentration of voiced and unvoiced regions in the word. Spectral contour mismatches for HMM training are obtained as follows. First, a spectral contour estimate is obtained for a neutral and a stressed utterance (same speaker, same text) by computing a second-order least squares estimate of the spectrum on a frame-by-frame basis, as shown in Fig. 3. An average spectral contour is then computed across the whole utterance for each neutral and stressed input token. The spectral mismatch between a neutral and stressed word is obtained by calculating the frequency difference between the two average spectral contours. The spectral mismatch profiles are quantized at equally spaced bins of 500 Hz over a 4 kHz bandwidth. Depending on the voiced-to-unvoiced concentration of the input neutral word, the quantized data is then used for training the appropriate one-state, single mixture, nine-parameter Gaussian HMM. A second-order estimate of the spectrum is used in order to capture the general spectral variations that occur under a stress condition as opposed to modeling the detailed spectral structure which may be specific to a certain word or speaker.

**C. Voiced Duration Modeling**

The duration modeling consists of representing the duration variation present in the voiced regions between neutral and stressed speech, as shown in Fig. 2. This variation is modeled as the ratio of stressed-to-neutral voiced duration. These ratios are then used to construct a voiced duration perturbation PMF. Once the PMF has been constructed, all scaling values greater than 2 were set back to the value 2, and scaling values less than a factor of 0.5 were set equal to 0.5. These two constraints are imposed by the speech quality limitations of the time-scale modification algorithm.

**V. HMM-Based Perturbation Models**

The HMM’s resulting from pitch perturbation training are given in Fig. 4 for angry, loud, and Lombard effect conditions. The models are Gaussianly distributed, and are plotted as bar graphs where each bar represents a total span of 6σ conditioned on a positive perturbation scaling. Each white center represents the mean of the Gaussian distribution. The plot shows the distribution of all three HMM states. The leftmost bar in each state represents the angry pitch perturbation model, the middle bar represents the Lombard perturbation model, and finally, the right-most bar represents the loud pitch perturbation model. These models show that the required pitch variation from neutral at the beginning of an utterance is larger under angry and loud conditions than under Lombard
effect (almost double the variation). At the closing of an utterance, the required pitch variation is wider for Lombard than for angry or loud, suggesting that speakers on the average will increase the variability of pitch under Lombard effect. In summary, these models show that under angry and loud conditions the required mean pitch perturbation and variance is large at the beginning with a reduction in the necessary pitch modification at the end of an utterance. For Lombard effect, the required pitch perturbation varies less at the beginning of an utterance and experiences a wider variation at the end. Therefore, it can be said that the HMM pitch perturbation model reflects both the mean pitch shift under stress, as well as an estimate of the change in pitch profile shape under stress.

Next, the spectral perturbation models for a voicing degree greater than 50% are shown in Fig. 5 for angry, loud, and Lombard effect. The spectral mismatches are similar for angry and loud, but are different for the Lombard effect. The needed spectral variations for loud and angry conditions increase at higher frequencies. As voice level increases, the energy in high-frequency components are raised much more than for the low-frequency components. The average spectral perturbation decreases for frequencies below 500 Hz for loud and angry conditions. The drop in energy at low frequencies is supported by William and Stevens [25] as follows. The low-frequency bands are in a frequency region occupied by the fundamental frequency $F_0$ for voiced speech. If $F_0$ is low, then it remains in the low-frequency band, and hence there is appreciable energy at lower frequencies. If $F_0$ is high, then there is less energy at low frequency. Therefore, the energy levels at low frequencies provide a rough indication of the average fundamental frequency. Under the Lombard effect, the spectral perturbation mean varies across frequencies, while the variance is almost constant across all frequency values.

The last area for stress perturbation modeling is duration. The variation in required duration perturbation under angry conditions is more uniformly distributed and wider than for loud, or Lombard. This indicates that under angry conditions, a voiced section can increase or decrease in duration depending on its relative position in the utterance. The duration scaling distributions for loud and Lombard are more Gaussianly distributed. It is also noted that the distributions become slightly peaked for duration modification of 2, since due to constraints in the resulting speech quality for the speech duration modification algorithm, the duration shift is limited to a maximum of 2.

