N-Channel Hidden Markov Models for Combined Stressed Speech Classification and Recognition

Brian David Womack and John H. L. Hansen, *Senior Member, IEEE*

***Abstract*— Robust speech recognition systems must address variations due to perceptually induced stress in order to maintain acceptable levels of performance in adverse conditions. One approach for addressing these variations is to utilize front-end stress classification to direct a stress dependent recognition algorithm which separately models each speech production domain. This study proposes a new approach which combines stress classification and speech recognition functions into one algorithm. This is accomplished by generalizing the one-dimensional (1-D) hidden Markov model to an N-channel hidden Markov model (N-channel HMM). Here, each stressed speech production style under consideration is allocated a dimension in the N-Channel HMM to model each perceptually induced stress condition. It is shown that this formulation better integrates perceptually induced stress effects for stress independent recognition. This is due to the sub-phoneme (state level) stress classification that is implicitly performed by the algorithm. The proposed N-channel stress independent HMM method is compared to a previously established one-channel stress dependent isolated word recognition system yielding a 73.8% reduction in error rate. In addition, an 82.7% reduction in error rate is observed compared to the common one-channel neutral trained recognition approach.**

***Index Terms*— Lombard effect, N-channel Markov model, speech recognition, stress classification.**

I. Introduction

In THE formulation of algorithms for classification and recognition of speech under stress, it may first be useful to distinctly define stress in our context. Stress can be defined as any condition that causes a speaker to vary speech production from neutral conditions. If a speaker is in a "quiet room" with no task obligations, then the speech produced is considered *neutral*. With this definition, two stress effect areas emerge: perceptual and physiological. Perceptually induced stress results when a speaker *perceives* his environment to be different from "normal" such that his *intention* to produce speech varies from *neutral* conditions. The causes of perceptually induced stress include emotion, environmental noise (i.e., the *Lombard effect*[1]), and actual task workload (e.g., a pilot in an aircraft cockpit). Physiologically induced stress is the result of a *physical impact* on the human body that results in deviations from neutral speech production *despite intentions*. Causes of physiological stress can include vibration, G-force, drug interactions, sickness, and air density. In this study, the following four perceptually induced stress conditions from the SUSAS database [10] (see the evaluations in Section III) are considered: *angry, clear, Lombard,* and *neutral*.

Stress classification is an automatic means of detecting the presence of perceptually induced speaker stress in an utterance. Stress directed recognition relies upon a stress classifier to detect the type of stress in an unknown utterance and to direct a codebook of stress dependent recognizers. Such a recognition system employs a stress dependent recognizer that is trained with data spoken under only one stress class. Hence, such an approach is better able to model the unique set of characteristics common to that stress condition. Until recently, the problems of stress classification [3], [12], [24], [25] and recognition of speech under stress [7], [11], [13] were never considered simultaneously.

For the problem of stress classification, there are two major application areas: 1) objective stress detection/assessment and 2) improved speech processing. Objective stress assessment is applicable to stressed speech token generation and stress detection applications. For example, a stress detector could direct highly emotional telephone calls to a priority operator at a metropolitan emergency service. Speaker stress assessment is therefore useful for applications such as emergency telephone message sorting and aircraft voice communications monitoring. A stress classification system could also provide meaningful information to speech algorithms for recognition, speaker verification, synthesis, and coding.

The main problem we address in this study is to achieve simultaneous stress classification and speech recognition of speech produced under perceptually induced stress. The substantial degradation in speech processing performance due to speaker stress has been well documented in a number of studies [8], [11], [14], [24], [26], [27], [30]. The motivation here is to formulate an algorithm that can detect the type of speaker stress and address this effect in the speech recognition task.

Manuscript received October 9, 1996; revised January 12, 1999. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Mazin Rahim.

B. D. Womack was with the Robust Speech Processing Laboratory, Center for Spoken Language Understanding, University of Colorado, Boulder, CO 80302 USA. He is currently with Speech Technologies Corporation, West Valley City, UT 84118 USA (e-mail: Brian@Speech-Tech.com).

J. H. L. Hansen is with the Robust Speech Processing Laboratory, Center for Spoken Language Understanding, University of Colorado, Boulder, CO 80309 USA (e-mail: jhlh@cslu.colorado.edu).

Publisher Item Identifier S 1063-6676(99)08076-1.

[1] The Lombard effect is the manner in which a speaker attempts to modify speech characteristics in an effort to improve human voice communication intelligibility in a noisy environment [7], [14], [18].

1063–6676/99$10.00 © 1999 IEEE

This study considers the problem of stress independent speech recognition using a multichannel hidden Markov model (N-channel HMM). The motivation for this study is to improve the robustness of speech recognition systems for speech under perceptually induced stress. The formulation and application of the N-channel HMM is presented in Section II; and, in Section III, evaluations of two potential applications of the N-channel HMM are considered for stress classification and improved stressed speech recognition. Finally, a summary of findings and recommendations for future studies are presented in Section IV.

A. Past Research Studies on Stress

A number of studies have been conducted on the analysis of speech under stress in an effort to identify meaningful relayers of stress [22], [23]. Unfortunately, many research findings at times disagree, due in part to the variation in the experimental design protocol employed to induce stressed speech and also due to differences in how speakers impart stress in their speech production. Past research experience suggests that no simple relationship exists to describe these changes [6], [8], [12]. Studies directed specifically at robust speech recognition have addressed intraspeaker variations via speaker adaptation [5], [15], front-end stress compensation [7], [8], [11], or wider domain training or token generation training sets [2], [9], [17]. While speaker adaptation techniques can address the variation across speaker groups under neutral conditions, they are not in general capable of addressing the variations exhibited by a given speaker under stressed conditions. Front-end stress compensation techniques such as morphological constrained feature enhancement with adaptive cepstral compensation (MCE-ACC) [7] employ adaptive cepstral compensation to address stress, and, morphologically constrained feature enhancement to address noise for improved recognition performance in noisy stressful environments. Finally, larger training sets have been considered for stressed speech in the training phase. Most notably, the multistyle training algorithm [17] has shown performance improvement for speaker *dependent* systems. An extension of multistyle training based on stress token generation from neutral training data has also shown improvement in stressed speech recognition [9]. However, for multispeaker systems, it has been shown that multistyle training results in a loss of performance over a neutral trained system [24], [26]. The cause of this is believed to be the additional stress-related interspeaker feature variations that the recognition models must now represent, resulting in a decrease in the discrimination ability across the vocabulary set. Additionally, a previous study did employ a stress directed speech recognition approach using neural networks, which was shown to provide a $+10.1\%$ improvement over conventionally trained neutral speech models [26].

