Feature Analysis and Neural Network-Based Classification of Speech Under Stress

John H. L. Hansen and Brian D. Womack

Abstract—**It is well known that the variability in speech production due to task-induced stress contributes significantly to loss in speech processing algorithm performance. If an algorithm could be formulated that detects the presence of stress in speech, then such knowledge could be used to monitor speaker state, improve the naturalness of speech coding algorithms, or increase the robustness of speech recognizers. The goal in this study is to consider several speech features as potential stress-sensitive relayers using a previously established stressed speech database (SUSAS). The following speech parameters will be considered: mel, delta-mel, delta-delta-mel, auto-correlation-mel, and cross-correlation-mel cepstral parameters. Next, an algorithm for speaker-dependent stress classification is formulated for the 11 stress conditions: *Angry*, *Clear*, *Cond50*, *Cond70*, *Fast*, *Lombard*, *Loud*, *Normal*, *Question*, *Slow*, and *Soft*. It is suggested that additional feature variations beyond neutral conditions reflect the perturbation of vocal tract articulator movement under stressed conditions. Given a robust set of features, a neural network-based classifier is formulated based on an extended delta-bar-delta learning rule. Performance is considered for the following three test scenarios: monopartition (nontargeted) and tripartition (both nontargeted and targeted) input feature vectors.**

I. Introduction

The problem of speaker stress classification is to assess the degree of a specific stress condition present in a speech utterance. "Stress," in this study, refers to perceptually induced variations on the production of speech. The variation in speech production due to stress can be substantial and will therefore have an impact on the performance of speech processing applications [4], [10]. A number of studies have focused on stressed speech analysis 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. A number of studies have considered the effects of stress on variability of speech production [1], [5], [6], [9]. One stress condition of interest is the *Lombard* effect [2], [3], [6], which results when speech is produced in the presence of background noise. 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 [5]. Extensive statistical assessment of over 200 parameters for simulated and actual speech under stress suggests that stress classification based on feature distribution separability characteristics is possible. One approach that has been suggested as a means of modeling stress for recognition is based on a source generator framework [3], [4]. In this approach, stress is modeled as perturbations along a path in a multidimensional articulatory space. Using this framework, improvement in speech recognition was demonstrated for noisy *Lombard* effect speech [3]. This approach has also been considered as a means for generating artificial stressed training tokens [4]. Finally, a tandem neural network stress classifier and HMM recognizer based on this framework has been shown to be effective for recognition under several stress conditions including the *Lombard* effect [10].

Although a number of studies have considered analysis of speech under stress, the problem of stressed speech classification has received little 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 [1]. 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 [3], which employ morphologically constrained feature enhancement, have demonstrated improved recognition performance. Next, larger data sets have been considered such as multistyle training [7] to improve performance in speaker-*dependent* systems. Additionally, an extension of multistyle training based on stress token generation has also shown improvement in stressed speech recognition [4]. However, for speaker *independent* systems, multistyle training results in performance loss over neutral trained systems [10] since it is believed that the HMM's cannot always span both the stress *and* speaker production domains. Hence, the problem of stressed speech recognition requires the incorporation of stress knowledge. This can be accomplished implicitly through robust features or front-end stress equalization. Alternately, stress knowledge could be incorporated explicitly by using a stress classifier to direct a codebook of stress-dependent recognition systems [10]. Several application areas exist for stress classification such as objective stress assessment or improved speech processing for synthesis or recognition. For example, a stress detector could direct highly emotional telephone calls to a priority operator at a metropolitan emergency service. A stress classification system could also provide meaningful information to speech systems for recognition, speaker verification, and coding.

In this study, the problem of speech feature selection for classification of speech under stress is addressed. The focus here is to develop a classification algorithm using features that have traditionally been employed for recognition. It is suggested that such a stress classifier could be used to improve the robustness of future speech recognition algorithms under stress. An analysis of vocal tract variation under stress using cross-sectional areas, acoustic tubes, and spectral features is considered. Given knowledge of these variations, five cepstral-based feature sets are employed in the formulation of a neural network-based stress classification algorithm. Next, the feature sets are analyzed using an objective separability measure to select the set that is most appropriate for stress classification. Finally, the stress classification algorithm is evaluated using an existing speech under stress database (SUSAS) for i) feature set selection and ii) monopartition stress classification algorithm performance.

II. Speech Under Stress

Speaker stress assessment is useful for applications such as emergency telephone message sorting and aircraft voice communications monitoring. Here, stress can be defined as a condition 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. With this

Manuscript received July 18, 1994; revised February 2, 1996. The associate editor coordinating the review of this paper and approving it for publication was Dr. Xuedong Huang.

The authors are with the Robust Speech Processing Laboratory, Department of Electrical Engineering, Duke University, Durham, NC 27708-0291 USA.

Publisher Item Identifier S 1063-6676(96)05073-0.

