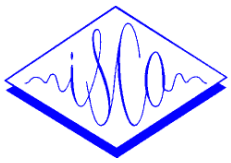

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

# Stress Independent Robust HMM Speech Recognition using Neural Network Stress Classification

**Brian D. Womack and John H.L. Hansen**

Robust Speech Processing Laboratory
Duke University Department of Electrical Engineering
Box 90291, Durham, North Carolina 27708-0291
E-Mail: womack@ee.duke.edu jhlh@ee.duke.edu

## Abstract

It is well known that the variability in speech production due to task induced stress contributes significantly to loss in speech recognition performance [6, 8]. If an algorithm could be formulated which estimates the speech stress condition, then such knowledge could be integrated to improve robustness of speech recognizers in adverse conditions. In this paper, the problem of automatic stressed speech recognition is addressed. The primary goal is to formulate a tandem HMM and neural network based algorithm for stress independent recognition. To motivate an effective stress classifier, an analysis is performed of speech produced across eleven stress conditions (e.g. *Angry*, *Clear*, *Fast*, *Lombard*, *Loud*, *Slow*, *Soft*, etc.). Features that differentiate stress using a previously established stressed speech database (*SUSAS*) are employed. A neural network algorithm is formulated to estimate a speech stress condition probability vector (with classification rates on the order of **59-100%**). The stress classification output probability vector is used to weight the outputs of a codebook of stress dependent HMM recognizers to generate an improved overall recognition score (for a **6-11%** improvement over neutral or multi-style trained recognition systems). It is suggested that this approach will accommodate the intra-speaker variability due to task induced stress in adverse conditions.

## 1 Introduction

The problem of speaker stress independent recognition requires the determination of the presence of emotionally or environmentally induced stress during human speech production. For example, when a talker is experiencing anger, the emotion is generally reflected by varying volume, duration, and pitch across the utterance [5]. The talker may also delete or emphasize portions of words while trying to express ideas in a rapid and firm manner. The physiological changes which occur in the vocal tract, and the resulting acoustic signal are dramatic and substantial. Based upon past research studies [5, 8, 14], it is often difficult to quantify the variations that occur. For example, when Apache helicopter pilots are flying, they undergo both task, physical, and emotionally induced stress (stress group $G_9$ in this study). Though a variety of studies have been performed on analysis of speech under emotion or stress, many research findings at times disagree. This is due in part to variations in how researchers simulate stressed and emotional speech, and because speakers can differ in how they vary speech production in order to convey their stress state. Hence, if a consistent range of production variations can be identified, speaker stress conditions can be grouped into classes. Past research experience indicates that there is no simple relation to describe these changes [5, 8, 14].

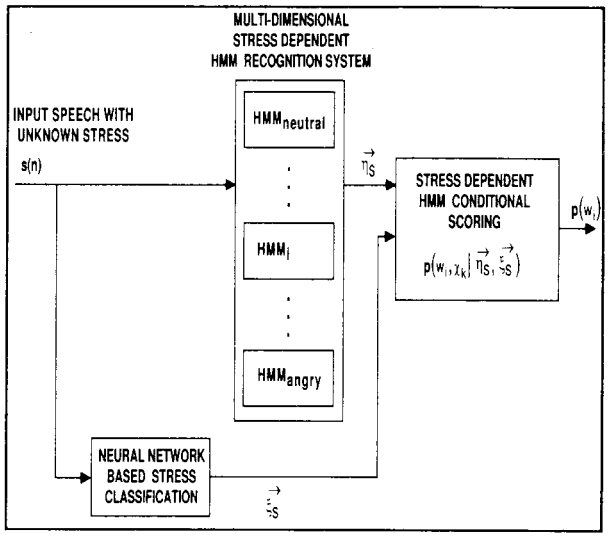


**Figure 1:** $K$**-dimensional Stress Dependent HMM Recognition System**

It is suggested that knowledge of the stress condition of a speech utterance will improve the performance of speech processing algorithms. Research in the recognition field on fast adaptation methods illustrate the benefits of modeling speaker differences. Other studies confirm the shortcomings of considering only inter-speaker variability in the speaker identification [13] and speech recognition [7] areas.