The effects of stress have been indirectly addressed by formulating a more accurate speech production representation of intraspeaker variability for the speech recognition problem [16]. Stressed speech analysis has yielded better modeling approaches for speech production which have been successfully applied to improve speech recognition performance [6], [8], [11], [17], [19], [25], [26]. Stress conditions considered in these studies include perceptually induced stress such as the Lombard effect or task workload (e.g., computer response tasks, aircraft fighter pilot stressed speech) as well as stressed speaking styles such as *fast, slow, clear, angry, loud, soft,* etc. The modeling framework for the present study is based upon a *source generator framework*, which allows for direct modeling of stress perturbation within a multidimensional feature space [7], [13]. In order to reveal the underlying nature of speech production under stress, an extensive evaluation of five speech production feature domains—including glottal spectrum, pitch, duration, intensity, and vocal tract spectral structure—was previously conducted [6]. Extensive statistical assessment of over 200 parameters for simulated and actual speech under stress suggests that stress classification based upon the separability of feature distribution characteristics is possible.

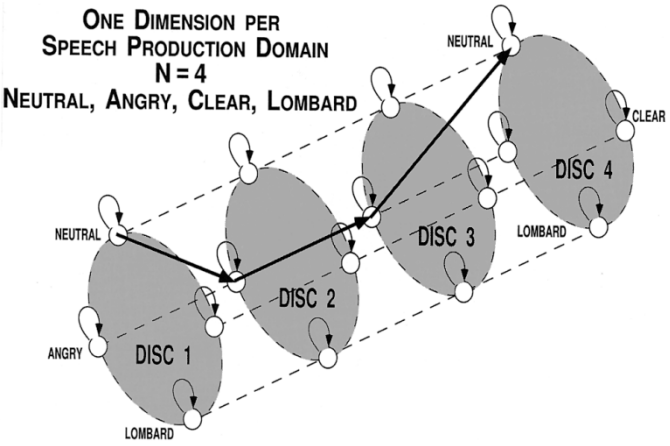


Fig. 1. Stress independent N-channel HMM recognition algorithm.

The idea to formulate a multidimensional HMM has been explored for speech recognition in noise using a two-dimensional (2-D) HMM similar in concept to the N-channel HMM presented in this study [28]. Another study considered a multichannel HMM [29] that had multiple one-channel HMM's in its formulation. This approach differs from the N-channel HMM presented in this study since, here, we allow the individual dimensions to reflect differences in how the *speaker* produces speech, versus dimensions used to reflect *noise/distortions* which corrupt an input speech signal.

II. N-CHANNEL HMM

In this study, we formulate a speech model for robust speech recognition capable of representing sub-phoneme trajectories across stressed speech production domains as illustrated in Fig. 1. The variation in sub-phoneme trajectories is motivated by the observation that a stress style is not uniformly observed over a word or sentence. Consider the word "help" under the Lombard effect condition. Here, the $/H/$ and $/P/$ phonemes[2] would reflect different stress attributes than the $/E/$ or $/L/$ due to voicing and phone class type. As such, Fig. 1 suggests

[2] In this study, we employ the single letter versions of the ARPAbet in [4, p. 118].

that the new N-channel hidden Markov model (N-channel HMM) would be better able to combine the benefits of a stress classification system with a traditional one-channel HMM for speech recognition. The key idea is to address the effects of both intraspeaker and interspeaker variability in one algorithm instead of separately as with tandem stress classification and stress dependent recognition algorithms. This gives the added benefit of a sub-phoneme speech model at the state level instead of the phoneme level that was used in a previous stress directed approach [26].

The fundamental idea of the N-channel HMM is to generalize the one-channel HMM to enable fast computation of multidimensional Markov speech processes. For example, a two-channel HMM could be formulated to model speech from male and female speakers with one dimension allocated for each gender. This would facilitate the integration of separate sub-phoneme statistics in addition to enabling state to state transitions across dimensions for gender. Consider a male speaker who temporarily produces female-like speech within a multisyllable word or phrase. A two-channel HMM would model this by placing the optimal state sequence in the female dimension during that portion of the utterance. Hence, the overall flexibility of the model is improved by allowing a combined model where the integrity of each dimension is preserved. It is suggested that this is better than two separate one-channel HMM models with two Gaussian mixtures because it provides greater separation in the model statistics.

Fig. 1 illustrates an N-channel HMM for the case when $N = 4$, which is designed to model the effects of three additional speaker stress effects. Suppose that this four-dimensional HMM is used to model a phoneme under the four perceptual stress conditions *neutral, angry, clear* and *Lombard* effect. Note that the four discs portrayed in the figure model portions of the phoneme across time from left to right. As each observation (or speech frame) is presented to the N-Channel HMM, a decision is made as to which stress dimension provides the highest score. This example shows a phoneme that is closer to the *neutral* statistics at the beginning and ending of the utterance, and *angry* for the second and third states (or discs).

A. Formulation

The N-channel HMM can be thought of as N single dimensional (one-channel) HMM's [20], [21] that allow state transitions across models.

1) Notation: Consider a Markov process that is modeled with I states (or sites) $S = \{S_1, \cdots, S_I\}$ where each state S_i has a neighborhood of states N_i. Given a sequence of observation vectors $\bar{\mathbf{y}} = \{\vec{y}_1, \cdots, \vec{y}_T\}$, with a K parameter observation $\vec{y}_t = \{y_{t,k} \mid k \in [1, \cdots, K]\}$ at time t, a sequence of states $\vec{z} = \{z_1, \cdots, z_T\}$ will be generated assuming an initial state distribution $I \times 1$ vector $\pi_i = P(z_1 = S_i)$. This model depends upon the state transition probabilities $A = \{a_{ij} = P(z_{t+1} = S_j \mid z_t = S_i, \lambda)\}$ and observation probabilities $B = \{b_j(t) = P(\vec{y}_{t,k} \mid z_t = S_j, \lambda)\}$ $\forall k$, both forming $I \times I$ matrices. The resulting parameters $\lambda = (A, B, \pi)$ therefore define an HMM. The goal here is to maintain a notation for N-channel HMM's that is compatible with the accepted notation for one-channel HMM's in the literature. This is illustrated next by relating N-channel to one-channel HMM's.

