ELSEVIER

Speech Communication 20 (1996) 131–150

SPEECH COMMUNICATION

Classification of speech under stress using target driven features [1]

Brian D. Womack [2], John H.L. Hansen *

Robust Speech Processing Laboratory, Duke University, Department of Electrical Engineering, Box 90291, Durham, NC 27708-0291, USA

Received 30 January 1996; revised 15 June 1996

Abstract

Speech production variations due to perceptually induced stress contribute significantly to reduced speech processing performance. One approach for assessment of production variations due to stress is to formulate an objective classification of speaker stress based upon the acoustic speech signal. This study proposes an algorithm for estimation of the probability of perceptually induced stress. It is suggested that the resulting stress score could be integrated into robust speech processing algorithms to improve robustness in adverse conditions. First, results from a previous stress classification study are employed to motivate selection of a targeted set of speech features on a per phoneme and stress group level. Analysis of articulatory, excitation and cepstral based features is conducted using a previously established stressed speech database (Speech Under Simulated and Actual Stress (SUSAS)). Stress sensitive targeted feature sets are then selected across ten stress conditions (including *Apache* helicopter cockpit, *Angry, Clear, Lombard effect, Loud*, etc.) and incorporated into a new targeted neural network stress classifier. Second, the targeted feature stress classification system is then evaluated and shown to achieve closed speaker, open token classification rates of 91.0%. Finally, the proposed stress classification algorithm is incorporated into a stress directed speech recognition system, where separate hidden Markov model recognizers are trained for each stress condition. An improvement of +10.1% and +15.4% over conventionally trained neutral and multi-style trained recognizers is demonstrated using the new stress directed recognition approach.

Zusammenfassung

Variation der Sprachproduktion wegen Stress und Rauschen tragen stark zu einer Verminderung der Sprachverarbeitungsleistung bei. Ein Ansatz zur Betrachtung von Produktionsvariationen wegen Stress ist, eine objektive Klassifikation von Spracherstress, basierend auf akustischen Sprachsignalen, vorzunehmen. Diese Untersuchung schlägt einen Algorithmus zur Abschätzung des induzierten Stress vor. Es wird vorgeschlagen, die resultierende Stressquelle in robuste Sprachverarbeitungsalgorithmen zu integrieren, um die Robustheit zu erhöhen. Zunächst werden die Ergebnisse einer früheren Stressklassifikationsstudie einbezogen und vorgestellt, um die Wahl der Zielmenge von Spracheigenschaften auf einer Stressgruppenebene zu motivieren. Eine Analyse von Artikulation und Aussprache-Eigenschaften wird durchgeführt unter Verwendung einer bereits früher aufgestellten Sprachdatenbank (Speech Under Simulated and Actual Stress (SUSAS)). Die stresssensitiven Zieleigenschaften werden dann aus einer Menge von 10 Stressumgebungen (eingeschlossen *Apache* Helikopter Cockpit, *wütend, klar, Lombard Effekt, laut*, etc.) ausgewählt und in ein neues stressklassifizierendes neuronales

* Corresponding author. E-mail: jhlh@ee.duke.edu; http://www.ee.duke.edu/Research/Speech.

[1] Audiofiles available. See http://www.elsevier.nl/locate/specom .

[2] E-mail: womack@ee.duke.edu.

PII S0167-6393(96)00049-0

Netzwerk integriert. Das betrachtete Stressklassifikationssystem wird dann ausgewerted und es wird gezeigt, dass geschlossene Sprecher, offene Tokenklassifikationsraten von 91.0% erreicht werden. Zum Schluss wird der vorgeschlagene Stressklassifikationsalgorithmus eingebaut in ein auf Stress ausgerichtetes Spracherkennungssystem, in dem separate versteckte Markov Modell Erkenner trainiert werden für jede Stresssituation. Eine Verbesserung von +10% und +15.4% gegenüber konventionell trainierten neutralen und multi-style trainierten Erkennern wird durch Verwendung dieses neuen stressgerichteten Ansatzes erreicht.

Résumé

Les variations dans la production de parole dues au stress induit contribuent de manière significative à la réduction des performances des systèmes de traitement de parole. Pour estimer ces variations, une approche consiste à établir une classification objective du stress du locuteur, basée sur le signal acoustique. Cette étude propose un algorithme pour l'estimation de la probabilité du stress induit. Le taux de stress prédit par cet algorithme peut être intégré dans des algorithmes de traitement de parole afin d'augmenter leur robustesse dans des environnements difficiles. Les résultats d'une étude précédente sur la classification du stress sont d'abord utilisés pour sélectionner un ensemble de paramètres de parole relatifs au phonème et au type de stress. Une analyse des paramètres articulatoires, d'excitation et cepstraux est conduite sur une base de données de parole sous stress ("Speech Under Simulated and Actual Stress" (SUSAS)). Les paramètres sensibles au stress sont ensuite sélectionnés pour dix conditions de stress (incluant le cockpit d'un hélicoptère *Apache*, *la colère*, *la parole claire*, *l'effet Lombard*, *la voix forte*, etc.) et sont incorporés dans un réseau de neurones appris pour classifier le degré de stress. Dans une deuxième partie, le système de classification du stress basé sur les paramètres précédents est évalué. Sur un ensemble fermé de locuteurs et pour un ensemble ouvert de stimuli de parole, il produit un taux de bonne classification de 91.0%. Finalement, l'algorithme de classification du stress est incorporé dans un système de reconnaissance de parole où un modèle de Markov est appris pour chaque condition de stress. Avec cette nouvelle approche de reconnaissance "dépendante du stress", on obtient une amélioration des performances de 10.1% et de 15.4%, respectivement, par rapport aux systèmes de reconnaissance appris avec de la parole neutre et avec différents styles de parole.

1. Introduction

The problem of speaker stress classification is to assess the degree to which a specific stress condition is present in a speech utterance. "Stress" in this study refers to perceptually induced variations on the production of speech. Past research studies indicate that it is difficult to quantify these variations. The change 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 (Womack and Hansen, 1995). A number of studies in the past have been performed on analysis of speech under stress in an effort to identify meaningful relayers of stress (Lieberman and Michaels, 1962; Simonov and Frolov, 1977; Williams and Stevens, 1972). 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 (Hansen, 1988, 1995b; Hansen and Womack, 1996).

Though a number of studies have considered analysis of speech under stress, the problem of stressed speech classification has received little if any attention in the literature. One exception is a study on detection of stressed speech using a parameterized response obtained from the Teager nonlinear energy operator (Cairns and Hansen, 1994). Previous studies directed specifically at robust speech recognition differ in that they estimate intraspeaker variations via speaker adaptation, front-end stress compensation, or wider domain training sets. 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 MCE-ACC (Hansen, 1994) employ adaptive cepstral compensation with morphologically constrained fea-

ture enhancement to improve recognition performance. Finally, larger training sets have been considered for stressed speech in the training phase. Most notably, the multistyle training algorithm (Lippmann et al., 1987) 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 (Hansen and Bou-Ghazale, 1995). However, for speaker *independent* systems, it has been shown that multi-style training results in a loss of performance over a neutral trained system (Womack and Hansen, 1995). The cause of this is believed to be due to additional stress related inter-speaker feature variations which the recognition models must now represent, resulting in a decrease in the ability to discriminate between words.

For the problem of stress classification, there are two major application areas: objective stress assessment and 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 useful for applications such as emergency telephone message sorting and aircraft voice communications monitoring. A stress classification system could provide meaningful information to speech algorithms for recognition, speaker verification, synthesis and coding.

The main focus of this study is to formulate a stress classification system as shown in Fig. 1. This general stress classification system assumes that input speech is parsed by phoneme class. With knowledge of the phone class, a set of stress differentiating targeted features could be formulated that is better able to detect stress characteristics. Next, a high level classifier could determine whether the input speech is spoken under perceptually or physiologically induced stress. Finally, a codebook of classifiers could detect each of the specific stress conditions under evaluation. In this study, phoneme group partitioning, targeted feature extraction and perceptually induced stress classifiers are evaluated as part of this theoretical system.

