IMPROVED SPEECH RECOGNITION VIA SPEAKER STRESS DIRECTED CLASSIFICATION

Brian D. Womack and *John H.L. Hansen*
Robust Speech Processing Laboratory
Duke University, Box 90291, Durham, NC 27708-0291
http://www.ee.duke.edu/Research/Speech womack@ee.duke.edu jhlh@ee.duke.edu

ABSTRACT

Speech production variations due to perceptually induced stress contribute significantly to reduced speech processing performance [2]. This study proposes an algorithm for estimation of the degree of perceptually induced stress. It is suggested that the resulting stress score could be integrated into speech processing algorithms to improve robustness in adverse conditions. First, results from a previous study motivate selection of a targeted set of speech features across phoneme and stress groups to improve stress classification performance. Analysis of articulatory, excitation, and cepstral based features is conducted using a previously established stressed speech database (SUSAS). Targeted feature sets are selected across ten stress conditions (including *Apache* helicopter, *Angry, Clear, Lombard* effect, *Loud*, etc.). Next, an improved targeted feature stress classification system is developed and evaluated achieving rates of **91.01%**. Finally, application of stress classification is incorporated into a stress directed speech recognition system. An improvement of **+10.14%** and **+15.43%** over conventionally trained neutral and multi-style trained recognizers is demonstrated using the new stress directed recognition system.

1 INTRODUCTION

Speaker stress assessment is useful for applications such as emergency telephone message sorting. Here, stress can be defined as any effect that causes the speaker to vary the production of speech from neutral conditions. Neutral speech is defined as speech produced assuming that the speaker is in a "quiet room" with no task obligations but to speak. With this definition, two stress effect areas emerge: perceptual and physiological. Perceptually induced stress result from the speaker's perception that the environment is not "normal" such that the speech production *intention* varies from neutral conditions. Causes of perceptually induced stress include emotion, environmental noise (*Lombard* effect), actual task workload (*Apache* helicopter cockpit), statement context (*Question*), and speaking tempo (*Fast*, *Slow*). Physiologically induced stress is the result of a physical impact on the human body which results in deviations from neutral speech production despite intentions. Causes of physiological stress include vibration, G-force, drug interactions, and air density. In this study, ten perceptually induced stress conditions are considered (*Angry, Apache, Clear, Fast, Lombard, Loud, Neutral, Question, Slow, Soft*).

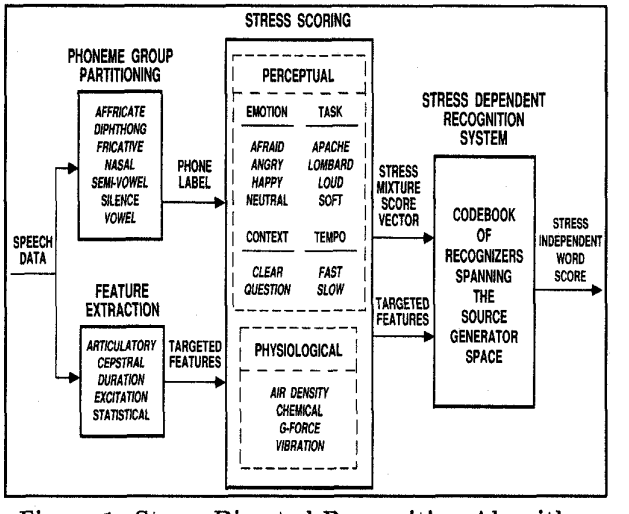

Figure 1: Stress Directed Recognition Algorithm

The problem of speaker stress classification is to assess the degree to which a specific stress condition is present in a speech utterance. Past research studies indicate that it is difficult to quantify these variations [2]. The variation in speech production due to stress can be substantial; and, will therefore have a direct impact upon the performance of speech processing applications if not addressed [7]. There have been a number of studies performed on analysis of speech under stress in an effort to identify meaningful relayers of stress. Unfortunately, many research findings at times disagree, due in part to the variation in the experimental design protocol employed to induce stressed speech; and, to differences in how speakers impart stress in their speech production. Past research experience suggests that no simple relationship exists to describe these changes [2, 7]. Though a number of studies have considered analysis of speech under stress, the problem of stressed speech classification, to our knowledge, has not been addressed in the literature except for one study on detection of stressed speech using the Teager nonlinear energy operator [1]. Previous studies directed specifically at robust speech recognition differ from this study in that they estimate intra-speaker variations via speaker adaptation, front-end stress compensation, or wider domain training sets [3, 5, 7].