2) Relationship to One-Channel HMM's: Let S be a set of N sites with a prior $P(\vec{z} \mid \lambda)$ which is the probability of the state sequence $\vec{z}$ given the HMM parameters λ. This prior depends upon the state transition probabilities as,

$$\begin{aligned} P(\vec{z} \mid \lambda) &= \exp\left[\sum_{t=1}^{T} \ln P(z_t \mid z_{t-1}, \lambda)\right] \\ &= \pi_1 a_{12} a_{23} \cdots a_{T-1,T} \end{aligned} \tag{1}$$

such that $P(z_1 \mid z_0, \lambda) = P(z_1 \mid \lambda)$. The likelihood $P(\bar{\mathbf{y}} \mid \vec{z}, \lambda)$ of a given observation sequence $\bar{\mathbf{y}}$ therefore depends upon the observation probabilities as follows:

$$\begin{aligned} P(\bar{\mathbf{y}} \mid \vec{z}, \lambda) &= \prod_{t=1}^{T} P(\vec{y}_t \mid z_t, z_{t-1}, \lambda) \\ &= b_1(\vec{y}_1) b_2(\vec{y}_2) \cdots b_T(\vec{y}_T). \end{aligned} \tag{2}$$

Finally, the probability of an observation sequence $\bar{\mathbf{y}}$ given the model parameters λ is

$$\begin{aligned} P(\bar{\mathbf{y}} \mid \lambda) &= \sum_{\forall z_t, z_{t-1}} P(\bar{\mathbf{y}} \mid z_t, z_{t-1}, \lambda) P(z_t, z_{t-1} \mid \lambda) \\ &= \sum_{\forall z_t, z_{t-1}} \pi_1 b_1(\vec{y}_1) a_{12} b_2(\vec{y}_2) \ldots a_{T-1,T} b_T(\vec{y}_T). \end{aligned} \tag{3}$$

For convenience, we suggest an inner product notation on the characteristics of the hidden state variable z, where k is the observation parameter, m is the mixture [see (17)], and i and j are state indices

$$\begin{aligned} \langle z_{t,j,m} \rangle &\approx \gamma_t(j, m) \\ &= P(z_t = S_j \mid k, \bar{\mathbf{y}}, \lambda) \end{aligned} \tag{4}$$

$$\begin{aligned} \langle z_{t,i} z_{t+1,j} \rangle &\approx \xi_t(i, j) \\ &= P(z_t = S_i, z_{t+1} = j \mid \bar{\mathbf{y}}, \lambda) \end{aligned} \tag{5}$$

$$\begin{aligned} \langle z_{t,j} \mid \bar{\mathbf{y}}, \lambda \rangle &\approx \gamma_t(j) \\ &= \sum_{m=1}^{M} \gamma_t(j, m). \end{aligned} \tag{6}$$

3) Reestimation Equations: In the reestimation, one must find the best set of model parameters $\lambda = (A, B, \pi)$ such that the observation probability $P(\bar{\mathbf{y}} \mid \lambda)$ is maximized. The well known forward-backward variables α and β are used to reestimate the joint probability $\langle z_{t,i} z_{t+1,j} \rangle$ of being in state i at time t, and state j at time $t+1$ as

$$\begin{aligned} \langle z_{t,i} z_{t+1,j} \rangle &\approx P(z_t = S_i, z_{t+1} = S_j \mid \bar{\mathbf{y}}, \lambda) \\ &= \frac{\alpha_t(i) a_{ij} b_j(\vec{y}_{t+1}) \beta_{t+1}(j)}{P(\bar{\mathbf{y}} \mid \lambda)} \\ &= \frac{\alpha_t(i) a_{ij} b_j(\vec{y}_{t+1}) \beta_{t+1}(j)}{\sum_{i=1}^{I} \sum_{j=1}^{I} \alpha_t(i) a_{ij} b_j(\vec{y}_{t+1}) \beta_{t+1}(j)} \end{aligned} \tag{7}$$

which implies that

$$\langle z_{t,i} \mid \bar{\mathbf{y}}, \lambda \rangle = \sum_{j=1}^{I} \langle z_{t,i} z_{t+1,j} \rangle. \tag{8}$$

With this, the following are the reestimation equations for the HMM model $\hat{\lambda}$:

$$\hat{\pi}_i = \langle z_{1,i} \mid \bar{\mathbf{y}}, \lambda \rangle \tag{9}$$

$$\hat{a}_{ij} = \frac{\sum_{t=1}^{T-1} \langle z_{t,i} z_{t+1,j} \rangle}{\sum_{t=1}^{T-1} \langle z_{t,i} \mid \bar{\mathbf{y}}, \lambda \rangle} \tag{10}$$

$$\hat{b}_j(\vec{y}_t) = \sum_{m=1}^{M} c_{jm} N(\vec{y}_t, \mu_{jm}, \sigma_{jm}), \quad j \in [1, \cdots, I] \tag{11}$$

where for the continuous observation densities $b_j(\vec{y}_t)$, there are M Gaussian mixtures with gains

$$\hat{c}_{jm} = \frac{\sum_{t=1}^{T} \langle z_{t,j,m} \rangle}{\sum_{t=1}^{T} \sum_{m=1}^{M} \langle z_{t,j,m} \rangle} \tag{12}$$

$$= \frac{\sum_{t=1}^{T} \langle z_{t,j,m} \rangle}{\sum_{t=1}^{T} \langle z_{t,j} \mid \bar{\mathbf{y}}, \lambda \rangle} \tag{13}$$

that form an $I \times M$ matrix. The mean $\vec{\mu}_j$ and variance $\vec{\sigma}_j$ vectors in each row of the $I \times M$ matrices $\hat{\mu}$ and $\hat{\sigma}$ are given by

$$\hat{\mu}_{jm} = \frac{\sum_{t=1}^{T} \langle z_{t,j,m} \rangle \cdot \vec{y}_t}{\sum_{t=1}^{T} \langle z_{t,j,m} \rangle} \tag{14}$$

$$\hat{\sigma}_{jm} = \frac{\sum_{t=1}^{T} \langle z_{t,j,m} \rangle \cdot (\vec{y}_t - \mu_{jm})(\vec{y}_t - \mu_{jm})^T}{\sum_{t=1}^{T} \langle z_{t,j,m} \rangle} \tag{15}$$

where

$$\langle z_{t,j,m} \rangle \approx \left[\frac{\alpha_t(j)\beta_t(j)}{\sum_{i=1}^{I} \alpha_t(i)\beta_t(i)} \right] \times \left[\frac{\hat{c}_{jm} N(\vec{y}_t, \mu_{jm}, \sigma_{jm})}{\sum_{m=1}^{M} \hat{c}_{jm} N(\vec{y}_t, \mu_{jm}, \sigma_{jm})} \right]. \tag{16}$$

The multidimensional m mixture Gaussian distribution

$$N(\vec{y}_t, \mu_{jm}, \sigma_{jm}) = \frac{\exp\left[-\frac{1}{2}(\vec{y}_t - \mu_{jm})^T \sigma^{-1} (\vec{y}_t - \mu_{jm})\right]}{(2\pi)^{K/4} \sqrt{|\sigma|}} \tag{17}$$

is normalized with the term K, which is the length of the observation vector $\vec{y}_t$. Having presented the relations for iterative estimation of the N-channel HMM parameters, higher level training issues are discussed next.