Before venturing into the formulation of a stressed speech classification algorithm, it would be useful to identify areas where speech processing research has centered on speech under stress. The effects of stress have been indirectly addressed by formulating a more accurate speech production representation of intra-speaker variability for the speaker identification (Soong and Rosenberg, 1988) and speech recognition (Lee and Tsoi, 1995) problems. Stressed speech analysis has yielded better modeling approaches for speech production which have been successfully ap-

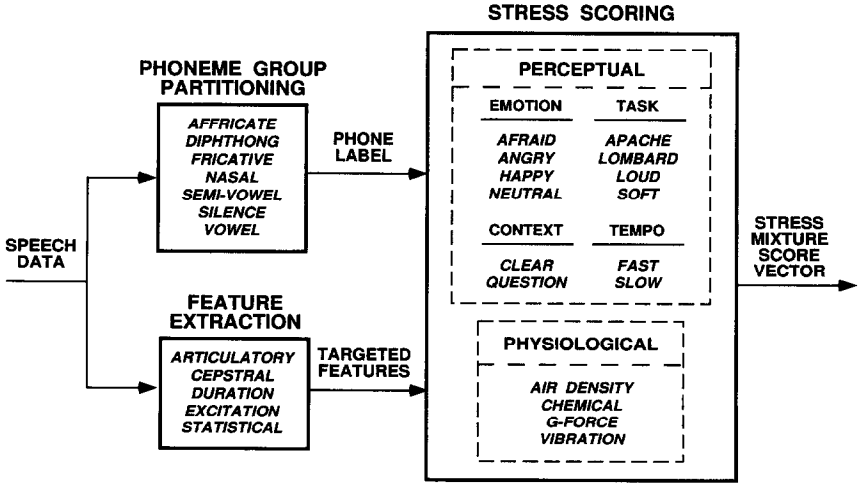

Fig. 1. Stress classification formulation.

plied to improve speech recognition performance (Hansen, 1995a; Hansen and Clements, 1995; Womack and Hansen, 1995, 1996). The incorporation of stressed speech modeling into speech processing algorithms has been applied previously to improve the performance of recognition systems (Hansen, 1995a; Lippmann et al., 1987; Womack and Hansen, 1995). Stress conditions considered in these studies include perceptually induced stress such as *Lombard* effect or task workload (e.g., computer response tasks, F-16 fighter pilot stressed speech (Stanton et al., 1989)). In another study, a novel stress equalization scheme was formulated using a tandem neural network and hidden Markov model recognition system which was shown to be effective for keyword recognition under several stress conditions including *Lombard* effect (Clary and Hansen, 1992). 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 (Hansen, 1993, 1994; Hansen et al., 1994). 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 (Hansen, 1988, 1995b). 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.

In this study, the problem of classification of speech under stress is addressed. Since stress can influence a variety of factors in speech production (i.e., physical production, speaker rate, word selection, sentence construction, etc.), the focus here is only on isolated words and stress exhibited from an overall perspective on a limited male speaker set. The first phase of this study requires that speech production, analysis and recognition features be analyzed with respect to their ability to differentiate speaker stress (Section 2). Given this knowledge, a set of targeted feature sets is determined, and employed in the formulation of a neural network based stress classification algorithm (Section 3). Next, in Section 4, the stress classification algorithm is evaluated using a speech under stress database (SUSAS) for (i) feature targeting, (ii) stress classification, and (iii) speech recognition. Finally, conclusions are summarized in Section 5.

2. Classification features for stressed speech

Before embarking on our study of stressed speech classification features, it may be useful to distinctly define stress in our context. Stress can be defined as any condition which causes a speaker to vary their 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. With this definition, two stress effect areas emerge: perceptual and physiological. Perceptually induced stress results when a speaker perceives their environment to be different from "normal" such that their *intention* to produce speech varies from *Neutral* conditions. The causes of perceptually induced stress include emotion, environmental noise (i.e., *Lombard* effect (Junqua, 1993; Lombard, 1911)), 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 which 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 ten perceptually induced stress conditions from the SUSAS database are considered: *Angry, Apache, Clear, Fast, Lombard, Loud, Neutral, Question, Slow, Soft.*

In order to formulate algorithms for stress classification, it would be useful to consider the type of speech production variations that occur in response to perceptually induced speaker stress. It is hypothesized that better stress classification performance can be achieved by characterizing stress induced production variations for each stress and phoneme group; so that stress sensitive feature sets may be selected. Previous studies have considered features from speech production domains such as pitch, duration, intensity, glottal source effects, and vocal tract spectrum. In this study, the focus is upon features derived from speech produced in the following three domains: (i) articulatory, (ii) excitation and (iii) cepstral. To accomplish this, it is assumed that the input speech has been parsed consistently by phoneme

group using a previously established phone class parser (Pellom and Hansen, 1996). The input speech under test is therefore automatically parsed and labeled (details on the parsing algorithm are presented in Section 3.1) using the following seven phoneme groups: SI: Silence, FR: Fricatives, VL: Vowels, AF: Affricates, NA: Nasals, SV: Semi-Vowels, and DT: Diphthongs. Speech features are then extracted in order to investigate the ability to perform stress classification across different partitioning levels. Frame-level features include articulatory, excitation and spectral characteristics of the speech signal. Partition-level features are used to provide statistics of the frame-level features over an entire partition. Finally, word-level features incorporate broad aspects of the word.

2.1. SUSAS speech database

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, 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 (i) psychiatric analysis data (speech under depression, fear, anxiety), (ii) talking styles [3] (*Angry, Clear, Fast, Loud, Slow, Soft*), (iii) single tracking task (mild task *Cond50*, high task *Cond70* computer response workload) or speech produced in noise (*Lombard effect*), (iv) dual tracking computer response task, and (v) subject motion-fear tasks (G-force, *Lombard* effect, noise, fear). 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 multi-syllabic words which are highly confuseable. Examples include /go–oh–no/, /wide–white/ and /six–fix/. A more complete discussion of SUSAS can be found in the literature (Hansen, 1994, 1995b).

The SUSAS speech employed in this study consists of a thirty-five word aircraft vocabulary from nine male speakers under simulated stress and two male speakers under actual stress. Simulated stressed speech conditions considered include *Angry, Clear, Fast, Lombard, Loud, Question, Slow, Soft* speech. Actual stressed speech conditions considered include *Apache* helicopter cockpit speech during warmup on a runway and in flight.

2.2. Feature targeting methodology

In a previous study (Womack and Hansen, 1995), articulatory, excitation and cepstral based feature domains were considered for application to stress classification. A master feature set was created from which subsets of targeted features could be selected. This selection was based on a separability distance measure and feature ranking using statistical and subjective measures. In the present study, a subset of these features is selected for each phoneme group and stress condition in order to formulate a targeted feature stress detection system. Next, the resulting codebook of stress detectors (i.e., across each phone group and stress condition) are combined to form the overall stress classification algorithm.

In order to rank order the set of speech features for stress classification, a performance criterion is needed. Here, the term "good" or "useful" is used to describe how reliable a feature is for stress detection using a feature separability score. The remainder of this section describes a feature ranking system. The process of feature targeting for each stress condition and phoneme group requires three stages: (i) feature differentiability across stress conditions, (ii) compilation of the best features for each stress condition, and (iii) compilation of the best features for a combined phoneme group and stress condition.

A feature's ability to differentiate stress conditions is graded (A, B or C) based upon how well a single feature is capable of distinguishing one or

[3] Approximately half of SUSAS consists of style data donated by Lincoln Laboratories (Lippmann et al., 1987).

more stress conditions. In order to achieve the A ranking, a feature must be able to *clearly* differentiate (implies separable) more than two stress conditions for a given phoneme group. This decision is based upon analysis of the statistical distribution of the feature for each stress condition across multiple speakers and utterances for a given phoneme group. A *C* ranking indicates that a selected feature can *detect* at least one stress condition. Note that a *B* ranking is subjectively placed between the *A* and *C* rankings. The ranking "–" denotes a feature with little if any stress separability. Next, these feature rankings are employed to target subsets of features (those with *A* and *B* rankings only) for each stress condition and phoneme group.

2.3. Articulatory based features

The first classifier feature domain considered is the parameterized cross-sectional area of the speech production system. These features are considered since it is believed that physical speech production variations due to stress will be reflected in vocal tract articulator variation, and therefore should be represented in the formulation of a stress classification algorithm. Articulatory vocal tract information is estimated from the acoustic speech signal using a single portion of data which is typically 4–32 ms in duration. Previous articulatory studies have illustrated methods by which to estimate the vocal tract configuration based on the acoustic speech signal