VI. HMM-BASED SPEECH PARAMETER REGENERATION

The models developed in the previous section are employed to generate perturbation vectors to be used for modifying neutral speech. First, pitch perturbations are statistically generated using the pitch perturbation HMM. In Section VI-A, the voiced duration distribution, pitch perturbation derivative, and state occupancy model are combined in a single algorithm to generate pitch perturbation profiles. In Section VI-B, the method for obtaining the spectral perturbations are described using the one-state nine-parameter spectral mismatch HMM model.

A. HMM-Generated Pitch Perturbations

The HMM-pitch perturbation model is used to generate pitch perturbation contours to impart stress traits onto neutral speech. To achieve this, three steps are proposed. First, the total number of observations to be produced by the HMM is determined; second, the length of time spent in each state is determined; finally, a procedure to order these observations is established. It is important to remember that in HMM modeling, it is assumed that the observation sequence is statistically independent, which for a pitch profile is not desirable. The total number of frame observations to be
produced for an input utterance represents the desired length of the pitch perturbation profile. The desired number of frame observations can be computed by multiplying the initial voiced speech duration of the input neutral speech by the duration scaling factor. The duration scaling factor accounts for the duration variation from neutral to a stressed condition, and is randomly generated from the duration scaling PMF’s that correspond to the desired stress condition to be imparted.

The next step is to determine the number of observations to be produced for each state, or the length of time to be spent in each state, using the state occupancy models formulated in Section IV-A. First, depending on the desired speaking condition, the appropriate probability distribution is sampled to determine the probability of either visiting or totally skipping the first state. Unless the first state is skipped, observations are generated from the first state according to the PMF distribution associated with that state. Next, and observations are generated from the second and third states simultaneously, according to their PMF distributions. At this point, the necessary set of pitch perturbation observations has been produced.

The final step is to formulate a procedure for ordering the perturbation observations. Perturbation observations produced from state 1 are ordered according to the distribution of the initial pitch perturbation derivative. The observations produced in subsequent states are ordered in either an ascending or descending order based on a minimum difference continuity criterion. The ordering is chosen so as to minimize the difference between observations at the boundaries. The proposed pitch ordering is illustrated in Fig. 6, and is governed by the following equations. Assume that the function \( f(x) \) denotes the ordered observations generated at interval \( I - 1 \), or by the HMM state \( q - 1 \). Let the functions \( g_p(x) \) and \( g_n(x) \) represent the ascending and descending ordered observations generated in the subsequent state \( q \) or interval \( I \). The derivative of \( g_p(x) \), \( g'_p(x) \), is positive at every point of the interval \( I \), while the derivative of \( g_n(x) \), \( g'_n(x) \), is negative at every point of the interval \( I \). The objective is to choose one of the two functions \( g_p(x) \) or \( g_n(x) \), in order to minimize the discontinuity at the boundary between the previous and the following interval. The function \( g_p(x) \) that represents the ascending ordered set of observations would be chosen if the following relation is satisfied:

\[
|f(x_{i-1}) - g_p(x_i)| < |f(x_{i-1}) - g_n(x_i)|
\]

Alternatively, a descending observation order is chosen, and hence the function \( g_n(x) \), if the following equation is satisfied:

\[
|f(x_{i-1}) - g_p(x_i)| > |f(x_{i-1}) - g_n(x_i)|
\]

The minimum difference criterion ensures a minimal discontinuity in the pitch perturbation profile during a transition from one state to another. The pitch profile is assumed to be continuous. This constraint is valid here since all utterances under test in the training corpus consist of a single voiced region with no interleaved unvoiced sections (e.g., words such as zero, help, or degree).

B. HMM-Generated Spectral Perturbations

Using the set of single state nine-parameter spectral mismatch HMM models formulated in Section IV-B, a spectral perturbation vector can be obtained by sampling the distribution associated with each parameter. Before this can be accomplished, the degree of voicing of the input neutral word is calculated in order to use the appropriate HMM from the three possible models. Using the selected model, a second-order least squares fit to the generated observation sequence is obtained. This frequency perturbation can be used directly in the frequency domain, or in the time domain by designing an \( n \)th-order finite impulse response (FIR) filter. The advantage of a time-domain filter is that it modifies the phase information along with the magnitude, whereas the frequency domain filtering keeps the phase unmodified. For this application, the spectral perturbation was done directly in the frequency domain as shown in Fig. 8 to prevent any mismatch or error that may result from the time-domain filter design. The frequency-domain perturbation was found to be superior to time-domain filtering based on an informal listener evaluation of the perturbed speech. The spectral perturbation is used to modify the spectral slope as well as the overall energy mismatch.

