ESD-TR-86-098

Technical Report
753

Cepstral Domain Talker Stress Compensation for Robust Speech Recognition

Y. Chen

10 November 1986

Lincoln Laboratory

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

LEXINGTON MASSACHUSETTS

**Prepared for the Defense Advanced Research Projects Agency
under Electronics Systems Division Contract F19628-85-C-0002.**

Approved for public release; distribution unlimited.

The work reported in this document was performed at Lincoln Laboratory, a center for research operated by Massachusetts Institute of Technology. This work was sponsored by the Defense Advanced Research Projects Agency under Air Force Contract F19628-85-C-0002 (ARPA Order 5328).

The views and conclusions contained in this document are those of the contractor and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the United States Government.

The ESD Public Affairs Office has reviewed this report, and it is releasable to the National Technical Information Service, where it will be available to the general public, icluding foreign nationals.

This technical report has been reviewed and is approved for publication.

FOR THE COMMANDER

Thomas J. Alpert

Thomas J. Alpert, Major, USAF
Chief, ESD Lincoln Laboratory Project Office

Non-Lincoln Recipients

PLEASE DO NOT RETURN

Permission is given to destroy this document when it is no longer needed.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
LINCOLN LABORATORY

CEPSTRAL DOMAIN TALKER STRESS COMPENSATION FOR ROBUST SPEECH RECOGNITION

Y. CHEN
Group 24

TECHNICAL REPORT 753

10 NOVEMBER 1986

LEXINGTON MASSACHUSETTS

ABSTRACT

Automatic speech recognition algorithms generally rely on the assumption that for the distance measure used, intraword variabilities are smaller than interword variabilities so that appropriate separation in the measurement space is possible. As evidenced by degradation of recognition performance, the validity of such an assumption decreases from simple tasks to complex tasks, from cooperative talkers to casual talkers, and from laboratory talking environments to practical talking environments.

This report presents a study of talker-stress-induced intraword variability, and an algorithm that compensates for the systematic changes observed. The study is based on Hidden Markov Models trained by speech tokens spoken in various talking styles. The talking styles include normal speech, fast speech, loud speech, soft speech, and talking with noise injected through earphones; the styles are designed to simulate speech produced under real stressful conditions.

Cepstral coefficients are used as the parameters in the Hidden Markov Models. The stress compensation algorithm compensates for the variations in the cepstral coefficients in a hypothesis-driven manner. The functional form of the compensation is shown to correspond to the equalization of spectral tilts.

Preliminary experiments indicate that a substantial reduction in recognition error rate can be achieved with relatively little increase in computation and storage requirements.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

CEPSTRAL DOMAIN TALKER STRESS COMPENSATION FOR ROBUST SPEECH RECOGNITION

1. INTRODUCTION

Current speech recognition systems generally degrade significantly in performance if the systems are not both trained and tested under similar talking conditions. A major reason for performance degradation when testing and training conditions differ is that people speak differently under different conditions. Previous research has demonstrated that differences in speech patterns can be caused by psychological or emotional stress,[1-5] by the presence of intense background noise,[5-7] by a demanding perceptual-motor or mental task,[2-4] by physical exertion,[2] and by natural inconsistencies in pronunciation.[8,9] Despite the knowledge that speech patterns change in stress and in noise and the demonstration of degraded recognition performance in stress, little speech recognition research has been directed at modeling systematic changes observed and at developing recognition systems that are resistant to such changes.

This report presents a study of talker-stress-induced variations in speech cepstral coefficients, and an algorithm that compensates for systematic (but unknown *a priori*) changes observed. The study is based on an isolated-word Hidden Markov Model speech recognizer trained by speech spoken in various talking conditions. The organization of this report is as follows. In Section 2 a speech data base that has been used in the study is described. In Section 3 a baseline Hidden Markov Model (HMM) speech recognizer is defined. In Section 4 a multistyle training experiment is described. The success of the multistyle training experiment has prompted the study of the stress-induced changes in speech, which is discussed in Section 5, and the development of a stress compensation algorithm, which is discussed in Section 6. The compensation algorithm may be interpreted as an adaptive, word-hypothesis-driven form of spectral tilt equalization. Section 7 presents experimental results.

Except otherwise noted, the notations used in this report are as follows: boldfaced upper case letters such as "**A**" represent matrices, boldfaced lower case letters such as "**b**" represent column vectors, and lower case letters such as "c" represent scalars. Elements of matrices and vectors may be written as scalars with an appropriate number of subscripts. The lower case letters i and j are used as indices, the upper case letters N and M are used to indicate dimensionalities. Therefore, the matrix **A** may be written as $[a_{ij}]_{MN}$, the vector **b** may be written as $[b_i]_N$. Lower case letters followed by one or more arguments enclosed in parentheses, such as "f(x)," represent functions.

2. THE "SIMULATED STRESS" SPEECH DATA BASE

The studies and experiments conducted in this research were based on the "simulated stress"[10] speech data base recently collected by Texas Instruments, Inc.

In this data base stress-like degradations of the speech signal were elicited by asking the speaker to produce speech in a variety of styles (normal, fast, loud, soft, and shout) as well as with 95-dB pink noise exposure in the ear to produce the Lombard effect.[5] The vocabulary consisted of 105 words, including monosyllabic, polysyllabic, and confusing words. A complete list of the words is given in the Appendix.