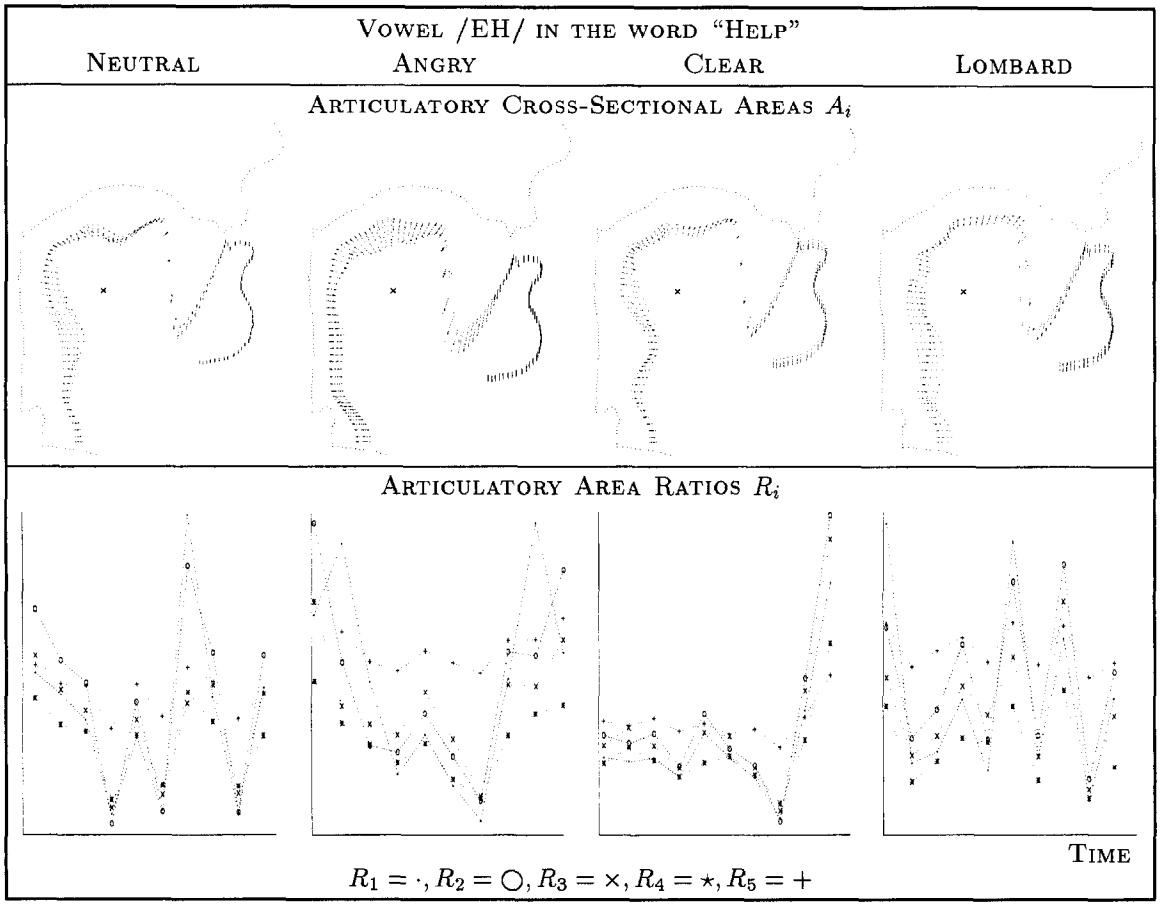

Fig. 2. Vocal tract structure variation for *Neutral*, *Angry*, *Clear*, *Lombard*.

(Kobayashi et al., 1991). In another study, the Distinctive Regions Model (DRM) was proposed for calculation of vocal tract shape from the acoustic speech signal (Mrayati et al., 1988). This method divides the vocal tract into eight regions, and imposes continuity constraints for adjacent acoustic sections (Richards et al., 1995). In a manner employed for the DRM, it is assumed that a restricted push/pull relationship exists between acoustic sections in the vocal tract (e.g., if the tongue moves forward and up, it cannot also move backward and down).

In order to illustrate vocal tract variation of speech produced under stress, cross-sectional vocal tract profiles for three stress conditions (*Angry, Clear, Lombard*) and *Neutral* are shown in Fig. 2 for a single speaker producing the selected vowel /EH/ in the word "help" [4]. The first row of this figure shows an estimate of the vocal tract shape, calculated from the linear predictive cepstral information for each frame in the selected phoneme (Hansen and Womack, 1996). It is clear that for *Angry* versus *Neutral* speech, the regions where the greatest variation occurs are reversed (i.e., pharynx cavity versus the mid pharynx to oral cavity). Differences in vocal tract shape are also apparent for *Clear* and *Lombard* effect profiles. Hence, features based upon this vocal tract shape representation should be useful for differentiating these stress conditions.

These observations motivate features that reflect cross-sectional area, A_i, of the vocal tract at selected "slices". Each slice, i, of the vocal tract is determined by a sequence of radial lines originating below the lips and across from the vocal chords (see Fig. 3). This partitioning is similar to the Distinctive Regions Model (DRM), except that ten regions of equal longitudinal size are used here.

2.3.1. Articulatory cross-sectional areas A_i

Cross-sectional areas, A_i, measure the distance from the soft to the hard pallate as illustrated in Fig. 3. The variation across phoneme groups are considered for ten slices of the vocal tract as approximated

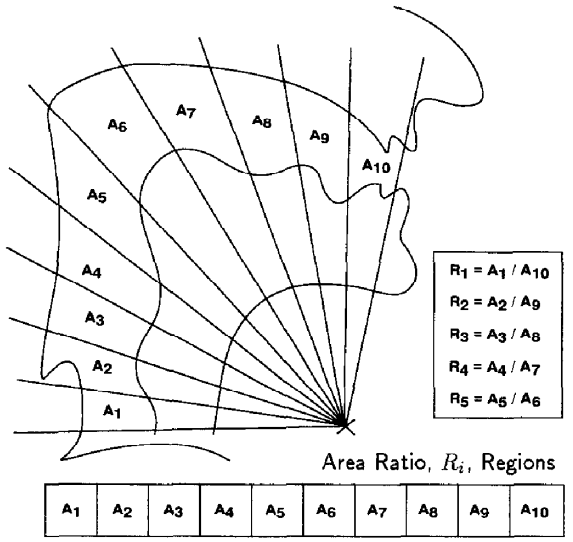

Fig. 3. Vocal tract cross-sectional area regions from the DRM model.

in the DRM (A_i: $i = 1,\ldots,10$). Assessment of cross-sectional areas indicate that articulatory parameters taken towards the end of a partition (e.g., the second half of a phoneme) are significantly more discriminative for detection of stress than those at the beginning. It is therefore suggested that some stress conditions have a greater effect on the ultimate phoneme target, rather than in the movement of the articulators toward that target. Five articulatory cross-sectional area terms are estimated for each phone class and stress condition. Feature differentiability rankings are then compiled for the articulatory cross-sectional areas and summarized in Table 1 for each phoneme and stress condition. Each cell of this table details the separability ranking for good (A rankings) versus moderate to poor (B or C rankings) detection of stress. From Table 1, we conclude that the cross-sectional area ratios of vowels, affricates, nasals and semi-vowels are the best at stress discrimination for virtually every stress condition.

2.3.2. Articulatory area ratios R_i

The articulatory cross-sectional area ratios are formulated using the DRM framework. Ten regions span the entire vocal tract from the glottis to the lips (see Fig. 3). Complementary area ratios are obtained using mean region cross-sectional areas. Each ratio,

[4] Example SUSAS audio files for a male speaker producing the word "help" under the four stress conditions from Fig. 2 is available at http://www.elsevier.nl/locate/specom.

Table 1
Articulatory targeted feature rankings

Stress classification feature targeting rankings; articulatory cross-sectional area A_i								
Stress group	Separability ranking (A,B ±)							
	FR	VL	AF	NA	SV	ST	DT	OVERALL
Angry G_1	–	A +	A +	A +	A +	–	–	A +
Normal G_2	–	A +	A +	A	–	–	B	A
Fast G_3	–	A +	A +	A +	A +	–	–	A +
Question G_4	–	A +	A +	A	A +	–	B	A +
Slow G_5	–	A +	A +	A +	A +	A +	–	A +
Clear G_6	–	A +	A +	A	A +	–	–	A +
Lombard G_7	B	A +	A +	A +	A +	–	–	A +
Soft G_8	–	A +	A +	A +	A +	A +	–	A +
Apache G_9	B	A +	A +	A +	–	–	–	A
Loud G_{10}	–	A +	A +	A +	A +	–	–	A +

R_i, is based on a mean area from one of the first five regions to one of the corresponding last five as follows:

$$R_i = \frac{A_i}{A_{11-i}} \quad \text{for } i = 1,\ldots,5. \tag{1}$$

Application of the area ratio, R_i, in evaluating stressed speech will be considered using contour plots for selected phonemes. Contour profiles are used to represent the relative area changes in regions of the vocal tract (summarized in the second row of Fig. 2). For a given frame of speech, each vocal tract configuration is estimated using sixty equally spaced cross-sectional area slices which are subsequently grouped into ten regions. The variation of each area ratio over time is modeled by obtaining a ratio average on a per phoneme basis for ten equal time periods during isolated word production. Using a frame width of 4 ms and frame separation rate of 4 ms, the average area ratio is obtained. Since these areas are estimated from the speech signal, they are only estimates of how the true vocal tract would actually behave under stress. Other methods involving imaging techniques (MRI, X-ray, etc.) would be needed to obtain actual vocal tract configurations. The present method is consistently applied to speech from all stress conditions. Therefore, any algorithm weaknesses would have an equal impact on the resulting estimated vocal tract shapes under stress (e.g., note the particular sharp tongue shape present in all stress conditions in Fig. 2).

The second row of Fig. 2 illustrates the variation of R_i over the /EH/ vowel in the word "help". It is noted that for *Neutral* and *Clear* speech, a bimodal ratio characteristic results, whereas for *Angry* and *Lombard* effect speech, a nearly unimodal characteristic variation is observed. The shape of the articulatory area ratio contour is the key factor in evaluation of movement and area distribution in the vocal tract. For example, a region where the contour slope is flat indicates no shift in vocal tract areas (i.e., stationary articulators). However, a negative slope indicates that either the area in the front of the vocal tract is becoming smaller with time, or that the back area of the vocal tract is becoming larger. The reverse is true for a contour with a positive slope. Hence, it is possible to make overall statements about the time evolution of movement for each stress condition. Note that for *Angry*, the largest shifts in area are where the contour slopes are greatest at the beginning and end of the liquid /L/. This suggests that, at the beginning of the liquid, the tongue is moving farther from the hard pallate and then, at the end of the liquid, back to its starting position.