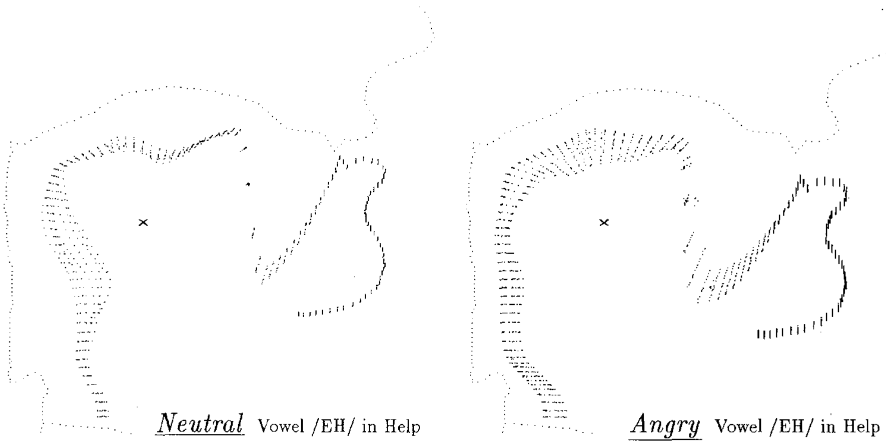


Fig. 1. *Neutral* versus *Angry* vowel /EH/ in "help" vocal tract variation.

definition, two stress effect areas emerge: perceptual and physiological. Perceptually induced stress results when a speaker perceives his environment to be different from "normal" such that speech production *intention* varies from neutral conditions. The causes of perceptually induced stress include emotion, environmental noise (i.e., *Lombard* effect), and actual task workload (pilot in an aircraft cockpit). Physiologically induced stress is the result of a physical impact on the human body that results in deviations from neutral speech production despite intentions. Causes of physiological stress can include vibration, G-force, drug interactions, sickness, and air density. In this study, the following 11 perceptually induced stress conditions from the SUSAS database are considered: *Angry, Clear, Cond50, Cond70,*[1] *Fast, Lombard, Loud, Neutral, Question, Slow,* and *Soft.*

A. SUSAS Database

The evaluations conducted in this study are based on data previously collected for analysis and algorithm formulation of speech analysis in noise and stress. This database refers to *speech under simulated and actual stress* (SUSAS) and has been employed extensively in the study of how speech production varies when speaking during stressed conditions.[2] A vocabulary set of 35 aircraft words make up over 95% of the database. These words consist of monosyllabic and multisyllabic words that 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 [3], [5].

III. Vocal Tract Modeling

Before a stress classification algorithm is formulated, it would be beneficial to illustrate how stress effects vocal tract structure. This section will demonstrate feature perturbations due to stress via

i) visualization of vocal tract shape
ii) analysis of acoustic tube cross sectional area
iii) speech parameter movements.

This analysis is based on a linear acoustic tube model with speech sampled at 8 kHz. The following sections are intended to show that a relation exists between speech production perturbation, acoustic tube analysis, and recognition feature variations.

A. Vocal Tract Shape

One means of illustrating the effects of stress on speech production is to visualize the physical vocal tract shape. The movements throughout the vocal tract can be displayed by superimposing a time sequence of estimated vocal tract shapes for a chosen phoneme. The vocal tract shape analysis algorithm assumes a known normalized area function and acoustic tube length. The articulatory model approach by Wakita [11] was used to consider changes in vocal tract shape under neutral and angry conditions, as illustrated in Fig. 1. Here, a set of vocal tract shapes are superimposed for each frame in the analysis window (10 frames for *Normal* and 18 frames for *Angry*, with 24 ms/frame). For the *Normal* condition, the greatest perturbation is in the pharynx cavity. However, for the *Angry* condition, the greatest perturbation is in the blade and dorsum of the tongue and the lips. This suggests that when a speaker is under stress, typical vocal tract movement is effected, resulting in quantifiable perturbation in articulator position.

B. Acoustic Tube Area

Next, a second experiment is performed to demonstrate that vocal tract variation due to stress results in vocal tract parameter variation. The experiment assumed fixed tube lengths in order to calculate the area coefficients for a 15-tube vocal tract model. These coefficients are calculated and logarithmically scaled for all frames of the /EH/ sound in the word "help" for *Normal* and *Angry* stressed speech. Fig. 2 shows the resulting change in acoustic tube models for cross-sectional areas. Each frame of the /EH/ phoneme is superimposed to show acoustic tube perturbations throughout the utterance. A greater perturbation in the acoustic tube area parameters for the *Angry* condition is observed. In addition, note the wide range of area perturbations across stress conditions.

C. Speech Parameter Variation Due to Stress

Finally, speech parameter variation due to stress is considered. In Fig. 3, one autocorrelation Mel AC_i (AC-Mel) speech analysis parameter is chosen to illustrate the variation due to stress. The key difference is observed by contrasting the gradual transitions across the utterance for the *Normal* compared with the *Angry* speech parameter contour. We also note the longer duration and approximately bimodal nature of the *Angry* contour.