0-7803-3192-3/96 $5.00©1996 IEEE

The assertion that a stress directed approach is better able to represent feature perturbations due to stress is supported by the observation that spectral, excitation, articulatory, and time domain speech features cluster in different domains between two stress conditions [7]. Accurate representation of intra-speaker variability due to stress and noise is a limitation of speech processing algorithms that has been demonstrated in related studies on speaker identification and speech recognition [4]. The incorporation of stressed speech into speech processing algorithms has been applied previously to improve the performance of recognition systems [2, 3, 5, 7].

In order to understand speech production under stress, an extensive stress evaluation on speech production features such as glottal spectrum, pitch, duration, intensity, and formant/spectral structure was previously conducted [2]. Extensive statistical assessment of over 200 parameters for simulated and actual speech under stress suggests that stress classification based upon feature distribution separability characteristics is possible. A subsequent evaluation of features for application to stress classification was conducted using five stressed speech feature representations [7]. These feature sets were considered with respect to (i) pair-wise stress class separability, and (ii) analysis of acoustic tube and vocal tract cross-sectional area variation under stress. Feature analysis suggests that perturbations in speech production under stress are reflected to varying degrees across multiple feature domains depending upon the stress condition and phoneme group. Hence, an algorithm that uses a front-end phoneme group classifier [6] can improve overall stress classification performance [7]. A tandem neural network and HMM recognition system has also been shown to be effective for recognition under several stress conditions including *Lombard* effect [7]. Finally, previous studies have also explored the effect of stress on recognition and approaches for compensation [2, 3, 5].

The goal of this study is to incorporate knowledge of speech stress content to improve recognition. A stress classifier is formulated which directs an HMM based stress dependent recognition system as shown in Fig. 1.

2 STRESS INDEPENDENT RECOGNITION FORMULATION

An algorithm for stress independent processing can be formulated with three major elements: (i) stress independent partitioning, (ii) stress classification, and (iii) robust stress dependent processing (recognition, enhancement, or speaker identification/verification).

2.1 Stress Independent Partitioning

A partitioning algorithm that provides consistent speech parsing is a difficult task due to gradual transitions between phonemes, the impact of stress, and coarticulation effects. However, in a previous study on robust speech partitioning [6], an algorithm is formulated using HMM and Viterbi decoding to parse noise corrupted speech by phoneme group. In a variation of this approach, the HMM models are trained using data from the neutral TIMIT and stressed SUSAS speech databases. Each continuous density phoneme group HMM model has five states per phoneme with two mixtures. The eight models (SI:Silence, FR:Fricatives, VL:Vowels, AF:Affricates, NA:Nasals, SV:Semi-Vowels, DT:Diphthongs) form trained word grammars composed of phoneme group sequences. The Viterbi decoding then matches the state sequence to the grammar for each input word to find phoneme boundaries.

2.2 Target Driven Features

In a previous study [7], articulatory, excitation, and cepstral based feature domains were assessed for application to stress classification. A master feature set is created from which subsets of targeted features are selected using a separability distance metric and feature separability ranking based on statistical and subjective measures. From the articulatory feature domain, cross sectional vocal tract areas are used in the master feature set. For the excitation feature domain, pitch and duration are used. Finally, from the cepstral domain, auto-correlation Mel-cepstral features and their statistics (mean, standard deviation, and slope) are included. For each phoneme group and stress condition, a subset of these features is selected for a targeted feature stress detection system. Next, this codebook of stress detection systems are combined to form the stress classification algorithm.