4) Training Phases: There are three distinct phases required to train an N-channel HMM, as follows:

1) one-channel stress dependent training;
2) N-channel state transition training;
3) N-channel model refinement.

The first phase requires training a codebook of stress dependent one-channel HMM's in the usual manner. In this study, k-means clustering, frame energy weighting, and selective token training [31] are employed. The k-means clustering modification to HMM initialization is simply a better means of obtaining initial state estimates and is well established in the literature. Frame energy weighting is a training enhancement that reduces the impact of low-energy speech frames on parameter reestimation. Finally, selective token training removes tokens from the training process that are clearly outliers. This idea is based on a selective training procedure originally described in [1] and [31]. A training token is considered an outlier after the second iteration of training if it produces a log score that is more than two times smaller than the last average mean score. This prevents outliers from overly modifying the Gaussian mixture models.

The second phase of training requires combining the stress dependent one-channel models into an N-channel model which is trained without k-means clustering, mean updating, or variance updating. Speech data from all N of the stress classes are used to train only the state transitions of the N-channel HMM at one time. Regular left-to-right transitions within one dimension are allowed in a manner similar to that in the original separate one-channel models. In Fig. 1, these are the transitions from disc to disc along one dimension (i.e., *neutral*). For a given state, transitions are also allowed within the same disc or, equivalently, across dimensions. Furthermore, transitions are allowed to any state in the next disc, which is not only a transition across time, but also from one stress class to another. If the N-channel HMM is to be employed for stress classification, then training is completed at this point. However, if the N-channel HMM is to be also used for speech recognition, then the third phase of training is employed. It requires training the N-channel HMM using the output model from the second phase of training; however, mean, variance, and state transition updating is enabled. Essentially, this phase takes a very good initial model and refines it slightly. It will be shown in the evaluations that this phase does not significantly affect the correlation of each of the dimensions with each corresponding stress condition in the N-channel model. It is suggested that this is due to the local convergence property of the EM (expectation-maximization) algorithm used here. The resulting initial model from the second phase can be said to be already in the "valley" of the cost function; hence, the third phase simply moves the model parameters to the minimum of that cost "valley."

B. Relationship to One-Channel HMM

The basic strength of the N-channel HMM compared to the one-channel HMM is that it provides a more flexible model. Alternatively, one could suggest that N-channel HMM's bear a similarity to an N mixture one-channel HMM, and therefore ask why are N-channel HMM's necessary? The reason is that an N mixture one-channel model does indeed have N separate means and variances for each state. However, it does not have the separate state transition probabilities available in the

N-channel HMM. Furthermore, the allowable state transitions and training method for the N-channel HMM is such that each speech production domain is clustered into each of the dimensions of the N-channel model; hence, making the model more flexible. It should also be noted that there is no reason why each state of the N-channel HMM could not also have multiple mixtures. For example, if it is desired to train an N-channel HMM for ten stress conditions for both male and female speakers, one could employ a ten-channel HMM with two mixtures per state.

Another strength of the N-channel HMM formulation becomes apparent when the problem of stress independent recognition is considered. In two previous studies [25], [26], a phoneme class-based stress classifier using neural networks was employed to direct a codebook of stress dependent one-channel HMM recognizers. Though these were isolated word recognizers, they could easily have been phoneme-based recognizers. In the case where the stress dependent speech recognizers are phone-based, the phone-based stress classifier would direct each phone to the appropriate recognizer. It is suggested that this would further improve the stress directed recognition performance. Such an experiment was not performed because we opted to take the next step to an N-channel HMM instead. An N-channel HMM works in a similar manner with the exception that the basic speech unit is of sub-phoneme or state duration. From our understanding of how speaker stress affects speech production, we suggest that this is a more appropriate way to address the effects of stress. This is due to the differing impact of stress on each phone class. For example, the perception of speaker stress for unvoiced consonant stops $(/p/, /k/, /t/)$ may be little due to limited temporal information, whereas vowels $(/@/, /E/, /I/, /R/, /U/$ [4]) are significantly affected. Furthermore, within a given phoneme class, the feature trajectories will differ for each stress condition across time. Hence, the N-channel HMM makes a stress class decision on a state (sub-phoneme) basis resulting in a better segmentation of the speech utterance. The results in the evaluations will show that this does indeed improve recognition performance.

C. Application to Stress Classification

As mentioned previously, the N-channel HMM integrates the stress classification decision into each state transition. Hence, a readily apparent means of generating a stress probability vector is to calculate the ratio of times the state trajectory passes through each stress dimension versus the total number of states. The neural network approach used in a previous study on stress classification [26] requires a dedicated structure for the speech processing units, whereas the N-channel HMM offers the opportunity to do stress classification in a variety of recognition scenarios. The structure of the neural network is more dependent upon the accuracy of a required front-end phoneme parser than an HMM would be for this task. The previous neural network approach also requires features that are both fixed in relative positions throughout a phoneme as well as features that summarize the statistical distribution of the phoneme.

III. Evaluations

The two applications of the N-channel HMM evaluated in this study are 1) stress classification and 2) stress resistant speech recognition. The performance of the N-channel HMM employed as a stress classifier will be compared to the neural network based stress classification system considered in a previous study [26]. Next, in order to assess the application of the N-channel HMM to stress independent recognition, a comparison to previously developed [24]–[26] stress dependent one-channel HMM's is investigated. Additionally, the performance of the one-channel stress dependent HMM will be compared to a *neutral* trained model for speech under stress to illustrate the need to model speaker stress effects.

A. Stressed Speech Data

It is important to establish the domain of the speech data employed in this study to understand the difficulties of conducting research on speech under stress. The evaluations conducted in this study employ data previously collected for analysis and algorithm formulation of speech under stress and noise. This database, called SUSAS [7], [6], [10] refers to *speech under simulated and actual stress*, and has been employed extensively in the study of how speech production varies when speaking during stressed conditions. SUSAS consists of five domains, encompassing a wide variety of stresses and emotions. A total of 44 speakers (14 female, 30 male), with ages ranging from 22 to 76, were employed to generate in excess of 16,000 utterances. The five stress domains include:

1) psychiatric analysis data (speech under depression, fear, anxiety);
2) talking styles[3] (*angry, clear, fast, loud, slow, soft*);
3) single tracking task (mild task *Cond50*, high task *Cond70* computer response workload) or speech produced in noise (*Lombard* effect);
4) dual tracking computer response task;
5) subject motion-fear tasks (G-force, *Lombard* effect, noise, fear).

Lombard effect was simulated by having speakers listen to 85 dB SPL of pink noise through headphones while producing speeh (i.e., speech tokens were noise free). The database offers a unique advantage for analysis and design of speech processing algorithms in that both *simulated* and *actual* stressed speech are available. A common vocabulary set of 35 aircraft communication words make up over 95% of the database. These words consist of mono- and multisyllabic words which are highly confusable. Examples include "go"—"oh"—"no;" "wide"—"white;" and "six"—"fix." A more complete discussion of SUSAS can be found in the literature [6]–[8], [10].