Further evaluations of the effects of speaker stress illustrate the serious impact on conventional recognition systems [4]. Stressed speech recognition algorithms have been formulated that attempt to adapt existing techniques to improve performance [4, 6, 7, 9, 11, 12]. These techniques attempt to compensate for the effects of stress or to simply study the effect of stress on performance. Stress conditions considered in these studies included *Lombard* effect, speaking styles, and task workload (e.g., computer response tasks, F-16 fighter pilot stressed speech).

The goal of this study is to incorporate estimated knowledge of the speech stress condition to improve recognition. This is accomplished by formulating a

stress classifier which directs an HMM based stress independent recognition system (see Figure 1). A tandem neural network & HMM recognition system has been shown to be effective for keyword recognition under *Lombard* effect [2]. Previous studies have explored the effect of stress on recognition applications and means to compensate for it [8, 9, 11]. The modeling framework for this study is based on a source generator framework, which allows for direct modeling of stress perturbation within a multidimensional feature space [6, 7, 8].

## 2 Stress Independent Recognition

An algorithm for stress independent recognition requires three major elements: stress independent partitioning, stress classification, and robust stress dependent recognition.

### 2.1 Stress Independent Partitioning

A separate HMM parsing algorithm is developed using training data from the TIMIT and *SUSAS* databases (see subsequent section) to partition the speech used in this study. This technique employs HMM models using five states and two mixtures. There are eight models (SI = Silence, FR = Fricatives, VL = Vowels, AF = Affricates, NA = Nasals, SV = Semi-Vowels, DT = Diphthongs) in which to represent phoneme groups.

### 2.2 Feature Analysis

An extensive evaluation on speech production features (glottal spectrum, pitch, duration, intensity, formant/spectral structure) was previously conducted [5]. An evaluation of five feature sets is conducted here to assess their potential as stress relayers [14]. The parameters are Mel-cepstral $C_i$, delta-cepstral $DC_i$, delta-delta-cepstral $D2C_i$, autocorrelation-cepstral $AC_i$, and cross-correlation-cepstral $XC_i$ parameters. The $AC_i$ and $XC_i$ features are new in that they provide a measure of the correlation between Mel-cepstral coefficients.

$$AC_i^{(\ell)}(k) = \frac{\sum_{m=k}^{m=k+L} C_i(m) * C_i(m+\ell)}{\sup_k AC_i^{(\ell)}(k)} \quad (1)$$

$$XC_{i,j}^{(\ell)}(k) = \frac{\sum_{m=k}^{m=k+L} C_i(m) * C_j(m+\ell)}{\sup_k XC_{i,j}^{(\ell)}(k)} \quad (2)$$

We considered the prospect of stressed speech classification using the five speech feature representations as stress relayers with respect to (i) pair-wise stress class separability, and (ii) analysis of acoustic tube and vocal tract cross-sectional area variation under stress. The findings suggest that perturbations in speech production under stress are reflected to varying degrees in the five speech feature representations [8, 14]. One of the proposed enhancements to the stress classification algorithm relies upon a conclusion in this study that asserts that phoneme groups (such as fricatives) are affected differently by stress. Hence, an algorithm that uses a front-end phoneme group classifier [1] should improve stress classification performance.

### 2.3 Neural Network Stress Classifier

The proposed classification algorithm involves four distinct steps to generate a stress class probability vector from raw speech data. Raw speech data is partitioned, parameterized, and classified. A neural network classifier is formulated using a Cascade Correlation network with an Extended Delta-Bar-Delta learning rule[10]. The stress classification algorithm includes three classes of features: single frame, partition, and word based parameters. This data set is created for both training and testing of the classification algorithm. Single frame based parameters are calculated for evenly spaced frames in relative positions within each source generator partition.

Partition based parameters provide summary information over an entire partition (phoneme). For all frames in partition $t$, the mean, variance, slope from leftmost (min/max) to rightmost (min/max), minimum, and maximum are calculated for each of the seven coefficients used (either $C_i$, $DC_i$, $D2C_i$, or $AC_i$). Word based parameters include the total number of frames (to incorporate a measure of word duration), and the duration of each frame.

For targeted feature stress classification, a fast back-propagation neural network is used. Both articulatory (vocal tract shape) and excitation (pitch) features are employed as well. A subset of features is targeted via a statistical evaluation of separability. A targeted parameter subset is obtained for each stress condition ($G_1$ $G_9$) and phoneme group (FR, VL, AF, NA, SV, DT) for a total of 54 subsets. Each subset has a neutral network trained to detect a given stress condition and reject all others (Table 2).