2.3 Stress Classifier Formulation

In formulating a stress classification system, it is necessary to realize that for a given stress condition, there are degrees of stress. Hence, it is necessary to estimate a stress probability response score to assess the stress level. However, a stress score for one stress condition alone is not sufficient, since it does not model mixtures of stress states. In order to model stress state combinations, also referred to as stress mixtures, a stress score must be obtained for each stress condition. This score is estimated by training a stress detector to recognize one stress condition given knowledge of the phoneme group. A codebook of these stress detectors will then provide an estimate of the degree of each stress state. The resulting vector of stress scores is then a more complete representation of the speaker's stress state. This formulation is based upon a mathematical framework that represents feature movement from one region in a source generator space to another (where each region is a stress state) [2].

Next, a general stress detection system employing neural networks is developed to estimate a stress score $p(\xi_k \mid w_i)$; which measures the degree of stress given utterance w_i spoken under stress condition k (Fig. 1).

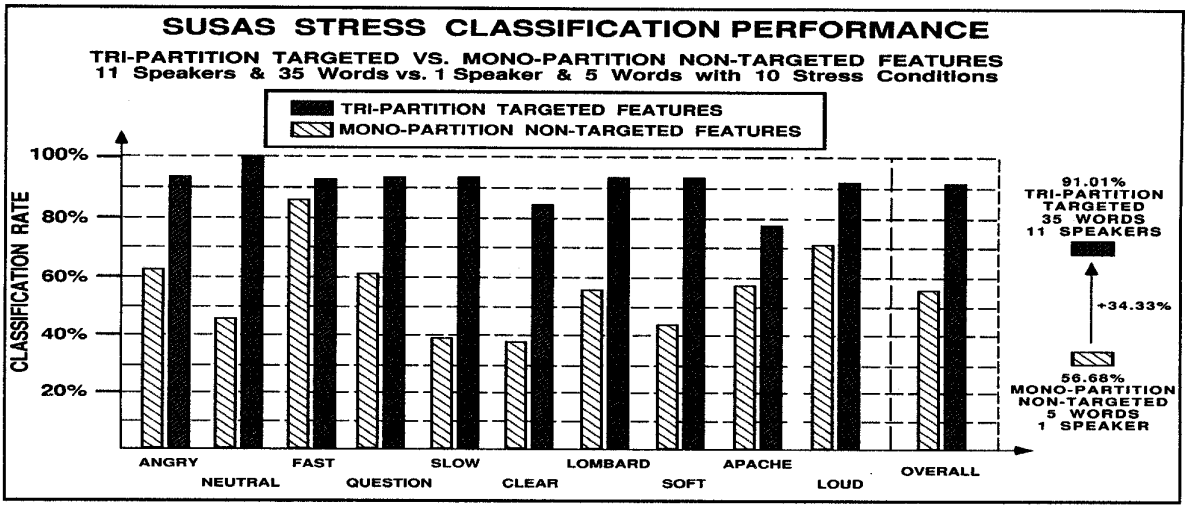

Figure 2: Stress Classification Performance

2.4 HMM Speech Recognition

In order to formulate a codebook of stress dependent recognizers, it is desirable to use an existing HMM speech recognition framework as shown in Fig. 1. The system incorporates stress class information in the source generator space by including data from each stressed speech region in the training of each stress dependent recognizer [2]. This is equivalent to maximizing the word log probability $p(w_i \mid \xi_k)$, given the overall word stress score ξ_k. The word stress score is calculated by averaging the scores across all partitions for a candidate word. These word score vectors, denoted $\vec{w}_k = \{w_{ik} \mid i = 1, \ldots, I\}$, are obtained from a codebook of speech recognizers spanning the source generator space. Once the maximum stress probability $p(\xi_{kmax} \mid w_i)$ has been calculated as in Eqn. (1), the speech features are passed to the stress dependent recognition system trained for stress condition $kmax$. The final utterance decision is then calculated with Eqn. (2) by maximizing $p(w_i \mid \xi_{kmax})$ over every word in the vocabulary.