Four stress conditions make up the domain of the evaluations considered here; hence, a four-channel HMM $(N = 4)$ with a total of sixty states $(I = 60)$ is partitioned into 15 states per dimension. The stress conditions include *neutral*, *angry*, *clear*, and *Lombard* which are allocated to each dimension (e.g., states 0–14 for *neutral*, 15–29 for *angry*, 30–44 for *clear*, and 45–59 for *Lombard* as shown in Fig. 2). The speech

[3] Approximately half of SUSAS consists of simulated style data donated by Lincoln Laboratory [17].

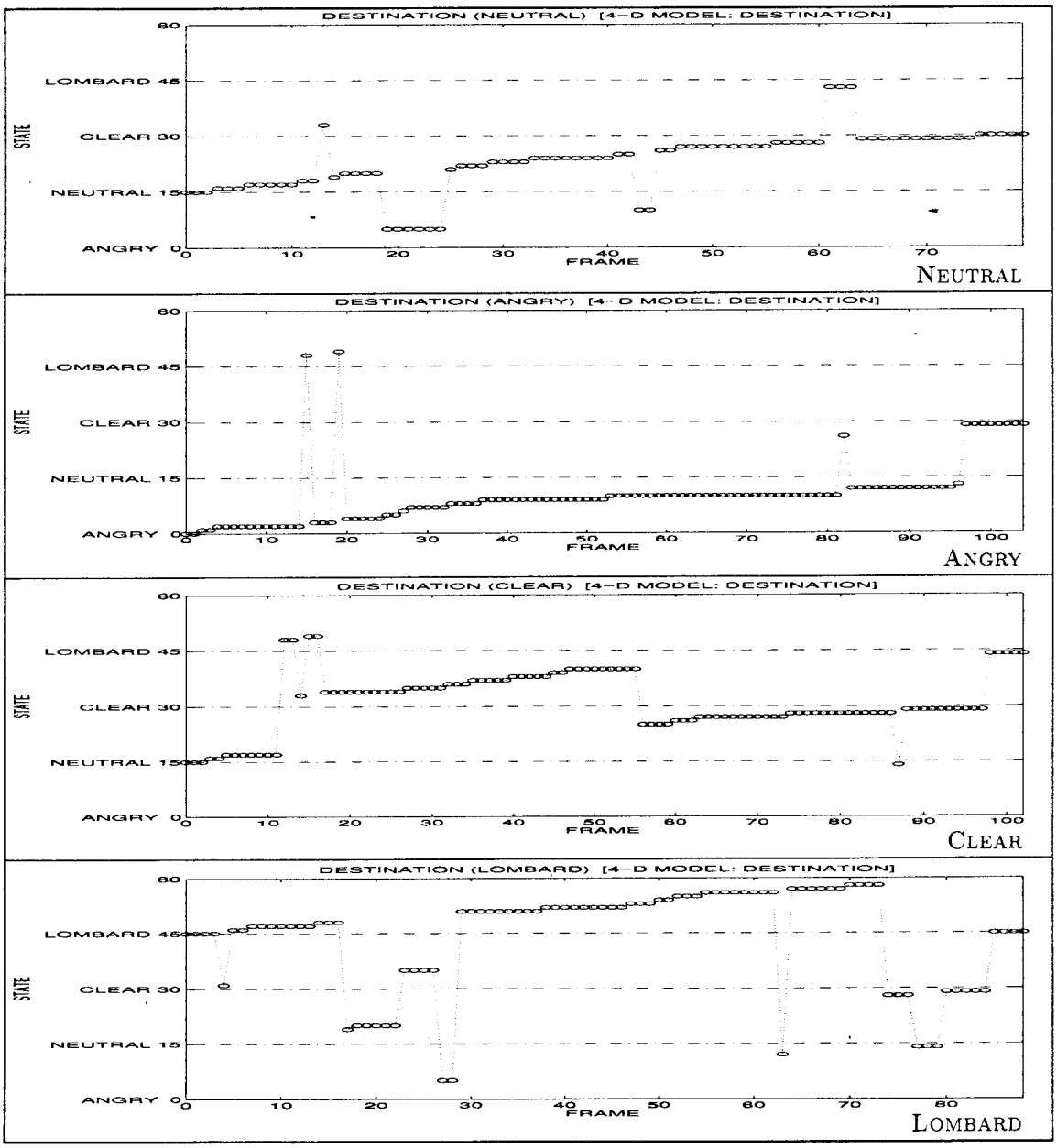


Fig. 2. N-channel HMM optimal state sequence of the word "destination" for *neutral, angry, clear,* and *Lombard.*

waveform is segmented into frames sampled at 8 kHz with a 30 ms window length and 10 ms skip rate. From the SUSAS database, six speakers are used for training and three for open testing of both the one-channel and N-channel HMM models. There are two tokens available for each stress condition in the 35 isolated word vocabulary employed. In the second and third phases of N-channel HMM training, the tokens for all four stress conditions are used.

B. Stress Classification

For stress classification, the N-channel HMM model is trained using the first two phases of training: 1) stress dependent one-channel HMM model generation, and 2) N-channel HMM state transition training. Fig. 2 illustrates the Viterbi decoded best state path through the N-channel HMM for the word "destination" under the four stressed speaking conditions. Each of the parts of Fig. 2 (i.e., *neutral, angry, clear, Lombard*) is generated using one of the three test speakers to find the best path through the single N-channel word model. It is apparent from this figure that stress classification occurs for the stress conditions of neutral, angry, and Lombard effect, since the clear majority of the observations belong to state transitions within those particular stress classes. We see that for the word "destination" under an angry stress condition, only three observation frames in the first 97 observations are associated with stress conditions other than angry; the remaining nine observation frames from the final portion of the nasal /N/ where associated with states from the neutral portion of the N-channel HMM. For the word "destination" spoken under *clear*, the observations occur from states associated with the *clear* dimension during the initial long voiced section (i.e., frames

TABLE I
ONE-CHANNEL AND N-CHANNEL HMM VERSUS NEURAL NETWORK STRESS CLASSIFICATION USING NINE SPEAKERS, 35 ISOLATED WORDS, AND AN OPEN THREE-SPEAKER TEST SET. THE SPEECH FEATURE VECTOR IS COMPOSED OF FIVE C-MEL, THREE DC-MEL, THREE D2C-MEL, AND THREE AC-MEL PARAMETERS