VII. SPEAKING STYLE MODIFICATION

The HMM-based models are integrated into a single overall algorithm employing pitch, duration, and spectral contour perturbation in order to generate stressed speech from neutral speech. In order to modify the speaking style of an input neutral word, the following steps are required (refer to Fig. 7). The duration of the input neutral word is computed and then multiplied by a duration scaling factor, which is obtained from a random generated output of the duration scaling PMF. This
determines the total length of the pitch perturbation profile to be generated. Next, using the state occupancy model, the number of required observations to be generated from each HMM state is computed. According to these state duration values, the necessary observations are generated from each pitch-perturbation HMM state. These observations are ordered according to the previously described minimal discontinuity criterion to form the pitch perturbation profile. This pitch perturbation profile is then used for perturbing the pitch of the input neutral speech as shown in Fig. 7. The pitch and duration of the input neutral utterance are modified in the time domain within a linear prediction framework. Pitch is modified by linear prediction residual resampling on a frame-by-frame basis, and duration is modified pitch-synchronously by varying the rate of the analysis and synthesis. After LP resynthesis, the spectral contour magnitude is modified in the frequency domain while the phase is kept unchanged as previously shown in Fig. 8.

VIII. Evaluations and Discussions

In these evaluations, the proposed integrated algorithm is first employed to obtain a corpus of neutral modified stressed synthetic speech. Having achieved this, evaluations in three separate areas are performed to demonstrate the effectiveness of the HMM-based perturbation system. The synthetic speech is presented to listeners in a formal listener test, which is described below, to judge the stress content of speech. Next, the stressed synthetic speech is presented to a stress classifier to judge whether our scheme is capable of modifying the neutral speech both perceptually and statistically. This supplements the subjective listener evaluation and provides an objective measure of the stress content. Finally, the generated stressed speech is used as training data for a stressed speech recognizer to assess the ability of improving its performance by training with the generated stressed speech tokens. These evaluations will help illustrate the performance and ability of our proposed HMM modeling scheme to accurately represent variations under angry, loud, and Lombard effect speaking conditions.

For both synthesis and recognition, linear predictive coding based (LPC-based) parameters are used as part of the feature set. LPC based cepstrum (LPCC) and delta cepstrum (ΔLPCC) are derived from the LPC coefficients using the following
equations:

\[ a_k = a_k + \frac{1}{k} \sum_{i=1}^{k-1} i \cdot a_{k-i} \quad 1 \leq k \leq p \quad (1) \]

\[ \Delta a_k(j) = \sum_{n=-l}^{l} n \cdot a_{k+j+n} \quad (2) \]

where \( a_k \) are the LPC coefficients and \( p \) is the LPC analysis order. The index \( l \) specifies the number of frames over which the \( \Delta \)LPCC were calculated \( (l = 2 \) in our experiment). Further details on the features used for classification and recognition are given in Sections VIII-C and VIII-D. A discussion of the synthetic speech follows in Section VIII-A.

**A. Generated Synthetic Speech**

A total of 6480 tokens (24 words/speaker \( \times \) 10 tokens/word \( \times \) 9 speakers/style \( \times \) 3 styles) are synthetically generated for the three stressed speaking conditions (2160 tokens per style). These tokens are generated by perturbing a 29-word vocabulary spoken by a group of nine general American English speakers where each word is repeated ten times. The vocabulary consists of mono- and multisyllabic words such as go, degree, and stand. The perturbation is applied to words that contain one main voiced island (i.e., a continuous pitch profile). In this context, our perturbation algorithm is text and speaker independent, but is unique for each speaking style. The synthetic angry, loud, and Lombard effect speech that resulted from neutral speech perturbation is presented to listeners to subjectively evaluate its stress content. The results are presented in the next section.

**B. Formal Subjective Listener Framework**

The listener test consisted of three separate evaluations. Each evaluation was targeted toward one of the three stressed speaking styles: angry, loud, and Lombard effect. During each of the three evaluations, listeners heard 20 sequences. A sequence consisted of a series of three isolated words spoken by the same speaker under the same speaking conditions. The speaking condition of a sequence was either neutral, synthetic stress, or original stress speech. Each sequence was played only once. Listeners were given instructions both verbally and on the screen of a workstation. However, they were not presented with sample words spoken under original neutral or original stressed speech prior to conducting the listener test. Such a demonstration was omitted in order to avoid the possibility of biasing the listeners opinion or perception of speech under stress. We suspected that if, for example, listeners were presented with tokens of actual angry speech prior to conducting the test, then their perception of angry speech may be influenced, or that we may be imposing a reference for listeners to use in their decision.