At this point, it is useful to compare both rows of Fig. 2 since they represent the same vowel variation for the word "help". For example, the *Neutral* utterance suggests a greater area movement towards the back of the vocal tract which represents greater shifts of R_2 and R_3. Furthermore, since little movement exists at the back of the tongue, R_5 should have a relatively flat area ratio contour. Both of these

observations are confirmed in Fig. 2. However, for the *Angry* utterance, this situation is reversed and, in addition, there is greater movement towards the front of the vocal tract.

Diphthongs are known to consist of vocal tract movement from one vowel target to another, requiring a carefully orchestrated series of articulatory muscle changes. Analysis of ratio contour of the /AW/ phoneme in the word "out" showed that for *Clear* speech conditions, the speaker does not produce a significant vocal tract shift across the diphthong. *Angry* and *Lombard* effect speech are also relatively constant compared to *Neutral* which has higher ratio shifts. While vowel and diphthong area ratios reflect vocal tract variation for voiced speech, stress could also impact production of consonants such as fricatives and affricates. For example, the affricate /CH/ in the word "change" showed a large bimodal contour shape for *Angry* with large starting and ending ratio variation; which confirms a large and rapid shift in vocal tract areas. All of the stress conditions show distinctly different contours for this speaker.

While results for three phonemes are discussed here, it should be noted that several hundred ratio profiles were considered. From these profiles, it was observed that the position within a phoneme directly affects stress class discrimination. In general, we conclude that articulatory features should be useful for stress classification.

2.4. Excitation based features

Since articulatory features reflect only vocal tract information, it is therefore appropriate to consider excitation characteristics. Three excitation related features are analyzed for the application of stress classification. Previous studies have assessed variation due to stress for speech features which include pitch, duration, intensity (Hansen, 1995b) and pitch synchronous analysis of the Teager nonlinear energy operator (Cairns and Hansen, 1994). This study employs both pitch and duration for stress classification using the observations outlined below.

2.4.1. Pitch

Previous studies suggest that pitch is one of the most visible features affected by stress. We recall that pitch differs from fundamental frequency in that it is a perceived value and not the actual rate of vocal fold movement. These studies are actually based on fundamental frequency measures. An analysis of statistical variation of mean pitch across stress conditions yields the following conclusions for application to stress classification.

- Pitch characteristics are useful for classification of *Apache, Clear, Lombard, Question, Slow* and *Soft* spoken speech.
- Mean pitch for voiced speech such as diphthongs (DT), nasals (NA) and vowels (VL) are good for classifying those stress conditions under consideration.

2.4.2. Phone class duration

While duration is not a direct excitation characteristic, it indirectly affects intensity and pitch due to speech rate and available forced vital capacity of the lungs. Evaluation of the duration distribution as represented by the number of frames per phoneme was conducted with the following observations:

- Phone class duration is best for classification of *Slow, Soft* and *Question* speech. It is also good for detection of *Angry, Fast* and *Loud* speech. It is not, however, useful for classification of *Clear* speech.
- Semi-vowel (SV) duration is extraordinarily useful.
- Duration for all phoneme groups with the exception of stops (ST) are good for classifying at least three or more stress conditions.

2.4.3. Intensity

The variation of intensity across whole words and individual phoneme classes was considered in a previous study (Hansen, 1995b). One key observation from that study was that intensity varies significantly for *Angry* and *Loud* speech, especially over vowels and voiced sections. In addition, is was shown that energy shifts from consonants toward vowels for *Angry, Lombard* effect and *Loud* speech.

2.5. Cepstral based features

Cepstral based features have been used extensively in speech recognition applications because they have been shown to outperform linear predic-

tive coefficients. Cepstral based features attempt to incorporate the nonlinear filtering characteristics of the human auditory system in the measurement of spectral band energies. The five feature sets under consideration here include Mel C_i (C-Mel), delta Mel DC_i (DC-Mel), delta-delta Mel $\mathrm{D2C}_i$ (D2C-Mel), auto-correlation Mel AC_i (AC-Mel) and cross-correlation Mel $\mathrm{XC}_{i,j}$ (XC-Mel) cepstral parameters. The first three cepstral features (C_i, DC_i and $\mathrm{D2C}_i$) have been shown to improve speech recognition performance in the presence of noise and *Lombard* effect (Hanson and Applebaum, 1990). Stress equalization using cepstral parameters has also resulted in significant recognition improvement for noisy *Lombard* speech (Hansen, 1994). The AC_i and $\mathrm{XC}_{i,j}$ features are new in that they provide a measure of the correlation between Mel-cepstral coefficients. Eqs. (2)–(6) summarize how these features are calculated for each frame k assuming l correlation lags, L frames per correlation window, and M Mel frequency warped (Mel(f)) bands with energy m_j.

$$C_i(k) = \sqrt{\frac{2}{M}} \sum_{j=1}^{M} m_j \cos\left[\frac{\pi i(j-0.5)}{M}\right], \tag{2}$$

$$\mathrm{Mel}(f) = 2595 \log_{10}[1 + f/700],$$

$$\mathrm{DC}_i(k) = \frac{\sum_{w=1}^{3} w\left[C_i(k+w) - C_i(k-w)\right]}{2\sum_{w=1}^{3} w^2}, \tag{3}$$

$$\mathrm{D2C}_i(k) = \frac{\sum_{w=1}^{3} w\left[\mathrm{DC}_i(k+w) - \mathrm{DC}_i(k-w)\right]}{2\sum_{w=1}^{3} w^2}, \tag{4}$$

$$\mathrm{AC}_i^{(l)}(k) = \sum_{m=k}^{k+L} \frac{\left[C_i(m)C_i(m+l)\right]}{\sup_k \mathrm{AC}_i^{(l)}(k)}, \tag{5}$$

$$\mathrm{XC}_{i,j}^{(l)}(k) = \sum_{m=k}^{k+L} \frac{\left[C_i(m)C_j(m+l)\right]}{\sup_k \mathrm{XC}_{i,j}^{(l)}(k)}. \tag{6}$$

Next, the statistical distribution for each feature set is calculated across stress conditions in order to obtain an overall measure of the differentiating capability of pairwise features (Hansen and Womack, 1996). This measure, denoted $d_2(x_a^0, x_b^0)$, estimates the distance between two feature vector indices a and b as

$$d_2(x_a^0, x_b^0) = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}\left[(\mu_{(a,i)} - \mu_{(b,i)})^2 + (\mu_{(a,j)} - \mu_{(b,j)})^2\right]}{\sum_{i=1}^{N}(\sigma_{(a,i)} + \sigma_{(b,i)})}. \tag{7}$$

This measure assesses the N-dimensional "distance" between all N stress classes under consideration. Here, i and j range over the N stress classes where x_a^0 and x_b^0 represent the feature cluster centers. The mean and standard deviation of the ith stress condition for speech features a are denoted $\mu_{(a,i)}$ and $\sigma_{(a,i)}$, respectively. It is important to note that the mean of a feature set is not necessarily the same as the cluster center, because the cluster center is chosen by the classification algorithm such that the separation between classes is maximized. The limitation of the d_2 distance measure is that it only summarizes the separation between a pair of features across the N stress conditions considered. In order to characterize the stress differentiating capabililty of a P dimensional feature set, the following measure is formulated:

$$d_3(x_*^0) = \frac{1}{P}\sum_{a=1}^{P}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\left[(\mu_{(a,i)} - \mu_{(a,j)})^2\right]}{\left[\sigma_{(a,i)} + \sigma_{(a,j)}\right]}. \tag{8}$$

Here, a small d_3 measure suggests reduced parameter diffusion across stress classes; while large measures suggest better separation between stress classes. The values of d_3 included in this study were calculated using seven ($P = 7$) cepstral parameters per feature set (i.e., $C_1, \ldots, C_7$). With this measure, a rank ordering of feature performance for stress classification is possible. In Table 2, we summarize the twenty most separable (note $d_3 \in [1.16, 4.38]$) and least separable (note $d_3 \in [0.18, 0.41]$) features. To explain this table, we use d_3 measure assessment of three feature subsets for pitch. Note that pitch is in the best feature set four times in this table. For example, $d_3(x_*^0) = 2.02, 3.46, 0.39$ for the (i) sampled, (ii) mean and (iii) slope pitch feature sets, respectively. First, the sampled partition feature set

Table 2
Selected best and worst stress classification features

Selected stress classification features; d_3 measure $\in$ [0.18, 4.38]; 6 speakers and 9 stress conditions; $C_1, \ldots, C_7$; $AC_1, \ldots, AC_7$; *Pitch*	
Top 20	Features
Best $d_3 \in [1.16, 4.38]$	Mean: C_1, C_2, C_4, C_5, AC_1, *Pitch* Slope: C_1, AC_2 $Sample_{25\%}$: C_1, C_4, *Pitch* $Samples_{50\%}$: C_1, C_4, *Pitch* $Sample_{75\%}$: C_1, C_4, AC_2, *Pitch* Duration
Worst $d_3 \in [0.18, 0.41]$	Mean: AC_5 Slope: $C_2, \ldots, C_7$, AC_1, AC_5, AC_6, AC_7, *Pitch* $Sample_{25\%}$: AC_6, AC_7 $Sample_{50\%}$: AC_6 $Sample_{75\%}$: C_7, AC_3, AC_5, AC_6

is composed of P features taken at equally spaced samples in a given phoneme (i.e., at 25%, 50% and 75% relative positions). Second, the mean partition feature set is simply the mean of each feature across every frame in the phoneme. Finally, the slope partition feature set is discussed later in this section. Next, a comparative assessment for each feature set is presented.