It has been shown that speech under stress causes variation in vocal tract structure, acoustic tube models, and speech parameters across time. In general, assessment of vocal tract shape is useful for the analysis of speech under stress. We note that the vocal tract model

[1] *Cond50* and *Cond70* refer to speech spoken while performing a moderate and high workload computer response task.

[2] Approximately half of the SUSAS database consists of style data donated by Lincoln Laboratories [7].

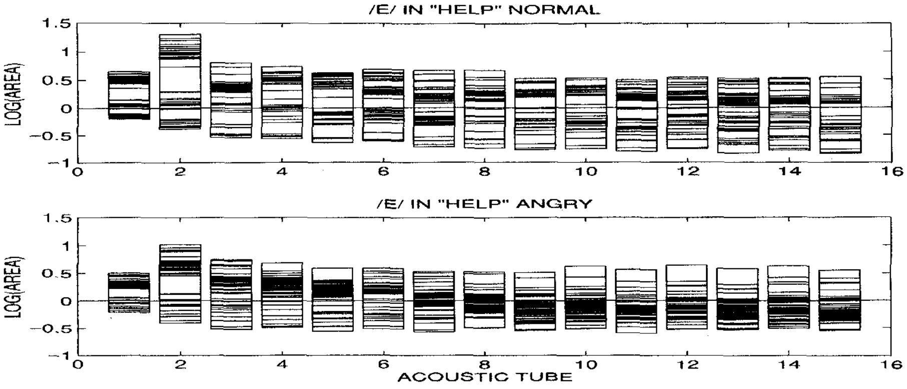


Fig. 2. Stress variation of $\log_{10}$ area coefficients.

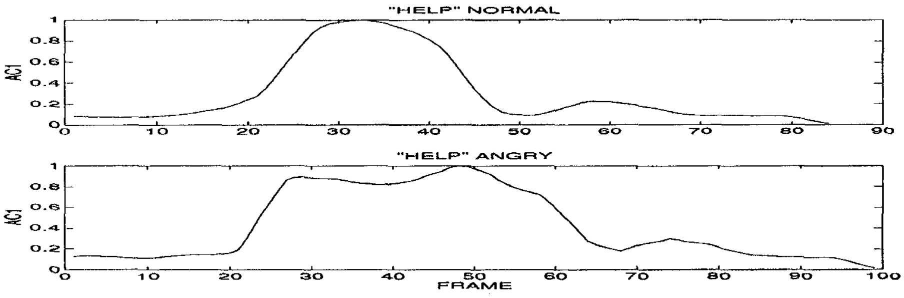


Fig. 3. Stress variation of AC_1 for *Normal* versus *Angry*.

employed in this study does not represent changes in the excitation of the vocal tract and, hence, the physical movements that control pitch. However, for the /EH/ phoneme in "help," it is known that the mean pitch of 142 Hz for *Normal* speech increases to 282 Hz for *Angry* speech. In addition, pitch is recognized as a good feature for stress analysis [1], [5]. However, in the present study, in order to limit front-end parameterization, only features typically used for recognition will be considered.

IV. Classification Features for Stressed Speech

In this section, speech production variation for cepstral features in response to perceptually induced speaker stress is considered. It is assumed that continuous speech has been parsed consistently by phoneme class across stress conditions. The primary focus is to determine which of five cepstral feature representations is better able to differentiate speaker stress.