The listener test was implemented using an interactive user interface on a computer workstation. Evaluators used high-quality headphones, and were seated in front of the workstation.
in a quiet office. A total of 16 listeners participated in these evaluations, comprising a combination of experienced speech researchers and naive listeners. Listeners consisted of both males and females. American English was the first language of all 16 listeners who had no reported history of hearing loss. After each sequence, listeners were prompted to make one of three choices. For example, when evaluating angry speech, listeners heard a series of three words, and were asked to pick one of the following three choices: i) the speech sounds neutral; ii) the speech sounds angry; or iii) the speech does not sound neutral. The sequences consisted of either neutral, original angry, or synthetic angry speech. All nine speakers in the data base were tested under all stressed conditions. The sequences were presented to listeners in a round robin scheme (i.e., training with eight speakers while reserving a speaker for open testing, in order to test all nine speakers). The training was repeated in a binary stressed/nonstressed decision. This was a difficult task for listeners since they were asked to judge the stress condition of a speaker whom they have never heard before. In other words, listeners had not heard the speaker under either neutral or any other stress condition. Their decision had to be based solely on the sequence of three words without any reference. The results of the listener test are discussed next.

1) Listener Test Results: Several conclusions are drawn from the formal listener evaluations. This listener test allowed one to determine the ability of listeners to identify original neutral and original stressed speech as well as to judge the stress content of synthetic stressed speech. On the average, original neutral speech was identified by listeners as sounding neutral approximately 80% of the time. Listeners correctly identified original angry speech 78.75% of the time (see Table I), original loud speech 75% of the time (see Table II), and original Lombard effect speech 60% of the time (see Table III). Original Lombard effect speech was the least identified by listeners because listeners may not have had a predefined perception of Lombard effect speech. The listener evaluation of synthetic stressed speech clearly demonstrated the performance of the perturbation algorithm. The results showed that only 6.94% of the synthetic angry speech, 9.72% of synthetic loud speech, and 8.33% of synthetic Lombard effect speech were judged as sounding neutral. This indicates the effectiveness of modifying the neutral speaking style. Listener results of synthetic stressed speech exhibited large variances. This was expected due to the nature of this listener test. Listeners had no reference for comparison, and hence had differing preconceived perceptions of stressed speech. The results of the synthetic stressed speech evaluation are tabulated in terms of mean, median, standard deviation, and maximum in Table IV. For example, the synthetic loud speech was judged as sounding loud 34% on average, while the median was 38.89%. The maximum indicates the highest percentage of synthetic loud speech judged as loud by any one listener. The maximum for synthetic loud speech, for example, was 77.78%. In summary, the listener test clearly reflects a movement in the perturbed speech toward the target stress speaking style. In the next section, the synthetic stressed speech is presented to a stress classifier to obtain an objective assessment.

C. Classification of the Generated Stressed Synthetic Speech

As a second evaluation, a stress classifier is employed to assess the generated stressed synthetic speech. Although, it is possible to design a more elaborate classifier based on nonlinear [26], multidimensional [27], or targeted [16] sets of parameters, the main goal here is to provide an independent objective assessment of the perturbed speech rather than to achieve the highest classification rates. Therefore, a speaker-independent stress classifier has been formulated for neutral, angry, loud, and Lombard effect speaking styles. The training consisted of a total of 1536 tokens (24 words/speaker × 2 tokens/word × 8 speakers/style × 4 styles) from eight out of the nine speakers. The training was repeated in a round robin scheme (i.e., training with eight speakers while reserving a speaker for open testing, in order to test all nine speakers). A 64-mixture one-state HMM was trained for each of the four speaking styles: neutral, angry, loud, and Lombard effect. The large number of mixtures is needed to account for the variability that exists among speakers. The speaker-independent HMM’s were trained with eight LPCC