### 2.4 HMM Speech Recognition

Each of the stress dependent recognizers used in this study are trained for a single stress condition across all words in the *SUSAS* database. During the recognition phase, the HMM output probabilities $\vec{\eta}_s$ are weighted using the stress condition output probability vector $\vec{\xi}_s$ from the neural network stress classifier (see Fig. 1).

The motivation for formulating an algorithm using weighted stress trained HMMs is that the neural network stress classifier will sometimes make errors. The application of a stress condition probability weighting vector allows a codebook of HMM recognizers to compensate speech that is a mixture of stress conditions. The probability $p(w_i)$ that a given word $w_i$ under stress condition $\vec{\xi}_{sk}$ across the isolated word dictionary ($i \in (1, \ldots, I)$) is calculated as follows:

$$p(w_i) = \frac{1}{K} \sum_{k=1}^{k=K} p(w_i, \gamma_k \mid \vec{\eta}_s, \vec{\xi}_s) p(\eta_{sk}) p(\xi_{sk}) + \max_k p(w_i, \gamma_k \mid \vec{\eta}_s, \vec{\xi}_s) \quad (3)$$

where $K$ is the number of stress groups. $\vec{\xi}_s$ is the probability vector from the neural network stress classifier, and $\vec{\eta}_s$ is the probability vector from the codebook of HMM recognizers trained for word $w_i$ under stress group $k$.

## 3 Evaluations

### 3.1 *SUSAS* Speech Database

The evaluations conducted are based upon data previously collected for analysis and algorithm formulation of speech recognition in noise and stress. This database, called *SUSAS*, refers to *Speech Under Simulated and Actual Stress*, and has been employed extensively in the study of how speech production and recognition varies when speaking during stressed conditions. *SUSAS* consists of five domains, encompassing a wide variety of stresses and emotions. A total of 32 speakers (13 female, 19 male), with ages ranging from 22 to 76 are employed to generate in excess of 16,000 utterances. The five stress domains include: i) psychiatric analysis data (speech under depression, fear, anxiety), ii) talking styles. (*Angry*, *Clear*, *Fast*, *Loud*, *Slow*, *Soft*), iii) single tracking task (mild task *Cond50*, high task *Cond70*) or speech produced in noise (*Lombard* effect), iv) dual tracking computer response task, v) subject motion-fear tasks (G-force, *Lombard* effect, noise, fear). The data base 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 data base. These words consist of mono- and multi-syllabic words which are highly confusable. Examples include /go-oh-no/, /wide-white/, and /six-fix/. Some of the *actual* stressed speech utilized in this study is the same 35 word vocabulary spoken in an Apache helicopter with two speakers both on the ground and in flight. A more complete discussion of *SUSAS* can be found in the literature [5, 6, 7, 14].

| **Stress Classification Performance** | | | | |
|---|---|---|---|---|
| Single Speaker, 5 Words, Stress Grouped "Brake", "East", "Freeze", "Help", "Steer" One Network per Phoneme Group Mono-Partition Non-Targeted Features | | | | |
| Stress Group | | Classification Rate (%) | | |
| | | $C_i$ | $DC_i$ | $D2C_i$ | $AC_i$ |
| *Angry, Loud* | $G_1$ | 85.29 | 73.53 | 88.89 | 94.12 |
| *Normal, Soft* | $G_2$ | 92.69 | 89.71 | 91.04 | 86.76 |
| *Fast* | $G_3$ | 70.59 | 70.59 | 87.50 | 85.29 |
| *Question* | $G_4$ | 70.59 | 76.47 | 76.47 | 82.35 |
| *Slow* | $G_5$ | 76.47 | 52.94 | 76.47 | 64.71 |
| *Clear* | $G_6$ | 76.47 | 41.18 | 60.00 | 70.59 |
| *Lombard* | $G_7$ | 76.47 | 70.59 | 57.89 | 94.12 |
| Overall | | 78.90 | 76.94 | 79.32 | 80.62 |