A. Cepstral-Based Features

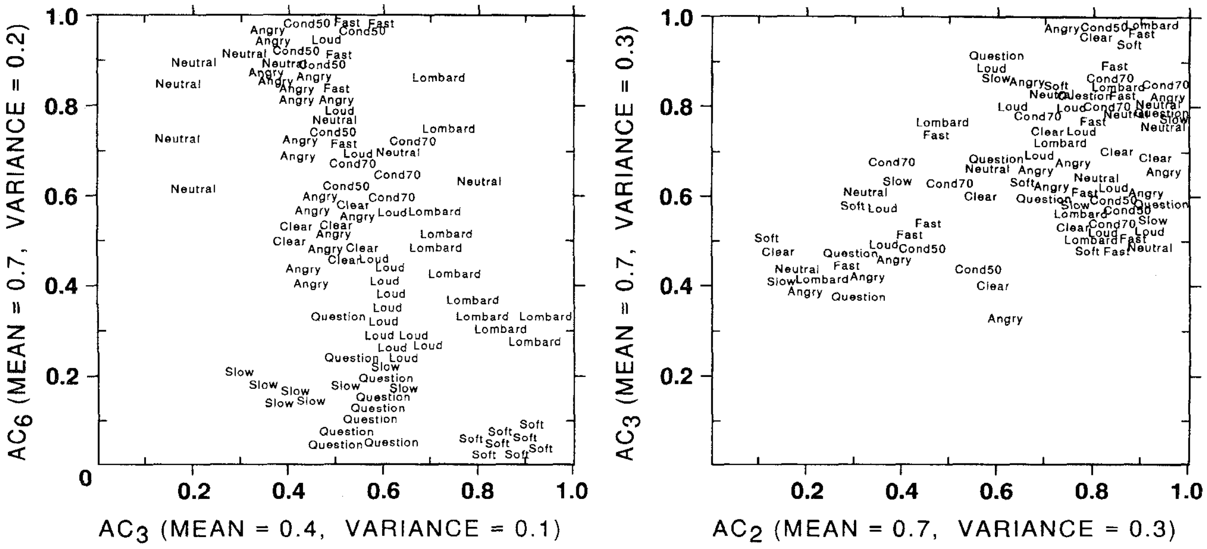
Cepstral-based features have been used extensively in speech recognition applications because they have been shown to outperform linear predictive coefficients. Cepstral-based features attempt to incorporate the nonlinear filtering characteristics of the human auditory system in the measurement of spectral band energies. The five feature sets under consideration here include Mel C_i (C-Mel), delta Mel DC_i (DC-Mel), delta-delta Mel $D2C_i$ (D2C-Mel), AC-Mel, and cross-correlation Mel $XC_{i,j}$ (XC-Mel) cepstral parameters. The first three cepstral features $(C_i, DC_i$, and $D2C_i)$ have been shown to improve speech recognition performance in the presence of noise and *Lombard* effect [2]. The AC_i and $XC_{i,j}$ features are new in that they provide a measure of the correlation between Mel-cepstral coefficients. The Mel-cepstral (C-Mel) parameters are well known as features that represent the spectral variations of the acoustic speech signal. It is suggested that such parameters are useful for stress classification, since, as has been seen, vocal tract and spectral structure vary due to stress. The C-Mel parameters are able to reflect these energy shifts. The DC-Mel and D2C-Mel parameters provide a measure of the "velocity" and "acceleration" of movement of the C-Mel parameters. These features are calculated by performing polynomial fitting of the C-Mel parameters and taking the derivative of the polynomial itself. This may differ from other studies that use a first- and second-order difference method to estimate DC_i and $D2C_i$, respectively. It is suggested that the reason delta parameters are more robust to stress variations is due to their reduced variance across stress conditions. This trait suggests that while these features are more useful for recognition, they may be less applicable to stress classification. It is suggested that the two new derived feature representations (AC-Mel and XC-Mel) could be more successful in representing variations due to stress. The AC-Mel features are calculated as follows:

$$AC_i^{(l)}(k) = \sum_{m=k}^{m=k+L} [C_i(m) * C_i(m+l)] \Bigg/ \sup_k AC_i^{(l)}(k) \quad (1)$$

where

k frame number;
L correlation window length;
l number of correlation lags;
i Mel coefficient index.

When $l = 0$, AC_i models the relative power between frequency bands. For $l > 0$, AC_i models spectral slope and changes in the frame-to-frame correlation variation due to stress. The XC-Mel coefficients are similar to the AC-Mel coefficients, except that the



(a)

(b)

Fig. 4. Separable and nonseparable stressed speech parameters. (a) SEPARABLE vowel /EH/ in "Help" (angry). (b) NON-SEPARABLE vowel /EH/ in "Help" (angry).

cross-correlation is found from one Mel coefficient C_i to another C_j across frames

$$XC_{i,j}^{(l)}(k) = \sum_{m=k}^{m=k+L} [C_i(m) * C_j(m+l)] \Big/ \sup_{k} XC_{i,j}^{(l)}(k). \qquad (2)$$

The XC-Mel parameters $XC_{i,j}$ provide a quantitative measure of the relative change of broad versus fine spectral structure in energy bands. Since the correlation window length $(L = 7)$ and correlation lags $(l = 2)$ are fixed in this study, the correlation terms are a measure of how correlated adjacent frames are over a 72-ms window (24 ms/frame and 8 ms skip rate). It is apparent that both AC-Mel and XC-Mel parameters provide a measure of correlation and relative change in spectral band energies over an extended window frame. From feature analysis, it is suggested that the AC-Mel parameters have similar properties to the XC-Mel parameters. In addition, the AC-Mel parameters can be directly compared with other selected feature sets since they are based on a single coefficient index i. Therefore, AC-Mel parameters will be used for stress classification instead of the XC-Mel parameters.