CLASSIFICATION ALGORITHM	STRESS CLASSIFICATION RATE (%)				
	Neutral	*Angry*	*Clear*	*Lombard*	AVERAGE
Neural Network	31.90	11.43	81.90	11.90	**34.28**
1-Channel HMM	53.81	84.76	51.43	44.29	**58.57**
N-Channel HMM (3rd Phase)	45.71	72.86	50.95	43.81	**53.33**
N-Channel HMM (2nd Phase)	46.67	78.57	54.29	50.95	**57.62**

18–55), but move to the *neutral* dimension for the majority of the ending portion of the utterance (i.e., frames 56–97, except for frame 88); resulting in an incorrect *neutral* classification. This is a reasonable transition across dimensions, since *clear* and *neutral* speech have similar characteristics for some phone classes. For *neutral, angry*, and *Lombard* utterances, there are several momentary transitions but the correct stress class is consistently identified in the optimal state path sequence. It is important to note that speakers may not always exhibit the same type/level of stress throughout an utterance or phrase; and, therefore, momentary transitions into other stress states may be possible. Therefore, this example supports the assertion that each class of phonemes and subphonemes is affected to varying degrees by stress relative to location and local phoneme content.

In a previous study on stressed speech detection, the Teager nonlinear energy operator was used for stress detection on extracted vowels on a *pairwise* basis to yield performance of 97.5%, 99.1%, 64.8%, and 86.1% for *neutral, angry, clear,* and *Lombard* effect stress conditions, respectively [3]. It should be noted these are stress *detection* results (pairwise decision), are not stress *classification* results (one of four decision); hence, the results are not directly comparable to those with higher stress dimensions [12], [24], [26].

For purposes of comparison, a neural network stress classifier is employed here, which is a nontargeted feature triphone based algorithm. This is different than a targeted feature system (employed in [26]) and has been chosen for the purpose of comparison since the N-channel HMM stress classifier uses nontargeted features (e.g., the same features are used for all classification decisions). The feature vector employed is based upon the duration, five C-Mel, three DC-Mel, three D2C-Mel, and three AC-Mel coefficients (C-Mel stands for Mel-cepstral, DC-Mel is delta C-Mel, D2C-Mel is delta-delta C-Mel, and AC-Mel is the autocorrelation C-Mel. A more complete discussion of these parameters is presented in [12]). The mean, variance, and slope of these parameters are calculated across each phoneme. All features are based on the center phoneme in every triphone; however, the two adjacent phonemes are used in obtaining the mean and variance of the features. This nontargeted feature neural network stress classifier achieves a performance of 31.9%, 11.4%, 81.9%, and 11.9% for *neutral, angry, clear,* and *Lombard* effect respectively for an average rate of 34.28% (see Table I). The classification performance for the *angry* and *Lombard* stress conditions are very low, which we determined was due to confusion between the *clear* and *Lombard* effect stress classes. It is also apparent that the *clear* condition is selected significantly more often than *angry* or *Lombard* conditions in these simulations. Since we might assume that all stress conditions have equal a priori probabilities, we have previously shown that for this reason, targeted feature stress classification is better able to differentiate confusable stress classes [26].

In Table I, stress classification rates for the N-channel HMM (with the second phase of N-channel training) are 46.7%, 78.6%, 54.3%, and 51.0% for *neutral, angry, clear,* and *Lombard* effect, respectively, for an average rate of 57.62%. Hence, the N-channel HMM stress classifier has a 23.34% higher performance than the neural network stress classifier. Note that the stress classification performance degrades slightly for the third phase of N-channel HMM training (model refinement) by -4.29% on average. This is an important observation because it shows that the integrity of the dimensions in the N-channel HMM have not been overly corrupted by the third phase of training. Hence, allowing mean and variance updating in the third phase of training simply refines the model without significantly reducing the correlation of each dimension to the stress condition with which it is associated from the first two phases of training.

Another useful comparison of the performance of the N-channel HMM as a stress classifier is to assess the ability of a codebook of stress dependent one-channel HMM's to classify speech under stress. The codebook in this case consists of four stress dependent one-channel HMM's. The stress class decision is based upon the model with the highest log score. The one-channel HMM yielded classification rates of 53.8%, 84.8%, 51.4%, and 44.3% for *neutral, angry, clear,* and *Lombard* effect respectively for an average rate of 58.57%. This is slightly higher on average than the N-channel HMM stress classifier. However, with the size of the training and test set, it is not known with statistical certainty which HMM will outperform the other. We suspect that the N-channel HMM will outperform the one-channel HMM because of the greater score separation seen in the recognition experiments in the next section. It is suggested that the N-channel HMM could improve stress classification performance with larger speaker sets, larger vocabularies, or in a phoneme based system. Next, the N-channel HMM is applied to recognition of speech under stress.

C. Stress Independent Recognition

In order to employ the N-channel HMM for stress independent speech recognition, the third phase of N-channel model refinement is performed after the same two phases required for stress classification. The two speech recognition evaluations in this study compare: 1) neutral versus stress dependent trained

TABLE II
HMM RECOGNITION RATES FOR ONE-CHANNEL AND N-CHANNEL MODELS USING NINE SPEAKERS, 35 ISOLATED WORDS, AND AN OPEN THREE SPEAKER TEST SET. TRAINING IS PERFORMED WITH 15 STATES, ONE MIXTURE, 12 C-MEL, FIVE DC-MEL, FIVE D2C-MEL, FIVE AC-MEL, AND LOG ENERGY PARAMETERS

HMM TRAINING TYPE	RECOGNITION RATE (%)				
	Neutral	*Angry*	*Clear*	*Lombard*	AVERAGE
1-Channel Neutral Trained	80.00	40.48	81.43	69.05	**67.74**
1-Channel Stress Dependent	80.00	71.90	82.38	80.48	**78.69**
N-Channel Stress Independent	96.19	89.52	96.67	95.24	**94.41**

TABLE III
HMM SCORE SEPARATION FOR ONE-CHANNEL AND N-CHANNEL MODELS USING NINE SPEAKERS, 35 ISOLATED WORDS, AND AN OPEN THREE SPEAKER TEST SET. TRAINING IS PERFORMED WITH 15 STATES, ONE MIXTURE, 12 C-MEL, FIVE DC-MEL, FIVE D2C-MEL, FIVE AC-MEL, AND LOG ENERGY PARAMETERS

HMM TRAINING TYPE	RECOGNITION LOGSCORE SEPARATION				
	Neutral	*Angry*	*Clear*	*Lombard*	AVERAGE
1-Channel Neutral Trained	4.21	-0.48	4.08	2.94	**2.69**
1-Channel Stress Dependent	4.21	2.49	5.08	4.64	**4.10**
N-Channel Stress Independent	6.87	5.97	8.13	7.32	**7.07**

1-Channel HMM's and 2) the stress independent N-channel HMM versus stress dependent one-channel HMM's.