2.5.1. C-Mel

The Mel-cepstral parameters C_i represent the spectral variations of the acoustic speech signal; hence, they should be useful for stress classification since vocal tract structure variation due to stress can cause movement in energy between spectral bands. A stress separability evaluation of Mel-cepstral parameters was performed for each feature and stress condition across selected phonemes. To illustrate each feature's ability to distinguish stress classes, the pairwise discriminitive measure in Eq. (2) was employed. For the purposes of multiple feature comparison, the objective stress distance measure value for C_i of $d_2(C_3^0, C_6^0) = 6.96$ and $d_3(x_*^0) = 1.12$, 1.90, 0.44 (sampled, mean and slope C-Mel, respectively) are used to compare the overall stress discrimination of this feature. Here, a larger score represents features which provide a wider separation under stressed speaking conditions.

2.5.2. DC-Mel and D2C-Mel

The delta Mel-cepstral DC_i and delta-delta Mel-cepstral $D2C_i$ parameters provide a measure of the "velocity" and "acceleration" of movement of the Mel-cepstral parameters C_i. These features are calculated using the regression in Eq. (3) on the C_i parameters. Previous studies have employed these velocity and acceleration parameters for recognition of *Lombard* effect speech (Hanson and Applebaum, 1990). It is suggested that the reason they are robust to stress variation is due to their reduced variance across stress conditions. This trait suggests that while these features are more useful for recognition, they are less applicable to stress classification. This is supported by the objective stress class separability distance measure values for DC_i and $D2C_i$ of $d_2(DC_3^0, DC_6^0) = 1.42$ and $d_2(D2C_3^0, D2C_6^0) = 1.69$ which are lower than for the Melcepstral parameters.

2.5.3. XC-Mel and AC-Mel

The cross-correlation of the Mel-cepstral parameters $XC_{i,j}$ provide a measure of the relative changes of broad versus fine spectral structure in energy bands from one Mel-cepstral parameter C_i to another C_j. Since the correlation window length ($L = 7$) and correlation lags ($l = 1$) are fixed in this study, the correlation terms are a measure of how correlated adjacent frames are over a 72 ms analysis window (24 ms/frame and 8 ms skip rate). This feature is potentially useful for stress classification, because it provides a quantitative correlation measure between broad versus fine speech spectral changes. Since this feature requires a sequence pair of Mel-cepstral parameters, an objective stress class separability distance measure could not be calculated because direct

comparison with other parameters (i.e., C-Mel, AC-Mel, etc.) would not be appropriate. However, the AC-Mel features are shown to have similar properties to XC-Mel features (Hansen and Womack, 1996). The auto-correlation of the Mel-cepstral parameters AC_i (i.e., AC-Mel) provide a measure of correlation and relative change in spectral band energies over an extended window frame. A separability feature assessment was conducted for AC-Mel resulting in a stress class separability distance measure of $d_2(AC_3^0, AC_6^0) = 7.24$, which is greater than all other cepstral based features studied. However, d_3 was slightly lower with a values of $d_3(x_*^0) = 0.61$, 0.94, 0.42 (sampled, mean and slope AC-Mel, respectively). In a previous study, the broader detail captured by the AC-Mel parameters was shown to be more reliable for stress classification (Hansen and Womack, 1996). Next, an assessment of the auto-correlation Mel-cepstral parameters and their derived features (mean, standard deviation and slope) are summarized with respect to stress classification.

- AC-Mel parameters estimated in the beginning of the phoneme group were significantly more useful than those estimated at the end of the phone group partition.
- Affricates (AF) are excellent for detection of all stress conditions considered with the exception of *Question* and *Clear* speech.
- Fricatives (FR) are good for detection of *Lombard* and *Apache* speech.

These observations are based upon an analysis of 6,580 words (35 word vocabulary, 2 tokens per word, 11 speakers, 10 stress conditions), with further analysis performed across phoneme partitions for mean, standard deviation, and slope.

2.5.4. Mean AC-Mel (MAC-Mel)

This feature provides the mean of the AC_i values across every frame in a partition. It therefore represents an average measure of the spectral structure in a phone group partition. The overall separability measure for this feature set is $d_3(MAC_*^0) = 0.94$, which is greater than $d_3(AC_*^0) = 0.61$. Mean AC-Mel parameters from:

- Semi-vowels (SV) are good for detection of *Lombard* and *Apache* speech.
- Diphthongs (DT) are good for detection of *Angry, Loud* and *Question* speech.
- Affricates (AF) are good for detection of *Neutral* speech.
- Fricatives (FR), nasals (NA), stops (ST) and vowels (VL) are not good for detection of stress.

2.5.5. Standard deviation AC-Mel (SDAC-Mel)

This feature provides the standard deviation of the AC_i values across every frame in a partition. The standard deviation of AC-Mel parameters from:

- Vowels (VL) are good for detection of *Apache, Clear* and *Lombard* effect speech.
- Fricatives (FR) are very good for detection of *Clear* speech.
- Diphthongs (DT) are good for detection of *Clear* and *Neutral* speech.

Table 3
Slope AC-Mel targeted feature rankings

Stress classification feature targeting rankings; slope AC-Mel SAC_i								
Stress group	Separability ranking (A,B ±)							
	FR	VL	AF	NA	SV	ST	DT	Overall
Angry G_1	–	A +	–	–	–	A	A +	A –
Normal G_2	B	A +	B +	A +	–	–	–	A
Fast G_3	–	A +	–	A +	–	A	A +	A
Question G_4	–	A +	B +	A	–	–	A +	A–
Slow G_5	–	A +	–	A +	–	–	A +	A
Clear G_6	B	A +	B +	A	–	A	–	A +
Lombard G_7	–	A +	–	A +	–	A	A +	A +
Soft G_8	–	A +	–	A +	–	–	A +	A
Apache G_9	–	A +	–	A +	–	A	A +	A
Loud G_{10}	–	A +	–	–	–	A	A +	A

Table 4
Overall targeted feature rankings

Stress classification feature targeting rankings; cepstral, excitation and articulatory domains							
Stress parameter	Separability ranking totals						
	A +	A	A	B +	B	B −	−
Articulatory	37	3	0	0	4	0	26
Pitch	30	0	0	0	0	1	39
Duration	8	0	18	3	2	0	39
AC-Mel	7	5	0	10	3	2	43
Mean AC-Mel	0	0	0	1	5	1	63
Std AC-Mel	1	0	0	0	5	2	62
Slope AC-Mel	26	6	0	3	2	0	33

- Affricates (AF) are good for detection of *Angry* and *Loud* speech.
- Nasals (NA), stops (ST) and semi-vowels (SV) are not good for detection of stress.

2.5.6. Slope AC-Mel (SAC-Mel)

This feature is based on the slope from the leftmost min/max AC-Mel parameter to the rightmost min/max AC-Mel parameter in the AC_i sequence for a phone group partition. It therefore provides an overall measure of the spectral movement across a partition. This feature can be compared to others using the overall separability measure value of $d_3(SAC_*^0) = 0.42$ which is slightly less than the slope C-Mel feature $d_3(SC_*^0) = 0.44$. An evaluation of this feature across the SUSAS database was performed to assess its stress discriminating ability. The results shown in Table 3 suggest that the slope of AC-Mel for vowels are consistently useful for differentiating all stress conditions. The slope AC-Mel feature for diphthongs, nasals and stops are also useful for stress differentiation whereas fricatives and affricates may be somewhat useful for stress detection.

2.6. Targeted stress classification features

In the previous sections, features from articulatory, excitation and cepstral domains were considered for their ability to achieve reliable stress classification. In the formulation of a neural network based stress classification algorithm, a codebook of targeted features will be assembled for each potential stress condition and phoneme class group. The targeted feature evaluation results in a parent set of features from these three domains. Table 4 summarizes the targeted feature rankings by listing the total number of times each rank appears for each feature set (i.e., the aggregate of Tables 1 and 3). From the articulatory feature domain, the cross-sectional vocal tract areas A_i, are selected for use in the parent feature set. For the excitation feature domain, pitch and duration are selected. Finally, from the cepstral domain, auto-correlation Mel-cepstral features and their statistics (mean, standard deviation and slope) are included in the parent feature set. 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 features is employed in the formulation of the stress classification algorithm.

3. Stress classification algorithm

Next, a stress classification algorithm is formulated using back propagation neural networks and targeted stress sensitive speech features. The stress classification system, as illustrated in Fig. 4, has three major components: (i) stress sensitive feature extraction, (ii) automatic stress independent phone group partitioning, and (iii) neural network stress scoring. Each area will be considered in detail.