**Table 1: 5 Word Phoneme Group Dependent Classification**

### 3.2 Stress Classification

Perceptually similar stress classes are grouped together for training data. The groups are indicated in Table 1 as $G_i$ where the index $i$ indicates the group number. Note that this grouping resulted from informal listening tests as to which stressed conditions are perceptually similar. Stress classes are grouped as follows: $G_1$ (*Angry, Loud*); $G_2$ (*Cond50, Cond70, Normal, Soft*); $G_3$ (*Fast*); $G_4$ (*Question*); $G_5$ (*Slow*); $G_6$ (*Clear*); and $G_7$ (*Lombard*). By grouping the classes into less confusable subgroups, performance of the classifier becomes more robust across a larger speech corpus and under noise for larger data sets. This classifier is evaluated using a 5 word stressed speech vocabulary from *SUSAS* with the results presented in Table 1. For example, under angry/loud stressed speech (group $G_1$), a parameterization using $AC_i$ provides better stress classification results than $DC_i$ (i.e., 94% vs. 73%). Similarly, for Lombard effect (group $G_7$), the stress classification rate is much higher (+36%) for $AC_i$ than for $D2C_i$. However, for $G_5$ (slow speech), the performance is lower for $AC_i$ than for $C_i$ or $D2C_i$. This is due to the durational information being filtered out by the neural network in order to minimize the global classification error.

| **Stress Classification Performance** | | | |
|---|---|---|---|
| 11 Speakers, 35 Words, Stress Grouped Targeted Features & Phoneme Dependent One Network per Phoneme & Stress Group Tri-Partition Targeted Features | | | |
| Stress Group | | Classification Rate (%) | |
| | | *Closed* | *Open* |
| *Angry, Loud* | $G_1$ | 100.00 | 58.84 |
| *Normal* | $G_2$ | 100.00 | 100.00 |
| *Fast* | $G_3$ | 100.00 | 100.00 |
| *Question* | $G_4$ | 100.00 | 100.00 |
| *Slow* | $G_5$ | 100.00 | 100.00 |
| *Clear* | $G_6$ | 100.00 | 100.00 |
| *Lombard* | $G_7$ | 100.00 | 100.00 |
| *Soft* | $G_8$ | 100.00 | 100.00 |
| *Actual* | $G_9$ | 66.67 | 62.75 |
| Overall | | 96.30 | 91.29 |

**Table 2: 35 Word Phoneme Group Dependent Classification**

The evaluation is extended to include 35 words and 11 speakers. Features used include the $AC_i$ and their derived features, durational, articulatory, and excitation features. One back-propagation network is trained for each phoneme group. Since the variability of the data and the size of the pattern vector are so large, this approach did not perform satisfactorily.

Next, a study is performed as previously discussed to find a targeted subset of features for each stress condition and phoneme group. This results in significantly smaller pattern vectors that contain more meaningful features for detection of stress, assuming knowledge of the phoneme group. Stress classes are regrouped as follows: $G_1$ (*Angry, Loud*); $G_2$ (*Normal*); $G_3$ (*Fast*); $G_4$ (*Question*); $G_5$ (*Slow*); $G_6$ (*Clear*); $G_7$ (*Lombard*); $G_8$ (*Soft*); and $G_9$ (*Actual*). Table 2 shows the classification performance of this ensemble of neural networks. Each neural network was tested on its ability to accept and reject a given stress condition for each speech token.

| *SUSAS* Recognition Performance | | | | | | | |
|---|---|---|---|---|---|---|---|
| 11 Speakers, 35 Words, Stress Grouped Simulated & Actual Stress<br>HMM Neutral, Multi-Style, and Stress Dependent Training | | | | | | | |
| Stress Group | | Neutral (%) | | Multi-Style (%) | | Dependent (%) | |
| | | *Closed* | *Open* | *Closed* | *Open* | *Closed* | *Open* |
| *Angry, Loud* | $G_1$ | 49.03 | 53.01 | 59.42 | 55.61 | 82.79 | 67.32 |
| *Normal* | $G_2$ | 95.78 | 79.35 | 69.81 | 66.02 | 95.78 | 79.35 |
| *Fast* | $G_3$ | 66.88 | 66.34 | 55.19 | 53.33 | 75.32 | 58.86 |
| *Question* | $G_4$ | 66.56 | 64.23 | 62.34 | 60.49 | 89.94 | 70.89 |
| | Overall | 69.56 | 65.73 | 61.69 | 58.86 | 85.96 | 69.11 |

**Table 3: Recognition Comparison**

### 3.3 Stress Independent Recognition

Next, a speaker dependent, isolated word, discrete-observation hidden Markov model recognizer is employed. A separate HMM is obtained for each word in the system dictionary. The HMM used is a five state left-to-right model, with each model beginning in state 1. In the training phase, each model is initiated with essentially random choices for non-zero elements and then iteratively adjusted to increase $P(\vec{\Phi}|\underline{\mathbf{M}})$, the probability of the observation sequence $\vec{\Phi}$ having been generated by model $\underline{\mathbf{M}}$. The training technique is based on the Baum-Welch forward-backward reestimation algorithm.