First, the performance of a codebook of one-channel single continuous Gaussian mixture density, fifteen state HMM's are evaluated. In a previous study [26], the one-channel stress dependent HMM showed an improvement of $+10.1\%$ over conventionally trained neutral, and $+15.4\%$ improvement over multistyle trained recognizers, respectively. In this study, the HMM training algorithm described in Section II-A4 is employed. The recognition rate for both neutral and stress dependent trained one-channel HMM's increased due to these training options; however, the improvement was still $+10.95\%$ as shown in Table II. These training techniques also increase the separation between the correct target word log score and the subsequent second highest log score as detailed in Table III. The separability measure is normalized by the number of training tokens and provides a measure of the size of the difference between the correct word score and the next highest score. Therefore, even when recognition rates are similar, it is possible to assess how "close" a given model is to making an error.

It is typical practice in the speech processing field to neglect the effects of speaker stress by training neutral speech models for speech recognition in adverse environments. The cost of this decision is made clear by studying the performance of the *neutral* one-channel HMM models applied to speech under stress. The degradation in recognition performance is -31.42%, a reduction from 71.90% for the stress dependent model to 40.48% when *angry* speech was presented to a *neutral* speech model. However, the recognition rate for *clear* speech only dropped -0.95% from 82.38% to 81.43% which is not statistically significant. Finally, for *Lombard* speech, performance dropped by -11.43% from 80.48% to 69.05%. Hence, an average loss of -10.95% in recognition performance occurs when using a *neutral* trained versus a stress dependent one-channel HMM recognizer.

The results show that (see Table II) the N-channel HMM considerably outperforms the one-channel stress dependent HMM by an average of 15.72%. Another interesting feature of the N-channel HMM is that it normally provides greater separation between output HMM log scores based upon the separability measure shown in Table III. This implies that, with a larger positive separability score, the separation between the correct token score and the second highest token score summarizes how well a model accepts the correct and rejects the incorrect tokens. For the N-channel HMM, the separability measure is 7.07 which is consistently greater than the 4.10 separation measure for the one-channel stress dependent HMM as shown in Table III. This property of greater separability in output HMM log scores for the N-channel HMM could lead to more robust models for speech under stress.

IV. SUMMARY OF FINDINGS AND CONCLUSIONS

The problem of stressed speech classification and stress independent recognition has been considered using a modified Markov process model. This new multidimensional hidden Markov model (N-channel HMM) has been formulated to generalize a set of N single dimensional (one-channel) HMM's to allow transitions across individual models. By employing a more general Markov model, it has been shown that reliable stress classification and improved speech recognition performance of speech under stress can be achieved simultaneously. While stress classification rates increased by 23.34% over an approach similar in structure to a previously formulated nontargeted feature neural network classifier [26], the stress classification performance was comparable to the one-channel HMM stress classifier (i.e., training separate HMM recognizers on speech from known stress conditions, and selecting the highest probable model). However, N-channel HMM stressed speech recognition rates did increase by 15.72% over a previously tested stress dependent one-channel HMM approach [26] (i.e., an overall recognition rate of 94.41% versus 78.69%). We must emphasize here that both stress classification and speech recognition evaluations were conducted on a small vocabulary set (35 words), and a small speaker set (speaker sets of less than ten speakers). However, the difficulty in collecting, organizing and calibrating a large speech under stress database has prevented evaluations by any group on

larger data sets. At the present time, SUSAS [10] represents the largest speech under stress database available to the speech research community.[4] Understanding the limitations of the present database, versus much larger corpora available for large vocabulary continuous speech recognition, we can make some concluding remarks. The formulation and evaluations presented here have suggested that the N-channel HMM is able to achieve higher levels of speech processing performance by integrating intraspeaker stress and inter-speaker nonstress characteristics into a single model. Previously, this task would normally be modeled separately with a stress classifier and a codebook of stress dependent recognition systems.

In the future, it would be useful to determine if such a formulation could be scaled up to address large vocabulary speech recognition under stress.[5] Such an evaluation for the N-channel HMM would require collecting a tremendously large speech corpus which would be labeled based on stress content across speakers. In addition, other speech problems could be considered such as applying the N-channel HMM to problems such as gender identification, phone recognition, parsing, and speaker identification. Performance could be compared to multiple Gaussian mixture models or separate model templates. It would also be interesting to assess the performance of both monophone and triphone based N-channel HMM speech recognizers and stress classifiers for speech produced under adverse conditions.

In a manner similar to the problem of stress classification, speaker verification could be approached using a N-Channel HMM formulation. For example, in a speaker verification application, each dimension of the N-channel HMM could be allocated for each of N speakers. The identity of the speaker could then be confirmed by monitoring state transitions within the target speaker's dimension.

REFERENCES

[1] L. M. Arslan and J. H. L. Hansen, "Improved HMM training and scoring strategies with application to accent classification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1996, vol. 2, pp. 589–592.

[2] S. E. Bou-Ghazale and J. H. L. Hansen, "HMM-based stressed speech modeling with application to improved synthesis and recognition of isolated speech under stress," *IEEE Trans. Speech Audio Processing*, vol. 6, pp. 201–216, May 1998.

[3] D. A. Cairns and J. H. L. Hansen, "Nonlinear analysis and detection of speech under stressed conditions," *J. Acoust. Soc. Amer.*, vol. 96, pp. 3392–3400, 1994.

[4] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*. Englewood Cliffs, NJ: MacMillan Series for Prentice-Hall, 1993.

[5] V. V. Digalakis, D. Rtischev, and L. G. Neumeyer, "Speaker adaptation using constrained estimation of Gaussian mixtures," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 357–366, Sept. 1995.

[6] 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, GA, p. 428, July 1988.

[4] Researchers interested in this area are encouraged to see the NATO IST/TG-01 Speech Under Stress web page: http://www.cslu.colorado.edu/rspl/stress.html. SUSAS is now available from LDC. See their web page: http://www.ldc.upenn.edu/.

[5] An evaluation by DERA, U.K. using their large vocabulary continuous speech recognizer (trained on *Wall Street Journal* SI284, about 80 hours of training data) showed an increase in word error rate from 16.6% for neutral, to 21.5% for *clear*, 27.9% for *Lombard*, and 52.4% for *angry*. This clearly confirms the need to address speaker stress variability for large vocabulary speech recognition.

[7] J. H. L. Hansen, "Morphological constrained feature 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.

[8] J. H. L. Hansen, "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition," *Speech Commun.*: *Spec. Issue Speech Under Stress*, vol. 20, pp. 151–173, Nov. 1996. Also in *Proc. ESCA-NATO Workshop on Speech Under Stress*, Sept. 1995, Lisbon, Portugal, pp. 91–98.

[9] 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.