3.1. Stress independent partitioning

A speech partitioning algorithm that provides consistently parsed speech across time is a difficult task

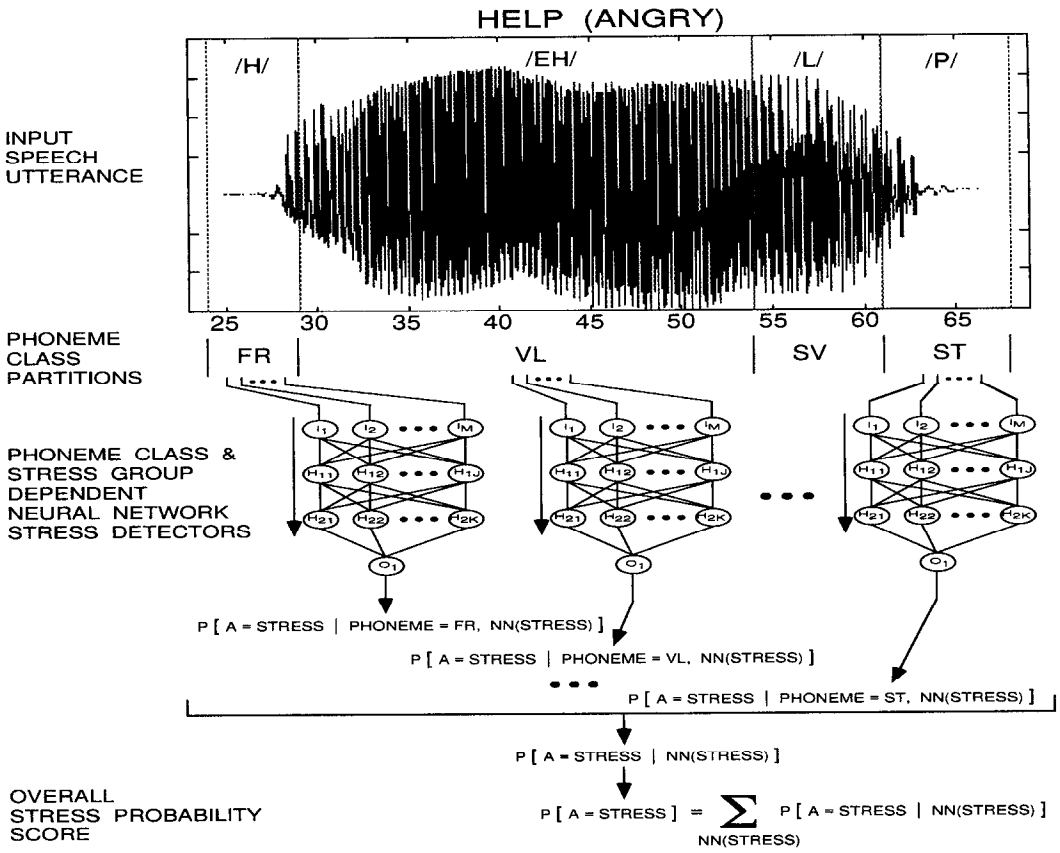

Fig. 4. Stress classification algorithm.

due to nonunique transitions between phonemes, the impact of stress, and coarticulation effects (Arslan and Hansen, 1994). However, in a previous study on robust speech partitioning (Pellom and Hansen, 1996), an algorithm was formulated using hidden Markov models and Viterbi decoding to parse speech by phoneme group. Though this algorithm was used to direct constrained speech enhancement, it was also shown to be useful for speech partitioning under stress. The speech partioning based HMM models for this study were trained using neutral speech data from the TIMIT (Fisher et al., 1986) and stressed speech SUSAS databases. Each HMM is trained for one phoneme group using continuous density distributions with five states per phoneme and two mixtures. The seven models (SI: Silence, FR: Fricatives, VL: Vowels, AF: Affricates, NA: Nasals, SV: Semi-Vowels, DT: Diphthongs) were trained using word grammars composed of phoneme group sequences. Viterbi decoding is then used to match the state sequence to the grammar for each input word to estimate the phoneme boundary sequence. This portion is incorporated in the overall stress classification algorithm as illustrated in Fig. 4.

3.2. Stress classifier formulation

In formulating an algorithm for stress classification, it should be noted that a range of stress or emotion may exist for a given speaking condition. Hence, it is necessary to estimate a stress probability response vector to assess the different degrees and types of stress. A stress score is estimated by training a stress detector to recognize one stress condition given knowledge of the phoneme group determined from the partitioning task. A codebook of these

stress detectors can then be used to provide an estimate of each stress condition. This formulation is based upon a mathematical framework that represents feature movement from one region in a source generator space to another, where each speech production region is represented as a stress state (Hansen and Cairns, 1995). Next, the general stress detection system shown in Fig. 4 employs neural networks to estimate the stress score $p(\xi_k \mid w_i)$; which measures the degree of stress given that utterance w_i is spoken under stress condition k. Two particular approaches using this general framework for stress classification are presented: (i) mono-partition (MPSC) and (ii) triple-partition (TPSC).

Two types of neural networks are considered for single and triple-partition classification. Mono-partition classification uses the cascade correlation network (Minai and Williams, 1990) with an extended delta-bar-delta (EDBD) learning rule. Triple-partition classification employs the commonly used fast back-propagation learning rule (Hansen and Womack, 1996). The motivation for using a more complex neural network training algorithm (EDBD) for single-partition classification is that training data for each class is less separable and larger than for the triple-partition case. Details on how these neural network classifiers were implemented will be presented in Section 4.

Both stress classification algorithms include three types of features: single frame, partition and word based parameters. The MPSC and TPSC algorithms differ in several ways, but most notably in the speech features that drive the algorithms. In the MPSC system, a stress detector is formulated for each stress condition and across all phoneme groups; however, the feature sets are not targeted. For the TPSC system, a stress detector is formulated for each stress condition and phoneme group using targeted features. Sections 4.1 and 4.2 will present results on the performance of these two approaches for stress classification. Next, the TPSC system will be employed in the formulation of a stress directed speech processing system.

3.3. Stress directed speech recognizer formulation

Here, the application of stress classification is considered in an effort to show that knowledge of stress could provide improvement in overall recognition performance. A flow diagram is shown in Fig. 5, where separate stress dependent recognizers are employed in combination with a stress classification system. Hence, it is proposed that a speech recognizer trained for one stress class will better model differences between words, since it is not required to model the additional variations due to stressed speaking conditions.

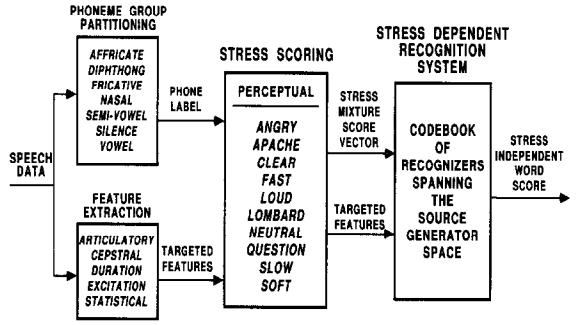

Fig. 5. Stress directed recognition algorithm.

Next, the notation associated with the stress directed recognition algorithm is presented. First, the stress classification system outputs a K-dimensional vector of stress scores, denoted $\vec{\xi} = \{\xi_k \mid k = 1, \ldots, K\}$, since there are K stress conditions. Second, since there are I words in the vocabulary, there is an $i \times k$ dimensional matrix of possible stress score vectors ξ_k in each column, such that each matrix term is denoted $w_{ik} = p(\xi_k \mid w_i)$. Next, the probability that the stress condition is k, denoted $p(\xi_k)$, is calculated using the matrix weight term w_{ik}. Fourth, a word recognizer score, denoted $p(w_i \mid \xi_k)$, is obtained for each word w_i in the vocabulary given that the stress condition is k. Finally, the highest probability that the word is $imax$, denoted $p(w_{imax})$, is calculated using these probabilities with the following procedure.

In order to formulate a codebook of stress dependent recognizers, it is desirable to use the existing HMM speech recognition framework as shown in Fig. 5. 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 (Hansen et al., 1994). 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 largest stress probability term $p(\xi_{kmax} \mid w_i)$ has been calculated as in Eq. (9), the speech features are passed to the stress dependent recognition system trained for stress condition *kmax* as illustrated in Fig. 5. The final utterance decision is then calculated as follows by maximizing $p(w_i \mid \xi_{kmax})$ over every word in the vocabulary:

$$p(\xi_{kmax} \mid w_i) = \max_k p(\xi_k \mid w_i), \tag{9}$$

$$p(w_{imax}) = \max_i p(w_i \mid \xi_{kmax}). \tag{10}$$

Next, the four preceeding components are integrated as shown in Fig. 5, for the overall stress independent recognition system. To achieve reliable performance, phonemic class partitioning is still 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 (MPSC) versus context dependent (TPSC) stress classification using the partitioned phoneme sequence.

4. Evaluations

4.1. Mono-partition stress classification (MPSC)

The MPSC system is formulated using three key factors: (i) a single-partition data window, (ii) perceptually grouped stress conditions, and (iii) a common feature set. Hence, the stress classifier is provided with only one partition of data for the stress class decision. Furthermore, in an attempt to minimize the number of classifier models, stress classes which were found to be perceptually similar are grouped together. Each stress group is denoted as G_i where index i indicates the group number. Note that this grouping is based upon informal listening tests of perceptually similar stressed conditions. Stress classes are grouped as follows: G_1 (*Angry, Loud*); G_2 (*Cond50, Cond70, Neutral, Soft*); G_3 (*Fast*); G_4 (*Question*); G_5 (*Slow*); G_6 (*Clear*); and G_7 (*Lombard effect*). Finally, a common set of speech features is used for stress classification of all phoneme groups.