B. Cluster Analysis of Parameters as Potential Stress Relayers for Classification

A clear visualization of parameter distributions is a beneficial first step for the determination of an optimal stress classification feature set. This is accomplished by obtaining, for a chosen phoneme, pairwise parameter scatter distributions for each frame and each stress condition to be studied. An evaluation over the five parameter representations (C, DC, D2C, AC, XC) considered each feature set's ability to reflect stress variation. Scatter distributions were used to visualize the degree of separability for a selected pair of parameters versus time (i.e., 10 coefficients per parameter set and 495 scatter plots per parameter domain for a total of 2475 possible scatter plots per phoneme). After considering an extensive number of scatter distributions such as that illustrated for the sample /EH/ phoneme in Fig. 4, a number of clear trends emerged that confirmed which speech parameters are better suited for stress classification. In this example, the figure illustrates two pairs of features: the first pair is well separated, and the second is poorly separated. For example, *Lombard* and *Soft* speech is shown to be well separated from all other stress conditions. AC-Mel parameters consistently showed high degrees of separability for those phones considered. Similar analysis was conducted for the DC-Mel and D2C-Mel parameters, indicating a consistently high degree of overlap; hence, they are less appropriate for stress classification. This implies that these parameters are less sensitive to stress effects and, hence, will be more useful for speech recognition [2].

C. Separability Distance Measure

Due to the wide range of features and stress conditions, it is desirable to establish an objective measure to predict stress classification performance. Hence, a measure that assesses a parameter's classification ability is one that increases when the distance between cluster centers increase and variances decrease. The following measure is suggested:

$$d_1(i,j)_{a,b} = \sqrt{(\mu_{(a,i)} - \mu_{(a,j)})^2 + (\mu_{(b,i)} - \mu_{(b,j)})^2} \Big/ [\sigma_{(a,i)} + \sigma_{(a,j)} + \sigma_{(b,i)} + \sigma_{(b,j)}], \quad \text{given} \sum \sigma \neq 0 \qquad (3)$$

where $i = 1, \cdots, n$ and $j = 1, \cdots, n$ are the numbered possible stress conditions, and x_a^0 and x_b^0 are the cluster centers for parameter indexes a and b (for Table I $a = 3$ and $b = 6$). Here, $\mu_{(a,i)}$ reflects the mean and $\sigma_{(a,i)}$ the standard deviation of the ith stress condition for speech feature a. This measure forms a 2-D distance between two speech parameterization classes that is easily visualized. The main underlying assumption of this measure is that the features under test form a Gaussianly distributed convex set. An example of the objective measure of separability is calculated for the two cluster centers x_3^0 and x_6^0 given AC_3 and AC_6 for the /EH/ phoneme in the word "help." The values calculated using (3) are summarized in Table I, providing a pairwise comparison of the separability of two stress conditions. Each mean provides an overall measure of the degree of overlap between a given stress condition and all other stress conditions. Values for d_1 that are greater than the mean indicate better separability than values less than the mean. For example, in

TABLE I

TWO-DIMENSIONAL DISTANCE MEASURE d_1 FOR THE PARAMETERS AC_3 VERSUS AC_6 FOR THE VOWEL /EH/ IN "help"

STRESS CLASS SEPARABILITY MEASURES Single Speaker, AC_3 vs. AC_6, "help", /EH/											
SPEECH	DISTANCE MEASURE										
STYLE	*Angry*	*Clear*	*Cond50*	*Cond70*	*Fast*	*Lombard*	*Loud*	*Normal*	*Question*	*Slow*	*Soft*
Angry	0	0.37	0.56	0.45	0.37	0.64	0.47	0.20	0.89	0.82	1.44
Clear	0.37	0	1.00	0.28	0.53	0.43	0.17	0.43	0.45	0.40	1.00
Cond50	0.56	1.00	0	1.05	0.35	1.00	1.03	0.20	2.03	2.00	2.85
Cond70	0.45	0.28	1.05	0	0.40	0.37	0.20	0.37	0.75	0.79	1.35
Fast	0.37	0.53	0.35	0.40	0	0.52	0.50	0.14	1.00	1.02	1.46
Lombard	0.64	0.43	1.00	0.37	0.52	0	0.29	0.53	0.47	0.60	0.67
Loud	0.47	0.17	1.03	0.20	0.50	0.29	0	0.45	0.40	0.45	0.89
Normal	0.20	0.43	0.20	0.37	0.14	0.53	0.45	0	0.85	0.83	1.27
Question	0.89	0.45	2.03	0.75	1.00	0.47	0.40	0.85	0	0.25	0.71
Slow	0.82	0.40	2.00	0.79	1.02	0.60	0.45	0.83	0.25	0	0.90
Soft	1.44	1.00	2.85	1.35	1.46	0.67	0.89	1.27	0.71	0.90	0
MAXIMUM	1.44	1.00	2.85	1.35	1.46	1.00	1.03	1.27	2.03	2.00	2.85
MEAN	0.62	0.51	1.21	0.60	0.63	0.55	0.48	0.53	0.78	0.81	1.25

Table I, $d_1 = 1.44$ for *Angry* and *Soft*, which is much higher than the mean of 0.62, thus indicating that the AC_3 and AC_6 parameters are well separated for these two stress conditions. Finally, having obtained pairwise intrafeature distance measures $d_1(i,j)_{a,b}$ as shown in Table I, it is desirable to have an overall measure that provides a summary of the differentiating capability of pairwise features across stress conditions. This measure, which has been denoted $d_2(x_a^0, x_b^0)$, estimates the distance between two stress classes a and b as follows:

$$d_2(x_a^0, x_b^0) = \sum_{i=1}^{n}\sum_{j=1}^{n}[(\mu_{(a,i)} - \mu_{(b,i)})^2 + (\mu_{(a,j)} - \mu_{(b,j)})^2] \bigg/ \sum_{i=1}^{n}(\sigma_{(a,i)} + \sigma_{(b,i)}). \tag{4}$$