<table>
<thead>
<tr>
<th>Input Speech</th>
<th>JUDGED AS:</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Neutral</td>
<td>Neither</td>
<td>Angry</td>
<td></td>
</tr>
<tr>
<td>Original Neutral</td>
<td>83.33</td>
<td>13.54</td>
<td>3.13</td>
<td></td>
</tr>
<tr>
<td>Synthetic Angry</td>
<td>6.94</td>
<td>75.00</td>
<td>18.06</td>
<td></td>
</tr>
<tr>
<td>Original Angry</td>
<td>5.00</td>
<td>16.25</td>
<td>78.75</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Input Speech</th>
<th>JUDGED AS:</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Neutral</td>
<td>Neither</td>
<td>Loud</td>
<td></td>
</tr>
<tr>
<td>Original Neutral</td>
<td>81.25%</td>
<td>11.46%</td>
<td>7.29%</td>
<td></td>
</tr>
<tr>
<td>Synthetic Loud</td>
<td>9.72%</td>
<td>56.25%</td>
<td>34.03%</td>
<td></td>
</tr>
<tr>
<td>Original Loud</td>
<td>11.25%</td>
<td>13.75%</td>
<td>75.00%</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Synthetic Style</th>
<th>Mean</th>
<th>Median</th>
<th>Standard Deviation</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angry</td>
<td>18.06%</td>
<td>11.11%</td>
<td>21.80%</td>
<td>66.67%</td>
</tr>
<tr>
<td>Loud</td>
<td>34.03%</td>
<td>38.89%</td>
<td>23.47%</td>
<td>77.78%</td>
</tr>
<tr>
<td>Lombard</td>
<td>50.70%</td>
<td>44.44%</td>
<td>24.33%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Table I: Listener Evaluation Results of Neutral, Synthetic Angry, and Original Angry Speech

Table II: Listener Evaluation Results of Neutral, Synthetic Loud, and Original Loud Speech

Table III: Listener Evaluation Results of Neutral, Synthetic Lombard, and Original Lombard Effect Speech

Table IV: Detailed Listener Evaluation Results of Synthetic Angry, Synthetic Loud, and Synthetic Lombard Effect Speech. The first column tabulates the average in percent that a synthetic stressed style is correctly identified by listeners. The second column represents the median value. The standard deviation in the third column indicates the variation in listeners’ judgment. The last column indicates the highest percent identification of a style given by a listener.
parameters, one normalized-energy, and one normalized-pitch. The energy and pitch were normalized to ensure similar variances in the training parameters. The classifier was not trained with delta parameters since these were found to degrade a stress classifier performance [16].

To test the performance of the classifier, a pairwise comparison test is conducted to differentiate between original neutral and one stressed condition at a time. The test classifies original neutral and stressed speech, and establishes a reference for the following evaluations. The classifier results for unmodified neutral and stressed speech are presented below. Then the classifier is used in a pairwise comparison mode to evaluate the generated stressed synthetic speech.

1) Classifier Performance: A total of 1728 tokens (24 words/speaker × 2 tokens/word × 9 speakers/style × 4 styles) are classified in each test. The pairwise comparison results are presented in Tables V–VII. In a pairwise comparison between one stressed speaking style and neutral, the correct classification rates are 93.29% for angry, 97.00% for loud, and 96.53% for Lombard effect speech. In the same pairwise comparison test, the neutral speech is correctly classified 96.53% when compared to angry, 97.00% when compared to loud, and 96.53% for Lombard effect speech. In the next section, the classifier is used to classify the generated stressed synthetic speech.

2) Synthetic Stressed Speech Classification: All the synthetic speech evaluations are speaker-independent (i.e., the models have not been trained with any speech from the input test speaker). A corpus of 6480 (24 words/speaker × 10 tokens/word × 9 speakers/style × 3 styles) synthetic stressed tokens were classified from all 3 stressed conditions. The perturbed speech was classified without any prior screening or testing. The synthetic angry speech was classified 65.34% as angry. The loud synthetic speech was classified 53.51% as loud. Finally, the Lombard synthetic speech was classified 46.82% as Lombard. A comparison of the classification results for the original neutral speech before and after perturbation is given in Fig. 9. For example, the original neutral speech, which was initially classified 96.53% as neutral, and 3.47% as angry, was classified 65.34% as angry after it had been perturbed. The results indicate that the perturbation was able to move 61.87% of the original neutral tokens into the angry domain (since 3.47% of the neutral tokens were already classified as angry). Comparable results were achieved for loud and Lombard effect speech. The neutral speech was initially classified 96.30% as neutral, and 3.7% as loud. After perturbation, 53.51% of the synthetic loud speech was classified as loud. The synthetic Lombard speech was classified 46.82% as Lombard.