$$p(\xi_{kmax} \mid w_i) = \max_k p(\xi_k \mid w_i) \quad (1)$$

$$p(w_{imax}) = \max_i p(w_i \mid \xi_{kmax}) \quad (2)$$

At this point, the four preceeding components including stress independent partitioning, target driven features, stress classification, and HMM formulation can be integrated as shown in Fig. 1, for a stress independent recognition system. To achieve reliable performance, phonemic class partitioning is used to temporally divide input speech into a source generator sequence. With these phoneme labels, targeted features are extracted and passed to the stress classification algorithm. It is of interest to determine the importance of isolated versus context dependent stress classification using a partitioned phoneme sequence. To accomplish this, the following two stress classifiers are evaluated in Section 3: mono-partition and tri-partition. Given a reliable stress classifier, the last step is to determine whether such information can improve the robustness of speech recognition under stress.

3 EVALUATIONS

To illustrate the application of stress information to speech recognition, a series of simulations are performed using the SUSAS speech under stress database [7].

3.1 Stress Classification (Tri-Partition)

Mono-partition and tri-partition stress classifier results are shown in Fig. 2 using a 35 word vocabulary set. The following features are made available to both classifiers: auto-correlation Mel-Cepstral parameters and their derived features, durational, articulatory, and excitation. In order to reduce data requirements for tri-partition classification, a targeted feature subset is selected for each stress condition and phoneme group. This results in a smaller and more meaningful feature set for stress detection. Each stress classifier consists of a codebook of neural networks, one for each phoneme group and stress condition. As Fig. 2 illustrates, when using isolated phonemes (mono-partition), measurable stress classification performance can be achieved. However, when the stress classifier is based upon a context dependent phoneme sequence (tri-partition), performance significantly improves by **+34.33%** [7].

3.2 HMM Stress Independent Recognition

Next, a speaker dependent, isolated word, continuous density hidden Markov model recognizer is used. The HMM training method in this study employs a state tying initialization based upon the degree of similarity between mean mixture vectors in successive states. The models assume left-to-right transitions without skips. The training algorithm is based on the Baum-Welch forward-backward reestimation algorithm.

Three stressed speech recognition evaluations are considered (Fig. 3). To establish a baseline level of performance, the first evaluation employs neutral