This measure assesses the n-dimensional "distance" between all n stress classes under consideration. A stress separability evaluation of Mel-cepstral parameters was performed using the d_2 measure for each feature and stress condition across selected phonemes. The global $d_2(x_3^0, x_6^0)$ scores for $(C_i, DC_i, D2C_i,$ and $AC_i)$ were (6.96, 1.42, 1.69, and 7.24), respectively. Hence, the AC-Mel features are the most separable spectral feature set considered based on this distance measure. It is suggested that the broader detail captured by the AC-Mel parameters is more reliable for stress classification.

V. STRESS CLASSIFICATION

There are three issues addressed in this study to demonstrate the viability of a perceptually induced stress classification system. First, fine versus broad stress group definitions are considered to determine if improved stress classification can be achieved. Second, an evaluation is conducted using extracted mono-phones from a i) 35-word versus ii) five-word speech corpus vocabulary training and test set. The 35-word corpus is used to evaluate the performance of the single monophone stress classifier across a larger set of phonemes, whereas the five-word corpus is employed to evaluate a more limited set of phonemes. The first vocabulary is evaluated using a limited five-word versus larger 35-word test set to select the "best" feature set for stress classification. The second vocabulary, which consists of five words different from the first vocabulary, is assessed to establish the level of performance of the proposed stress classification algorithm. Finally, performance of the objective separability measures versus the stress classification rates are compared to select the "best" feature set for stress classification. The goal of the stress classification formulation and evaluations in this study is *not* to find the "best" classification system for stress but rather to obtain the "best" selection from five feature sets for classification.

A. Neural Network Classifier

The proposed neural network classifier consists of a single neural network that is trained with monopartition features (i.e., a single phone class partition). Each partition of speech features is propagated through two hidden layers of the neural network to an output layer that estimates the stress probability scores. The neural network training method employed in this study is the cascade correlation backpropagation network using the extended delta-bar-delta learning rule [8]. This method was selected due to its flexibility. Its strength is its ability to only use as many hidden units as are needed to perform optimal classification. Additionally, this algorithm is capable of forming the complex contoured hypersurface decision boundaries needed for the stress classification problem.

B. Stress Classifier Evaluation

The stress classification algorithm was evaluated using a collection of features derived from frame- to word-level features. Both fine and broad stress classes are evaluated to determine which is more effective for stress classification. The fine (i.e., ungrouped) stress classes are simply the 11 stress conditions in this study. Ungrouped stress class neural network classifier performance is summarized in Table II using the closed 35- and five-word test sets from the first vocabulary under evaluation. Classification rates ranged from 25% to 47% for the 35-word test set, which is greater than chance (i.e., 9%). It is clear that for some stress conditions, such as computer response tasks *Cond50/70, Fast,* and *Soft* spoken speech, significant classification performance is attained. By decreasing the first vocabulary size from 35 to five words, classification rates increased to 60%–61% as summarized in Table II(b). These increased classification rates support the assertion that phonemes are affected differently by stress since the smaller vocabulary has fewer phonemes, and the neural network classifier can then focus on particular variations due to stress.

TABLE II
CLASSIFICATION FOR UNGROUPED CLOSED (a) 35 AND (b) FIVE WORD TESTS

STRESS CLASSIFICATION PERFORMANCE				
Single Speaker, 35 Words, Stress Ungrouped "Brake", "East", "Freeze", "Help", "Steer" CLOSED VOCABULARY TEST SET				
STRESS CLASS	CLASSIFICATION RATE (%)			
	C_i	DC_i	$D2C_i$	AC_i
Angry	6.20	0.00	19.49	4.96
Clear	12.50	0.00	4.27	7.32
Cond50/70	42.47	59.76	56.08	14.61
Fast	7.26	1.65	44.53	68.10
Lombard	1.64	0.00	8.53	2.54
Loud	12.40	3.73	3.15	1.69
Normal	22.31	0.00	2.61	1.72
Question	14.05	0.00	5.15	4.76
Slow	19.01	4.76	27.56	4.31
Soft	16.53	33.09	0.00	2.38
OVERALL	33.14	24.69	47.31	32.57

(a)