Performance over a focused word set is one means of measuring a stress classification algorithm's ability to differentiate stressed speech. In this evaluation, a five word vocabulary from one speaker taken from the SUSAS database is used. The five word set chosen is: "brake", "east", "freeze", "help" and "steer". Note that for mono-partition based stress classification, the order of the phoneme classes will not influence stress classification performance. Therefore, the same three neural network stress detectors will be used for each phoneme in the two sample words "sam" and "mass". Comparative overall classification results for the feature sets across all stress groups are (C_i, DC_i, $\mathrm{D2C}_i$, AC_i) = (78.9%, 76.9%, 79.3%, 80.6%), suggesting that AC-Mel parameters are the best cepstral features for stress classification considered. Results will indicate for d_2 that both C_i and AC_i perform better than velocity and acceleration features. For example, for *Lombard* effect speech, a parameterization using AC_i provides better stress classification results than C_i, DC_i or $\mathrm{D2C}_i$ (i.e., 94% versus 76%, 71%, 58%).

Another measure for feature set comparison is the stress class separability distance measure from Eq. (7). The measure assesses the separation of stress groups for a given feature pair. The two chosen feature indices are $a = 3$ and $b = 6$ so that the distance measure $d_2(x_a^0, x_b^0)$ yields (C_i, DC_i, $\mathrm{D2C}_i$, AC_i) = (6.96, 1.42, 1.69, 7.24). It is clear that for index 3 versus 6 (roughly a comparison of global versus fine spectral structure for C_i), that C_i and AC_i are better able to reflect differences in stressed speech. These values are designed for comparison purposes only, hence, actual values do not have physical units. The results show that the AC_i features are the most separable feature set of those considered. Hence, d_2 provides a means by which to reduce the number of features in the original codebook set for stress classification.

For the MPSC system evaluation, it was determined that (i) perceptually grouped stress conditions may not translate to similarly produced stressed styles, (ii) a broad feature set is needed (such as articulatory and excitation), (iii) separate classifiers

should be employed for each phoneme group, and (iv) adjacent partition information should be incorporated to model cross-partition variation.

4.2. Tri-partition stress classification (TPSC)

Reduction of the size of the feature targeting search space is accomplished by using only the AC-Mel cepstral features. For the second classifier, stress classes are grouped as follows: G_1 (*Angry*); G_2 (*Neutral*); G_3 (*Fast*); G_4 (*Question*); G_5 (*Slow*); G_6 (*Clear*); G_7 (*Lombard* effect); G_8 (*Soft*); G_9 (*Apache*); and G_{10} (*Loud*). Note that the additional stress class termed *Apache* is added which represents actual helicopter cockpit stressed speech for comparison with other simulated stressed speech conditions.

MPSC and TPSC classifier results are compared with the following features made available to both classifiers: autocorrelation Mel-cepstral parameters and their derived features, durational, articulatory and excitation. In order to reduce data requirements for TPSC, 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.

The TPSC system consists of a codebook of neural networks, one for each phoneme group and stress condition. As Fig. 6 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 +34.3% (Womack and Hansen, 1995). Note that when only one back-propagation neural network is trained for each phoneme group, tri-partition classification using the master non-targeted feature set did not perform satisfactorily. The results also show that a phone sequence, stress and speaker independent stress detection system is not viable. This leads us to focus the problem such that the stress detectors are both stress and phoneme sequence dependent. Next, details of the improvement obtained with targeted feature sets is discussed.

Outstanding stress classification performance is achieved for vowels and diphthongs. Good performance is also achieved for nasals and stops which might be unexpected since they are more difficult to represent due to limited duration, mixed excitation, or derivation from an all-pole speech model than other phoneme groups. It is suggested that such

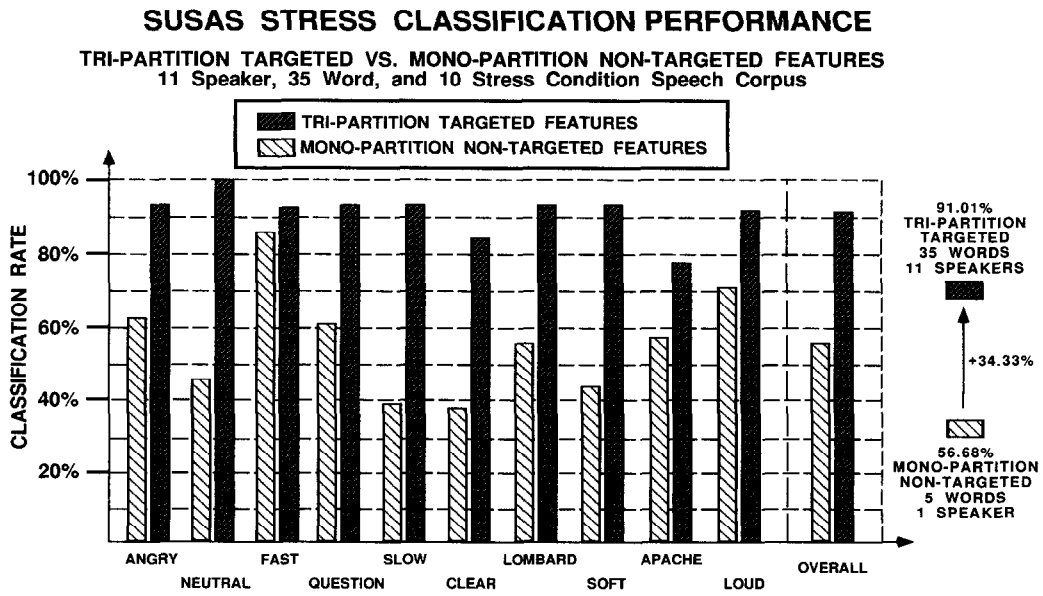

Fig. 6. Stress classification performance comparison using (i) mono-partition non-targeted and (ii) tri-partition targeted features.

performance is achieved because a mixture of excitation and articulatory features are employed in addition to adjacent partition information. A 7.2% difference between the open and closed test results suggests that the stress classification algorithm is able to generalize its decisions from testing data.

4.3. Automatic versus human stress classification

To put the performance of the triple-partition stress classification algorithm into perspective, a comparison is made with human listeners. A previous study on stressed speech synthesis employed a subjective listener test where the listener was asked to decide on a pairwise token basis the stress content (Bou-Ghazale and Hansen, 1995). In that study, an experiment was performed using SUSAS data in which human listeners were asked to select whether one, both or neither of the two tokens was spoken under stress. Here, only a single stress condition versus *Neutral* was considered. The listener's ability to detect *Angry*, *Lombard* and *Loud* versus *Neutral* speech was 97%, 82% and 85%, respectively. This contrasts with the performance of the automatic stress classifier which achieved 97%, 100% and 94%, respectively. Note that for *Lombard* effect speech, the stress classification system achieved 18% higher performance than human listeners. The potential reason *Angry* and *Loud* listener performance is closer to that of the stress classifier is that listeners may have more experience identifying these stress styles versus *Lombard* effect. The results in this study show that it is possible for an automatic stress classification system to perform as well or better than a human listener.

4.4. Application to stressed speech recognition

In this final section, we consider whether the proposed stress classification algorithm can provide additional knowledge to improve speech recognition under stressed conditions. The scores from the TPSC system are used to weight the outputs of a codebook of stress dependent recognizers. Hence, a recognizer must be formulated for each type of speaker stress. Here, a speaker dependent, isolated word, continuous density hidden Markov model recognizer is used. The HMM training method employs a state tying initialization based upon the degree of similarity between mean mixture vectors in successive states. The models assume left-to-right state transitions with no skips allowed. The training algorithm is based on

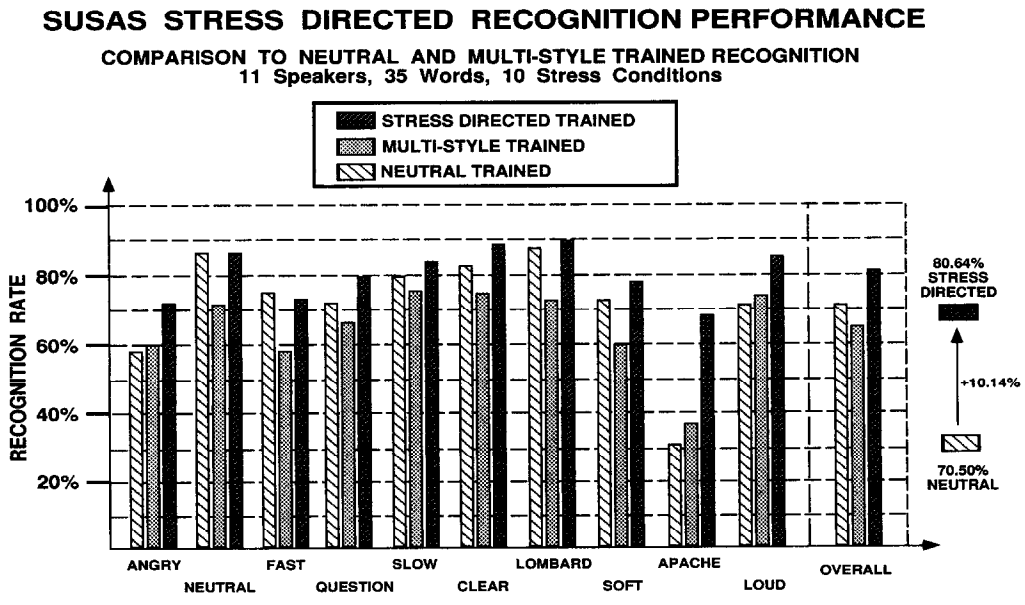

Fig. 7. Stress directed recognition comparison.

the Baum–Welch forward–backward reestimation algorithm.