The classification results presented so far clearly demonstrate the effectiveness of our perturbation algorithm. However, the classification rates of stressed synthetic speech can be increased. Since the perturbation scheme is capable of generating a number of different perturbed versions of the same input word, the classifier can therefore be used as a screening system to accept only those perturbed tokens which are classified as stressed. If so, then a data base of speech can be generated which possesses a high probability of characterizing actual stressed speech. Such a system is implemented here and is referred to as the “recursive synthesis” method.

In this scheme, input neutral speech is perturbed and then classified. Unless the perturbed word is classified as stress, the perturbation/classification procedure is repeated either until the
token is classified as being under stress or for a maximum of ten iterations, whichever occurs first. The average results over the entire word and speaker set are plotted in Fig. 10. Using the "recursive synthesis" approach, the classification rates of the synthetic stressed speech increased from 65.34% to 94.05% on average for synthetic angry speech. The classification of synthetic loud speech increased from 53.51% to 86.60%, while for synthetic Lombard speech it increased from 46.82% to 82.42%. This ability to generate an unlimited number of speech tokens characterizing a wide range of stress levels clearly demonstrates the strength of our modeling approach.

The modification of a neutral speech token is not simply a fixed perturbation vector. In fact, the same token can be modified to characterize stress conditions ranging from mild to severe.

D. Application to Stressed Speech Recognition

It has been well documented that the variability introduced by a speaker under stress causes recognizers trained with neutral tokens to fail [1]–[4]. Unlike the human auditory system, which is capable of extracting this variability as additional perceptual information of the speaker (i.e., emotion, situational speaker state), typical recognition algorithms do not attempt to extract this information and cannot overcome such speaking conditions.

It is desirable to improve stressed speech recognition by training an HMM-based speech recognizer using the generated stressed synthetic speech. The goal here is to investigate whether the stressed synthetic speech possesses sufficient stress characteristics that are capable of improving stressed speech recognition. Another advantage of training with stressed synthetic speech is that a potentially much larger number of training tokens would be readily available. This is due to the HMM regenerative property, which can be used to produce a large number of vectors that are used to perturb neutral speech, and hence generate stressed synthetic speech. This eliminates the need for collecting stressed tokens for training.

The next section discusses the HMM topology, and the feature set used for training the recognizer. The recognition evaluations are divided into four parts. The first part presents the performance of neutral trained HMM’s when tested with stressed speech. The second part presents recognition results of models trained and separately tested with original stressed speech. The third evaluation addresses the advantages of training with the corpus of synthetic stressed speech. Finally, the last evaluation studies the effects of using pitch as part of the feature set on the recognition performance of stressed speech.

1) HMM Training: In this study, all recognition evaluations were speaker independent, and considered only male speakers. A 29-word HMM-based recognizer was formulated using a variable state, left-to-right model, with two continuous mixtures per state. Two different sets of HMM’s are formulated here in order to evaluate the effect of using pitch as a feature in stressed speech recognition. Pitch is not normally used in speaker-independent speech recognition but has been used in text-dependent speaker recognition systems [28]. The common features used in both sets are eight LPCC, ΔLPCC, energy, and Δenergy. However, one set of models is trained with the additional parameters of pitch and Δpitch to investigate the effect of pitch on stressed speech recognition.

The HMM stressed models were trained with the corpus of perturbed speech of eight speakers, while the ninth speaker was left for open testing. A total of ten tokens per speaker were used for each neutral word, resulting in 80 training tokens per word for the neutral models. The training and testing were done in a round robin scheme to allow all speakers and tokens to be open tested. The neutral models were trained with 80 actual neutral tokens per word, while the actual stress models were trained with 16 tokens per word, representing all the available data.

2) Effects of Stress on Neutral Trained Models: The recognition performance of the speaker-independent neutral trained recognizer is 92.13% when tested with neutral speech. This is shown in Fig. 11 as the top left bullet. When neutral trained HMM’s are tested with angry, loud, and Lombard speech, recognition performance drops to 78.01% for angry, 81.25% for loud, and 89.35% for Lombard effect as illustrated by the lower dotted line of Fig. 11. These results therefore confirm earlier studies which show that stressed speech adversely impacts recognition performance.

3) Original Stressed Trained HMM Models: Next, we see that recognition of stressed speech improves when style-dependent models are used for recognition, as indicated by the

![Diagram](image-url)
Fig. 11. Performance of a speaker-independent, but style-dependent speech recognizer. This table compares the performance of a neutral trained model to six other models, three of which were trained individually with one of the three stressed styles, specifically angry, loud, and Lombard effect. The three remaining models (in the third column) were trained with synthetically generated stressed speech. The parameters used both for training and testing were eight LPCC, scaled energy, and delta parameters.