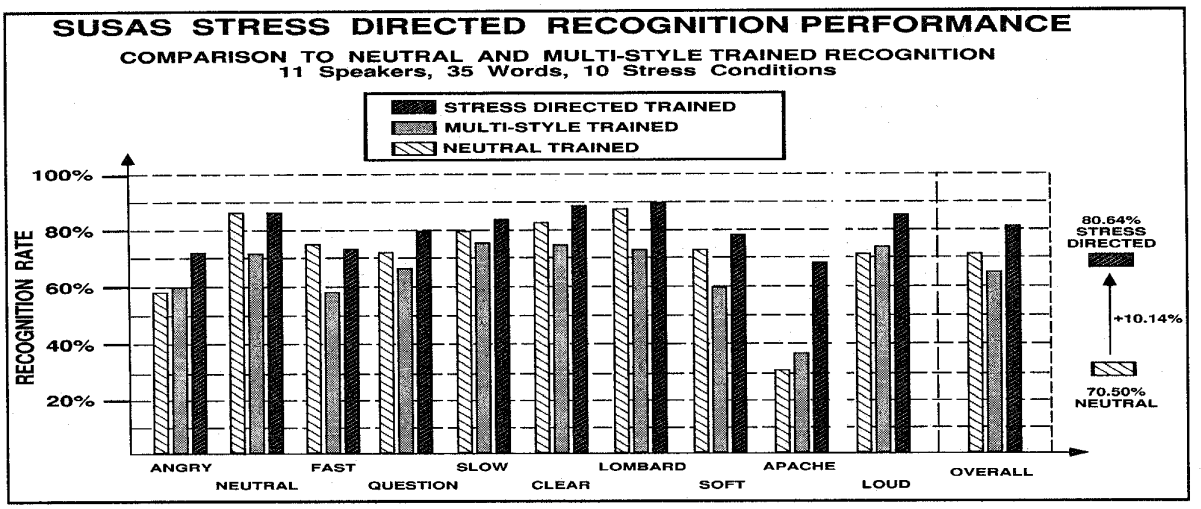

Figure 3: Stress Directed Recognition Comparison

trained HMMs that are tested with stressed SUSAS data. An overall open test recognition rate of **70.50%** is achieved, with performance ranging from 33% for *Apache* to 87% for *Neutral* speech. It is noted that recognition is most severely affected by *Angry* and *Apache* speech due to their feature distributions; which are well separated from neutral speech. The second evaluation focuses upon multi-style trained HMMs. This approach differs from a previous study [5] in that training is speaker independent and speech is sampled at 8kHz. An overall open test recognition performance of **65.21%** is achieved; which is **-5.29%** lower than the neutral trained HMM. The third evaluation assumes estimated knowledge of the speaker stress state from a tandem neural network stress classifier and HMM recognizer trained for each stress condition. The stress directed recognition rate is **80.64%**; which is **+10.14%** more than neutral trained, and **+15.43%** more than the multi-style trained HMM. This codebook of stress dependent HMM recognizers provides improved recognition performance using estimated stress classification knowledge. This evaluation has served to illustrate the benefit of a stress directed formulation which encompasses general speech production as reflected in a source generator space [2]. It should also be noted that we believe the speaker independent multi-style trained HMM attempts to represent a broader range of speech production with diminishing returns as the level of stress increases.

4 SUMMARY

In this study, we have considered (i) improved stress classification using targeted features and (ii) the formulation of a robust stress independent recognition system. A probability vector representing the degree of speaker stress is estimated by a classification algorithm. It was shown that context sensitive stress classification via tri-partitioning achieves better performance. The output stress probability vector can be employed to measure mixtures of speaker stress (e.g., combined *Fast* and *Loud* speech).

A robust stress independent recognition system was formulated, consisting of the following: (i) stress independent speech partitioning, (ii) feature targeted stress classification, and (iii) stress dependent recognition. Stress classification performance is improved when speech is partitioned by phoneme group, target driven features are used, and excitation and articulatory features are employed. A codebook of neural network stress detectors are used to estimate the speaker stress mixture probability vector. This vector is then used to select the most probable stress dependent HMM recognizer, with an improvement of **+10.14** to **+15.43%** over neutral and multi-style trained systems.

References

[1] D. A. Cairns and J. H. L. Hansen, "Nonlinear Analysis and Detection of Speech Under Stressed Conditions," *J. Acous. Soc. Am.*, vol. 96, pp. 3392–400, December 1994.

[2] 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, 1994.

[3] J. H. L. Hansen, "Analysis and Compensation of Speech Under Stress & Noise for Environmental Robustness in Speech Recognition," *ECSA-NATO Proc. of Speech Under Stress Workshop*, pp. 91–8, September 1995. Lisbon, Portugal.

[4] H. Lee and A. Tsoi, "Application of Multi-Layer Perceptron in Estimating Speech / Noise Characteristics for Speech Recognition in Noisy Environment," *Speech Communication*, vol. 17, pp. 59–76, August 1995.

[5] 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.

[6] B. L. Pellom and J. H. L. Hansen, "Text-Directed Speech Enhancement using Phoneme Classification and Feature Map Constrained Vector Quantization," *ICASSP*, 1995.

[7] B. D. Womack and J. H. L. Hansen, "Stress Independent Robust HMM Speech Recognition using Neural Network Stress Classification," *EuroSpeech*, pp. 1999–2002, 1995.