STRESS CLASSIFICATION PERFORMANCE				
Single Speaker, 5 Words, Stress Ungrouped "Brake", "East", "Freeze", "Help", "Steer" CLOSED VOCABULARY TEST SET				
STRESS CLASS	CLASSIFICATION RATE (%)			
	C_i	DC_i	$D2C_i$	AC_i
Angry	82.35	82.35	59.09	29.41
Clear	41.18	35.29	53.33	23.53
Cond50/70	79.41	79.41	74.29	31.43
Fast	76.47	76.47	95.24	100.00
Lombard	64.71	64.71	46.67	47.06
Loud	82.35	82.35	78.57	43.75
Normal	58.82	52.94	71.43	5.88
Question	70.59	70.59	84.21	17.65
Slow	47.06	47.06	45.45	17.65
Soft	52.94	47.06	64.29	11.76
OVERALL	59.98	60.51	61.16	61.40

(b)

TABLE III
CLASSIFICATION FOR GROUPED FIVE WORD (a) IN-VOCABULARY CLOSED AND (b) OUT-OF-VOCABULARY OPEN TESTS

STRESS CLASSIFICATION PERFORMANCE				
Single Speaker, 5 Words, Stress Grouped "Brake", "East", "Freeze", "Help", "Steer" CLOSED VOCABULARY TEST SET				
STRESS GROUP	CLASSIFICATION RATE (%)			
	C_i	DC_i	$D2C_i$	AC_i
G_1	85.29	73.53	88.89	94.12
G_2	92.69	89.71	91.04	86.76
G_3	70.59	70.59	87.50	85.29
G_4	70.59	76.47	76.47	82.35
G_5	76.47	52.94	76.47	64.71
G_6	76.47	41.18	60.00	70.59
G_7	76.47	70.59	57.89	94.12
OVERALL	78.90	76.94	79.32	80.62

(a)

STRESS CLASSIFICATION PERFORMANCE				
Single Speaker, 5 Words, Stress Grouped "Degree", "Enter", "Fifty", "Strafe", "Three" OUT-OF-VOCABULARY TEST SET				
STRESS GROUP	CLASSIFICATION RATE (%)			
	C_i	DC_i	$D2C_i$	AC_i
G_1	53.33	41.86	91.30	80.95
G_2	90.20	82.69	80.00	83.33
G_3	45.45	61.90	47.83	57.14
G_4	68.75	57.14	53.33	32.26
G_5	55.00	31.58	36.84	50.00
G_6	53.33	22.22	42.86	53.85
G_7	61.11	40.00	60.00	76.47
OVERALL	47.82	42.85	44.90	48.38

(b)

Next, broad (i.e., grouped) stress classes are evaluated by combining perceptually similar stress conditions that may cluster in similar domains. Note that this grouping resulted from informal listening tests as to which stressed conditions were perceptually similar, (i.e., G_1 (*Angry, Loud*), G_2 (*Cond50, Cond70, Normal, Soft*), G_3 (*Fast*), G_4 (*Question*), G_5 (*Slow*), G_6 (*Clear*), and G_7 (*Lombard*)). Employing stress class grouping, classification rates are further improved by +17%–20% to 77%–81% (compare Table II(b) with Table III(a)). Hence, it is shown that stress class grouping using less confuseable subgroups improves classifier performance. It is suggested that further improvement in classification could be accomplished using a two-step decision process in which grouped stress conditions are more finely discriminated in a second stage if a larger speech corpus is used or if noise is present. Finally, the performance of the stress classification system is evaluated using the second five-word out-of-vocabulary test set with similar phoneme content. Classification rates ranged from 43%–48% as shown in Table III(b), which is greater than chance (i.e., 14.3%). These results agree with the expected stress class differentiability of the AC-Mel feature set based on objective separability measures.

VI. CONCLUSIONS

In this study, a stress-sensitive feature set has been proposed for use in stress classification. Further, a monopartition stress classification system has been formulated using neural networks. An analysis was performed for five speech feature representations as potential stress relayers. Features were considered with respect to the following:

i) pair-wise stress class separability;
ii) a numerical pair-wise and global objective measure of feature separability across stressed conditions;
iii) 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 on stress condition and phoneme group. The results have demonstrated the effects of speaker stress on both micro (phoneme) and macro (whole word or phrase) levels. Phoneme classes are affected differently by stress. For example, the unvoiced consonant stops (/P/, /K/, /T/) are perturbed little by stress, whereas vowels (/AE/, /EH/, /IH/, /ER/, /UH/) are significantly effected. In addition, coarticulation effects are more critical for stressed speech since stress variation across a phoneme sequence is more pronounced for an isolated phoneme. Hence, an algorithm that uses a front-end phoneme group classifier could improve overall stress classification performance [10]. It was shown that the autocorrelation of the Mel-cepstral (AC-Mel) parameters are the most useful features considered for separating the selected stress conditions.

Next, a cascade correlation extended delta-bar-delta-based neural network was formed using each feature to determine stress classification performance. Classification rates across 11 stress conditions were 79% for in-vocabulary and 46% for out-of-vocabulary tests (which are both greater than chance 14.3%), further confirming that AC-Mel parameters are the most separable feature set considered. In conclusion, this study has shown that a particular speech feature representation can influence stress classification performance for different stress styles/conditions and that a neural network-based classifier over word or phoneme partitions can achieve good classification performance. It is suggested that such knowledge would be useful for monitoring speaker state, as well as ultimately contributing to improvements in speech coding and recognition systems [10].