The top solid line of Fig. 11. A style-dependent model refers to a model trained and tested with speech from the same speaking style. For example, the recognition rate of an angry trained and tested model is 85.19%. This represents an improvement of 7.18% over the neutral trained model. However, it is still measurably lower (6.94%) than the recognition rate of neutral speech. The recognition rate for loud speech is 90.05% when HMM models are trained and tested with loud speaking conditions. This represents 8.8% improvement over neutral trained models. Finally, Lombard effect speech recognition is improved by 3.24% when Lombard trained models are used as opposed to neutral trained models, achieving a recognition rate of 92.59%, which is the same rate obtained for neutral speech recognition. These results show that style-dependent training improves stressed speech recognition. However, stressed speech data is not always available for training. Therefore, when the training corpus consists of only neutral speech, the latter could be modified using the HMM perturbation models to generate stressed synthetic speech which is then used for training. This approach is discussed in the next section.

4) Synthetic Stressed Trained HMM Models: When original stressed data is not readily available for training, we propose training with synthetic stressed speech generated by perturbing the available neutral speech data. The hypothesis is that stressed synthetic trained models could outperform the neutral trained models for stressed speech recognition. However, there is a limiting upper bound as to the extent of improvement that can be achieved. It is suggested that this upper bound is determined by the recognition rates achieved by the original stress trained models. Therefore, the maximum improvements that could be attained are 7.18% for angry, 8.8% for loud, and 3.24% for Lombard effect. When the corpus of synthetic angry speech is used for training, recognition of angry speech is improved by 3.51%. This represents 51% of the possible total improvement that could be attained. Training with loud synthetic speech improves loud speech recognition by 4.4%, which is 50% of the total possible improvement. Synthetic Lombard trained models do not contribute to improved recognition of Lombard effect speech, though the change in noise-free Lombard recognition is less than for the other two stress conditions. The results for synthetic stress trained models are shown as the middle dotted line of Fig. 11. This figure illustrates the overall average recognition results across all nine speakers.

In order to study the performance effects of pitch on our stressed speech recognizer, a second set of models was trained with the same feature set used in the previous evaluations but with the addition of pitch. These recognition results are shown in Fig. 12. The performance of a neutral trained speaker-independent speech recognizer tested under neutral
conditions is the same whether or not pitch is included in the feature set. However, including pitch as a feature does significantly reduce the recognition performance of stressed trained as well as neutral trained models when tested with stressed speech. This was particularly true for angry speech, which suffered an additional 9.95% loss in recognition rate when pitch information was included. It is believed that the reason for this change is due to the increased irregularity of pitch structure under angry speaking styles. However, training with synthetic stressed speech still improves stressed speech recognition regardless of whether or not pitch is used as part of the feature set. When training with the synthetic speech, improvements over neutral trained models of 13.19%, 6.02%, and 2.32% were achieved for angry, loud, and Lombard effect testing conditions. The synthetic angry trained models achieved the same performance as the original angry-trained models. It is true that the average recognition performance drops when including pitch in the feature set. However, this is not true across all speakers. In fact, the synthetic angry trained models (with pitch) achieved significant improvements over the models trained with no pitch information for three of the nine speakers. For loud speech, the same is true. For Lombard effect speech, including or omitting pitch has no significant effect on the recognition performance.

In summary, we can conclude that using knowledge of how other speakers modify their speech under stress, and using this for perturbation of neutral training data can measurably increase recognition performance under stressed conditions.

**IX. CONCLUSIONS**

In this study, we have proposed and implemented a novel modeling approach for speech parameter variations under stress using HMM’s. The variations in overall pitch contour, voiced duration, and overall spectral contour were modeled for angry, loud, and Lombard effect speech. The models were trained with the variations, referred to as perturbations, in speech parameters from neutral to each stressed condition as opposed to training with actual speech parameters. A pitch-perturbation model was developed for each stressed condition using a three-state single-mixture Gaussian HMM. Three spectral mismatch models were trained for each stressed condition using a one-state, single mixture, nine-parameter Gaussian HMM. Each of the three models represents a different voiced-to-unvoiced concentration in an utterance. Voiced duration variation was modeled using one PMF for each stressed condition.