[10] J. H. L. Hansen and S. E. Bou-Ghazale, "Getting started with SUSAS: A speech under simulated and actual stress database," *Eurospeech'97*, Rhodes, Greece, Sept. 1997, vol. 4, pp. 1743–1746. Available at http://morph.ldc.upenn.edu/catalog/LDC99s78.html.

[11] J. H. L. Hansen and M. A. Clements, "Source generator equalization and enhancement of spectral properties for robust speech recognition in noise and stress," *IEEE Trans. Speech and Audio Processing*, vol. 3, pp. 407–415, Sept. 1995.

[12] J. H. L. Hansen and B. D. Womack, "Feature analysis and neural network based classification of speech under stress," *IEEE Trans. Speech Audio Processing*, vol. 4, pp. 307–313, July 1996.

[13] J. H. L. Hansen, B. D. Womack, and L. M. Arslan, "A source generator based production model for environmental robustness in speech recognition," *Int. Conf. Spoken Language Processing*, 1994, pp. 1003–1006.

[14] J. C. Junqua, "The Lombard reflex and its role on human listeners and automatic speech recognizers," *J. Acoust. Soc. Amer.*, vol. 93, pp. 510–524, Jan. 1993.

[15] C.-H. Lee, C.-H. Lin, and B.-H. Juang, "A study on speaker adaptation of the parameters of continuous density hidden Markov models," *IEEE Trans. Signal Processing*, vol. 39, pp. 806–814, Apr. 1991.

[16] H. Lee and A. Tsoi, "Application of multi-layer perceptron in estimating speech/noise characteristics for speech recognition in noisy environment," *Speech Commun.*, vol. 17, pp. 59–76, Aug. 1995.

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

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

[19] D. B. Paul, "A speaker-stress resistant HMM isolated word recognizer," in *IEEE Proc. Int. Conf. Acoust., Speech, Signal Processing*, 1987, pp. 713–716.

[20] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, pp. 267–295, Feb. 1989.

[21] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Englewood Cliffs, NJ: Prentice-Hall, 1993.

[22] P. V. Simonov and M. V. Frolov, "Analysis of the human voice as a method of controlling emotional state: Achievements and goals," *Aviation, Space, Env. Sci.*, vol. 1, pp. 23–25, 1977.

[23] C. E. Williams and K. N. Stevens, "Emotions and speech: Some acoustic correlates," *J. Acoust. Soc. Amer.*, vol. 52, pp. 1238–1250, 1972.

[24] B. D. Womack and J. H. L. Hansen, "Stress independent robust HMM speech recognition using neural network stress classification," in *Proc. Eurospeech*, Sept. 1995, pp. 1999–2002.

[25] B. D. Womack and J. H. L. Hansen, "Stressed speech classification with application to robust speech recognition," *IEEE Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 1, pp. 53–56, 1996.

[26] B. D. Womack and J. H. L. Hansen, "Classification of speech under stress using target driven features," *Speech Commun.*: *Speech Under Stress*, vol. 20, pp. 131–150, Nov. 1996.

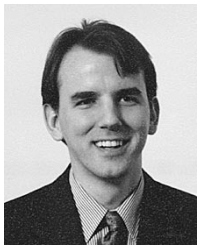
[27] B. D. Womack, "Classification and recognition of speech under perceptual stress using neural networks and N-D HMM's," Ph.D. dissertation, Duke Univ., Durham, NC, Dec. 1996.

[28] A. P. Varga and R. K. Moore, "Hidden Markov model decomposition of speech and noise," in *IEEE Proc. Int. Conf. Acoustics, Speech, Signal Processing*, 1990, vol. 2, pp. 845–848.

[29] D. Xu, C. Fancourt, and C. Wang, "Multi-channel HMM," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 1996, pp. 841–844.

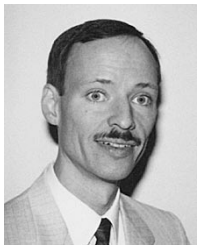
[30] J. H. L. Hansen *et al.*, *The Impact of Speech Under Stress on Military Speech Technology*, NATO Tech. Rep. AC/232/IST/TG-01, March 1999. Available at http://cslu.colorado.edu/rspl/stress.html.

[31] L. M. Arslan and J. H. L. Hansen, "Selective training in hidden Markov model recognition." *IEEE Trans. Speech Audio Processing*, vol. 7, pp. 46–54, Jan. 1999.



Brian David Womack was born in Endicott, NY, on October 3, 1966. He received the B.S. in electrical engineering degree from Florida Atlantic University, Boca Raton, in 1988, the M.S. degree from Texas A&M University, College Station, in 1990 (with focus on robust adaptive control), and the Ph.D. degree from Duke University, Durham, NC, in 1997 (with focus on robust speech processing under stress).

In 1998, he founded Speech Technologies Corporation, West Valley City, UT. His primary research interests include speech processing in noisy and stressful environments, speaker verification, speech enhancement, natural language parsing, and embedded real-time application development. Currently, he is developing speech applications in both the medical and military settings.



John H. L. Hansen (S'81–M'82–SM'93) was born in Plainfield, NJ. He received the B.S.E.E. degree (with 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, GA, in 1983 and 1988, respectively.

In 1988, he joined the faculty of the Department of Electrical Engineering, Duke University, Durham, NC, as an Assistant Professor, and later became Associate Professor. There he established and directed the Robust Speech Processing Laboratory (RSPL). He also held a secondary appointment in the Department of Biomedical Engineering. Prior to joining the Duke University faculty, he was employed by the RCA Solid State Division, Somerville, NJ, (1981–1982), and Dranetz Engineering Laboratories, Edison, NJ, (1978–1981). In January 1999, he moved RSPL to the University of Colorado, Boulder, where (with R. Cole and W. Ward) he established the new Center for Spoken Language Understanding (CSLU). He is presently an Associate Professor in the Department of Speech, Language, and Hearing Sciences and the Department of Electrical and Computer Engineering. He also serves as Associate Director of CSLU. He has served as a technical consultant to industry and the U.S. government, including AT&T Bell Laboratories, IBM, Sparta, ASEC, VeriVoice, and DoD, in the areas of voice communications, wireless telephony, robust speech recognition, and forensic speech/speaker analysis. His research interests span the areas of digital signal processing, analysis and modeling of speech under stress and/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, accent, and Lombard effect. He is the author of more than 100 journal and conference papers in the field of speech processing and communications, and is co-author 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), and was Tutorials Chair for IEEE ICASSP'96. He was an Associate Editor for the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING (1992–1998), and is presently serving as Associate Editor for the IEEE SIGNAL PROCESSING LETTERS. He also served as guest editor of the TRANSACTIONS ON SPEECH AND AUDIO PROCESSING October 1994 special issue on robust speech recognition. 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.