References

[1] D. A. Cairns and J. H. L. Hansen, "Nonlinear analysis and detection of speech under stressed conditions," *J. Acous. Soc. Amer.,* vol. 96, no. 6, pp. 3392–3400, 1994.

[2] B. A. Hanson and T. Applebaum, "Robust speaker-independent word recognition using instantaneous, dynamic and acceleration features: Experiments with Lombard and noisy speech," in *Proc. ICASSP*, Apr. 1990, pp. 857–860.

[3] J. H. L. Hansen, "Morphological constrained feature enhancement with adaptive cepstral compensation (MCE-ACC) for speech recognition in noise and Lombard effect," *IEEE Trans. Speech Audio Processing,* vol. 2, pp. 598–614, Oct. 1994.

[4] J. H. L. Hansen and S. E. Bou-Ghazale, "Robust speech recognition training via duration and spectral-based stress token generation," *IEEE Trans. Speech Audio Processing,* vol. 3, pp. 415–421, Sept. 1995.

[5] J. H. L. Hansen, "A source generator framework for analysis of acoustic correlates of speech under stress. Part I: Pitch, duration, and intensity effects," *J. Acous. Soc. Amer.,* submitted for publication.

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

[7] R. P. Lippmann, E. A. Martin, and D. B. Paul, "Multistyle training for robust isolated-word speech recognition," in *Proc. ICASSP*, Apr. 1987, pp. 705–708.

[8] A. A. Minai and R. D. Williams, "Back-propagation heuristics: A study of the extended delta-bar-delta algorithm," in *IJCNN*, June 17-21, 1990, pp. 595–600.

[9] B. J. Stanton, L. H. Jamieson, and G. D. Allen, "Robust recognition of loud and Lombard speech in the fighter cockpit environment," in *Proc. ICASSP*, May 1989, pp. 675–678.

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

[11] H. Wakita, "Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms," *IEEE Trans. Audio Electroacoust.*, vol. AU-21, pp. 417–27, Oct. 1973.

An Extended Clustering Algorithm for Statistical Language Models

Joerg P. Ueberla

***Abstract*— An existing clustering algorithm is extended to deal with higher order N-grams and a faster heuristic version is developed. Even though results are not comparable to back-off trigram models, they outperform back-off bigram models when many million words of training data are not available.**

I. Introduction

It is well known that statistical language models often suffer from a lack of training data. This is true for standard tasks and even more so when one tries to build a language model for a new domain, because a large corpus of texts from that domain is usually not available. One frequently used approach to alleviate this problem is to construct a class-based language model. Let $W = w_1, \cdots, w_m$ be a sequence of words from a vocabulary V and let $G{:}w \rightarrow G(w) = g_w$ be a function that maps each word w to its class $G(w) = g_w$. A class based bigram language model calculates the probability of seing the next word w_i as

$$p(w_i|w_{i-1}) = p_G(G(w_i)|G(w_{i-1})) * p_G(w_i|G(w_i)). \qquad (1)$$

In order to derive the clustering function G automatically, a clustering algorithm as shown in Fig. 1 can be used (see [2]). In the spirit of decision-directed learning, it uses as optimization criterion a function F that is very closely related or identical to the final performance measure one wishes to maximize. As suggested in [2], F is based in all our experiments on the leaving-one-out likelihood of the model generating the training data.

In Section II, the algorithm is extended so that it can cluster higher order N-grams. When such a clustering algorithm is applied to a large training corpus, e.g., the Wall Street Journal (WSJ) corpus, with tens of millions of words, the computational effort required can easily become prohibitive. Therefore, a simple heuristic to speed up the algorithm is developed in Section III. It can then be applied more easily to the WSJ corpus and the obtained results will be presented in Section IV.

II. Extending the Clustering Algorithm to N-Grams

As shown in [6], there are several ways of extending the algorithm to higher order N-grams. The method we chose uses two clustering functions G_1 and G_2:

$$\begin{aligned} &p(w_i|w_{i-N+1}, \cdots, w_{i-1}) \\ &\quad = p_G(G_2(w_i)|G_1(w_{i-N+1}, \cdots, w_{i-1})) * p_G(w_i|G_2(w_i)) \end{aligned} \qquad (2)$$

G_1 is a function that maps the current context $c = (w_{i-N+1}, \cdots, w_{i-1})$ into one of a set of context equivalent classes (or states). Thus any two contexts, which are mapped by G_1 to the same class, will have identical probability distributions. G_2, as the G of (1), maps

Manuscript received December 16, 1994; revised February 28, 1996. The associate editor coordinating the review of this paper and approving it for publication was Dr. Xuedong Huang.

The author is with Forum Technology, DRA Malvern, Malvern, Worc. WR14 3PS, U.K.

Publisher Item Identifier S 1063-6676(96)05074-2.