The pitch perturbation models suggest that under angry and loud conditions, the beginning of an utterance is produced with greater emphasis than the closing of an utterance. Under Lombard effect, greater emphasis is placed at the end of an utterance.

The spectral perturbations for loud and angry are very similar. The required average spectral perturbation increases for frequencies above 500 Hz for all three stressed conditions. The average spectral perturbation decreases for frequencies below 500 Hz for loud and angry conditions, and remains almost at 0 dB under Lombard conditions. The variation in spectral perturbation at low frequencies is lower than at higher frequencies for loud and angry. This implies that angry and loud conditions are manifested in a higher subglottal pressure and a narrower glottal pulse. Under Lombard effect conditions, the variation in spectral perturbation is almost constant across all frequencies.

The highest average duration variation is for loud speech, while the variation in the required duration perturbation is wider for angry speech. The duration perturbation under angry conditions is almost uniformly distributed between factors of 1 and 1.6, while it is more peaked for loud and Lombard effect.

The HMM perturbation models were employed for generating perturbation vectors used for modifying neutral speech. The resulting stressed synthetic speech was evaluated in three separate areas. First, the synthetic speech was presented to listeners in an effort to judge its stress content. Next, the synthetic speech was evaluated by a stress classifier for an objective assessment. Finally, the synthetic speech was used in training an HMM-based stressed speech recognition system in order to improve stressed speech recognition. The listener evaluation results indicated that the perturbed speech was achieving the desired stressed style. The listener results, however, exhibited large variances. Classification results of the stressed synthetic speech were very satisfactory, and proved the effectiveness of the perturbation algorithm in modifying the speech not only perceptually but statistically as well. Training with the generated stressed synthetic speech proved to be an effective method for improving stressed speech recognition. The algorithm and results presented in this study suggest that effective stress modification algorithms can be formulated that successfully impart stress or emotion onto neutral speech for improved performance as judged by human listeners for speech synthesis, as well as improved robustness for speech recognition application.

**REFERENCES**


Sahar E. Bou-Ghazale was born in Baabda, Lebanon. She received the B.S.E.E. degree (honors) from the University of North Carolina, Charlotte, in 1991, and the M.S. and Ph.D. degrees in electrical engineering from 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, where she worked on robust features for improving speech recognition under actual stressful environments. In November 1997, she joined the Speech, Audio, and Video Team of the Personal Computing Division, Rockwell Semiconductor Systems, Newport Beach, CA. Her research interests include modeling, synthesis, and robust recognition of speech under adverse conditions.

Dr. Bou-Ghazale has served as President of the North Carolina Delta Chapter of Tau Beta Pi in 1990. She 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 PhiEtaSigma, Tau Beta Pi, and Phi Kappa Phi.

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

In 1988, he joined the faculty of the Department of Electrical Engineering, Duke University, Durham, NC, where he established the Robust Speech Processing Laboratory. He has since received a secondary appointment in the Department of Electrical and Computer Engineering. Prior to joining the Duke faculty, he was employed by the RCA Solid State Division, Somerville, NJ, from 1981 to 1982, and Dranetz Engineering Laboratories, Edison, NJ, from 1978 to 1981. He has served as a technical consultant to industry and the U.S. government, including AT&T Bell Laboratories, IBM, VeriVoice, the U.S. Department of Defense, and Sparta in the areas of voice communications, wireless telephony, robust speech recognition, and forensic speech analysis. His research interests span the areas of digital signal processing, analysis and modeling of speech under stress or pathology, speech enhancement and feature estimation in noise, robust speech recognition with current emphasis on robust speech feature enhancement for voice communications, and source generator based speech modeling for robust recognition in stress, noise, and Lombard effect. He is the author of numerous journal and conference papers in the field of speech processing and communications, and is coauthor of the textbook *Discrete-Time Processing of Speech Signals*, (Englewood Cliffs, NJ: Prentice-Hall, 1993).

Dr. Hansen was an invited tutorial speaker for ICASSP’95 and the ESCA-NATO Speech Under Stress Research Workshop, Lisbon, Portugal. He has served as 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 ICASSP-96, and has served as Associate Editor for IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING since 1992; he served as Guest Editor of the October 1994 special issue on Robust Speech Recognition for the TRANSACTIONS. He was the recipient of a Whitaker Foundation Biomedical Research Award in 1993, a National Science Foundation’s Research Initiation Award in 1990, and has been named a Lilly Foundation Teaching Fellow.