Three stressed speech recognition evaluations are considered with results summarized in Fig. 7. To establish a baseline level of performance, the first evaluation employs neutral trained HMMs that are tested with stressed SUSAS data. An overall open test recognition rate of 70.5% is achieved, with performance ranging from 33% for *Apache* to 87% for *Neutral* speech. It is noted that recognition is most severely affected by *Apache* speech since the data represents actual stressed speech. The second evaluation focuses upon multi-style trained HMMs. For each word, an HMM is trained across all stress conditions and speakers in the training set. This approach differs from a previous study (Lippmann et al., 1987) in that training is speaker *independent* and speech is sampled at 8 kHz. An overall open test recognition performance of 65.2% is achieved; which is −5.3% lower than the neutral trained HMM. The third evaluation assumes estimated knowledge of the speaker stress state from a tandem TPSC neural network stress classifier and HMM recognizer trained for each stress condition. The stress directed recognition rate is 80.6%, which is +10.1% more than neutral trained and +15.4% more than the multi-style trained HMM. Results are particularly encouraging for *Apache* style stressed speech, with rates increasing from 31% to 69%. This suggests that improvement can be achieved for actual stressed speech. 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.

5. Summary

In this study, the problem of improved stress classification using targeted speech features has been considered. Two stress classification algorithms are proposed to estimate a probability vector representing the degree of speaker stress. It was shown that context sensitive stress classification via tri-partition (TPSC) achieves better performance than the mono-partition (MPSC) algorithm. Further, new features for stress classification from the articulatory and excitation domains were assessed. It is suggested that the output stress probability vector can also be employed to measure mixtures of speaker stress (e.g., combined *Fast* and *Loud* speech). A stress mixture model is suggested to be useful for applications such as emergency telephone message sorting or performance improvement in conventional speech processing systems. The stress classifier output stress score vector was then used to direct a stress dependent HMM recognizer. This resulted in an improvement of +10.1% to +15.4% over neutral and multi-style trained systems. In conclusion, stress classification using targeted features and neutral network classifiers have been shown to be viable for the estimation of the degree of speaker stress, as well as providing useful information for improving performance of a speech recognition algorithm.

References

L.M. Arslan and J.H.L. Hansen (1994), "A mimimum cost based phoneme class detector for improved iterative speech enhancement", *IEEE Proc. Internat. Conf. Acoust. Speech Signal Process.*, pp. 45–48.

S.E. Bou-Ghazale and J.H.L. Hansen (1995), "Source generator based stressed speech perturbation", *Proc. EuroSpeech*, pp. 455–458.

D.A. Cairns and J.H.L. Hansen (1994), "Nonlinear analysis and detection of speech under stressed conditions", *J. Acoust. Soc. Amer.*, Vol. 96, No. 6, pp. 3392–3400.

G.J. Clary and J.H.L. Hansen (1992), "A novel speech recognizer for keyword spotting", *Internat. Conf. on Spoken Language Processing*, pp. 13–16.

W.M. Fisher, G.R. Doddington and K.M. Goudie-Marshall (1986), "The DARPA speech recognition research database: Specifications and status", *Proc. DARPA Speech Recognition Workshop*, TIMIT database.

J.H.L. Hansen (1988), Analysis and compensation of stressed and noisy speech with application to robust automatic recognition, Ph.D. thesis, Georgia Institute of Technology, Atlanta, GA.

J.H.L. Hansen (1993), "Adaptive source generator compensation and enhancement for speech recognition in noisy stressful environments", *IEEE Proc. Internat. Conf. Acoust. Speech Signal Process.*, pp. 95–98.

J.H.L. Hansen (1994), "Morphological constrained feature enhancement with adaptive cepstral compensation (MCE-ACC) for speech recognition in noise and Lombard effect", *IEEE Trans. Speech Audio Process.*, Vol. 2, No. 4, pp. 598–614.

J.H.L. Hansen (1995a), "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition", *ESCA-NATO Proc. Speech Under Stress Workshop*, Lisbon, Portugal, pp. 91–98.

J.H.L. Hansen (1995b), "A source generator framework for analysis of acoustic correlatecs of speech under stress. Part 1:

Pitch, duration, and intensity effects", *J. Acoust. Soc. Amer.*, Submitted.

J.H.L. Hansen and S.E. Bou-Ghazale (1995), "Robust speech recognition training via duration and spectral-based stress token generation", *IEEE Trans. Speech Audio Process.*, Vol. 3, No. 5, pp. 415–421.

J.H.L. Hansen and D.A. Cairns (1995), "ICARUS: Source generator based real-time recognition of speech in noisy stressful and Lombard effect enviromnents", *Speech Communication*, Vol. 16, No. 4, pp. 391–422.

J.H.L. Hansen and M.A. Clements (1995), "Source generator equalization and enhancement of spectral properties for robust speech recognition in noise and stress", *IEEE Trans. Speech Audio Process.*, Vol. 3, No. 5, 407–415.

J.H.L. Hansen and B.D. Womack (1996), "Feature analysis and neural network based classification of speech under stress", *IEEE Trans. Speech Audio Process.*, Vol. 4, No. 4, pp. 307–313.

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

B.A. Hanson and T. Applebaum (1990), "Robust speaker-independent word recognition using Instantaneous, dynamic and acceleration features: Experiments with Lombard and noisy speech", *Internat. Conf. Acoust. Speech Signal Process.*, pp. 857–860.

J.C. Junqua (1993), "The Lombard reflex and its role on human listeners and automatic speech recognizers", *J. Acoust. Soc. Amer.*, Vol. 93, pp. 510–524.

T. Kobayashi, M. Yagyu and K. Shirai (1991), "Application of neural networks to articulatory motion estimation", *Internat. Conf. Acoust. Speech Signal Process.*, pp. 489–492.

H.S. Lee and A.C. Tsoi (1995), "Application of multi-layer perceptron in estimating speech/noise characteristics for speech recognition in noisy environment", *Speech Communication*, Vol. 17, Nos. 1–2, pp. 59–76.

P. Lieberman and S. Michaels (1962), "Some aspects of fundamental frequency and envelope amplitude as related to the emotional content of speech", *J. Acoust. Soc. Amer.*, Vol. 34, No. 7, pp. 922–927.

R.P. Lippmann, E.A. Martin and D.B. Paul (1987), "Multi-style training for robust isolated word speech recognition", *Internat. Conf. Acoust. Speech Signal Process.*, pp. 705–708.

E. Lombard (1911), "Le signe de l'élévation de la voix", *Ann. Maladies Orelle, Larynx, Nez, Pharynx*, Vol. 37, pp. 101–119.

A.A. Minai and R.D. Williams (1990), "Back-propagation heuristics: a study of the extended delta-bar-delta algorithm", *Internat. Joint Conf. on Neural Networks*, pp. 595–600.

M. Mrayati, R. Carre and B. Guerin (1988), "Distinctive regions and modes: A new theory of speech production", *Speech Communication*, Vol. 7, No. 3, pp. 257–286.

B.L. Pellom and J.H.L. Hansen (1996), "Text-directed speech enhancement using phoneme classification and feature map constrained vector quantization", *Internat. Conf. Acoust. Speech Signal Process.*

H.B. Richards, J.S. Mason, M.J. Hunt and J.S. Bridle (1995), "Deriving articulatory representations of speech", *Proc. EuroSpeech*, pp. 761–764.

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

F.K. Soong and A.E. Rosenberg (1988), "On the use of instantaneous and transitional spectral information in speaker recognition", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 36, No. 6, pp. 871–879.

B.J. Stanton, L.H. Jamieson and G.D. Allen (1989), "Robust recognition of loud and Lombard speech in the fighter cockpit environment", *Internat. Conf. Acoust. Speech Signal Process.*, pp. 675–678.

B.D. Womack and J.H.L. Hansen (1995), "Stress independent robust HMM speech recognition using neural network stress classification", *Proc. EuroSpeech*, pp. 1999–2002.

B.D. Womack and J.H.L. Hansen (1996), "Stressed speech classification with application to robust speech recognition", *Internat. Conf. Acoust. Speech Signal Process.*, pp. 53–56.

C.E. Williams and K.N. Stevens (1972), "Emotions and speech: Some acoustic correlates", *J. Acoust. Soc. Amer.*, Vol. 52, No. 4, pp. 1238–1250.