Our preliminary results (see Table 3) show a **6-11%** improvement in recognition performance over neutral or multi-style trained HMM models. The following four stressed speech recognition evaluations will be considered:

1. **Neutral HMM**
   Establishes a lower bound on performance using an HMM recognizer trained on *Normal* speech and evaluated across stressed speech conditions (see Table 3.
2. **Multi-Style Trained HMM**
   For the purposes of comparison with previous methods [9], a single multi-style trained HMM recognizer is also evaluated across all stress styles.
3. **Non-Weighted Stress Trained HMMs**
   It is possible that a codebook of stress dependent HMM recognizers could provide improved recognition performance without requiring an estimated stress classification probability vector. This evaluation will establish the benefit of employing a codebook of HMM recognizers that span a particular stressed speech region in a source generator space [6, 8]. This will contrast with a single multi-style trained HMM.
4. **Weighted Stress Trained HMMs**
   An estimated stress condition probability vector will be used to weight the codebook of HMM probability outputs. The key to this formulation is the stress classification performance. The results in Table 2 indicate that stress classification using targeted feature sets is possible.

## 4 Summary

The use of stress classification in combination with HMM recognition results in a viable robust stress independent speech recognition system. It is shown that partitioning of stressed speech by phoneme group is a contribution to the problem of stressed speech recognition that facilitates stress classification for stress independent recognition.

## References


[1] L. M. Arslan and J. H. L. Hansen, "A Mimimum Cost based Phoneme Class Detector for Improved Iterative Speech Enhancement," *ICASSP*, April 1994.

[2] G. J. Clary and J. H. L. Hansen, "A Novel Speech Recognizer for Keyword Spotting," *ICSLP*, pp. 13-16, Oct. 1992.

[3] Y. Gong, "Speech Recognition in Noisy Environments : A Survey," *Speech Communication*, vol. 16, no. 3, pp. 261-291, 1995.

[4] B. A. Hanson and T. Applebaum, "Robust Speaker-Independent Word Recognition Using Instantaneous, Dynamic and Acceleration Features: Experiments with Lombard and Noisy Speech," *ICASSP*, pp. 857-60, April 1990.

[5] J. H. L. Hansen, *Analysis and Compensation of Stressed and Noisy Speech with application to Robust Automatic Recognition*. PhD thesis, Georgia Institute of Technology, July 1988. 428 pages.

[6] 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 Proc.*, Oct. 1994.

[7] J. H. L. Hansen, "Adaptive Source Generator Compensation and Enhancement for Speech Recognition in Noisy Stressful Environments," *ICASSP*, pp. 95-98, April 1993.

[8] J. H. L. Hansen, B. D. Womack, and L. M. Arslan, "A Source Generator based Production Model for Environmental Robustness in Speech Recognition," *ICSLP*, pp. 1003-6, September 18-22 1994.

[9] R. P. Lippmann, E. A. Martin, and D. B. Paul, "Multi-Style Training for Robust Isolated-Word Speech Recognition," *ICASSP*, pp. 705-708, April 1987.

[10] A. A. Minai and R. D. Williams, "Back-Propagation Heuristics: A Study of the Extended Delta-Bar-Delta Algorithm," *IJCNN*, pp. 595-600, June 17-21 1990.

[11] D. B. Paul, "A Speaker-Stress Resistant HMM Isolated Word Recognizer," *ICASSP*, pp. 713-16, April 1987.

[12] P. K. Rajasekaran, G. R. Doddington, and J. W. Picone, "Recognition of Speech Under Stress and In Noise," *ICASSP*, pp. 733-36, April 1986.

[13] F. K. Soong and A. E. Rosenberg, "On the Use of Instantaneous and Transitional Spectral Information in Speaker Recognition," *IEEE Trans. on ASSP*, vol. 36, pp. 871-879, June 1988.

[14] B. D. Womack and J. H. L. Hansen, "Feature Analysis and Neural Network based Classification of Speech under Stress," *IEEE Trans. Speech & Audio Proc.* (submitted).