The data base was divided into training data and test data. Training data consisted of five samples of each of the 105 words collected in a random order under normal talking conditions, and test data consisted of two samples of each word under each simulated-stress condition. Data were collected from five adult males and three adult females using a 16-bit A/D converter, sampled at 20-kHz rate, in a quiet laboratory environment. The data were downsampled to 8 kHz for laboratory usage. The total number of test word tokens was 10,080.

To verify the effects of "simulated stress," a baseline Hidden Markov Model based recognizer (to be described in the next section) has been tested on this data base. Substitution rate (no rejection or deletion was allowed in the experiments) of this test is given in Table I. It is seen that, relative to normally spoken speech, the error rate increases significantly for the various style conditions. It is the purpose of this research to understand the causes for the performance degradation experienced, and to develop effective means to compensate for them.

TABLE I

Substitution Rate (Percent):
The "Simulated Stress" Data Base

Condition	Norm	Fast	Loud	Noise	Soft	Shout	Avg5*	Avg6†
Baseline HMM	1.0	6.1	29.1	19.6	13.5	86.4	13.9	25.9

* Avg5 is the average error rate of all talking conditons except shout.

† Avg6 is the average error rate of all talking conditions.

3. A HIDDEN MARKOV MODEL SPEECH RECOGNIZER

The theory of Hidden Markov Models and the application of HMM to automatic speech recognition can be found in a number of papers.[11-16]

Figure 1 shows the type of HMM we used in this research. The word model network is a linear sequence of nodes with no skip branches. The model is intended to be used on speech inputs consisting of one word with background (silence) at each end; the first and the last nodes are background nodes to provide a semi-open-endpoint recognizer.

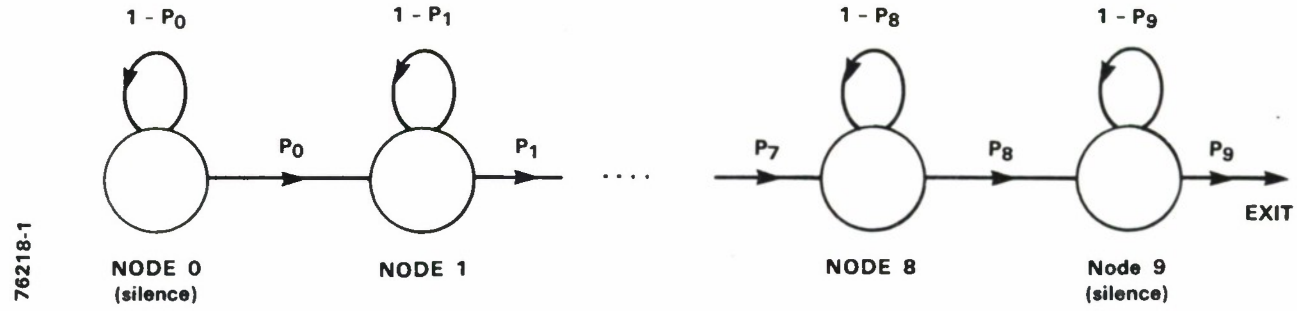

Figure 1. The 10-node left-to-right Markov model.

The HMM model can be described by a matrix-vector pair $\{\mathbf{A},\mathbf{b}\}$. $\mathbf{A} = [a_{ij}]$ is a bidiagonal state transition matrix, its elements are given by

$$a_{ij} = \begin{cases} p_i \quad , & j = i + 1 \quad , \\ 1 - p_i \quad , & j = i \quad , \\ 0 \quad , & \text{otherwise} \quad . \end{cases} \tag{1}$$

Note that p_i is the transition probability from node i to node i + 1, and $1 - p_i$ is the self-looping probability of node i. At each state a vector $\mathbf{v}$ of continuous variables is observed. The probabilistic nature of the vector $\mathbf{v}$ is described by a set of joint probability density functions $\mathbf{b} = [f_i(\mathbf{v})]$.

The observation vector $\mathbf{v}$ contains 12 mel-frequency cepstral coefficients, i.e., $\mathbf{v} = [v_j]_{12}$. The cepstral coefficients are similar to those used by Davis and Mermelstein.[17] To compute a cepstral vector, 160 speech samples were read from the input, padded with 96 zeros, windowed by a 256-sample Hamming function and transformed into the frequency domain via 256-point FFT. In the frequency domain, magnitude of the spectrum is squared, multiplied by the function

$$g(f) = 1 + \frac{f^2}{250000} \quad , \tag{2}$$

where f is frequency in hertz, to boost high frequency content. Logarithms of the frequency samples are then taken. A set of 24 triangular-shaped windows (see Figure 2) are then used to computer averaged log spectral parameters $\mathbf{x} = [x_i]_{24}$. Notice that the bandwidths of the windows increase as their center frequencies increase; the areas under the windows are kept constant.

From these averaged log spectral parameters $\mathbf{x}$ the cepstral coefficients v_j are computed as

$$v_j = \sum_{k=1}^{24} x_k \cos[j(k - \frac{1}{2}) \frac{\pi}{24}] \qquad j = 1,2, \ldots ,12 \tag{3}$$

Each node of the Hidden Markov Model is represented by a cepstral vector template which in turn is characterized by a jointly normal distribution with mean vector $\mathbf{c}$ and covariance

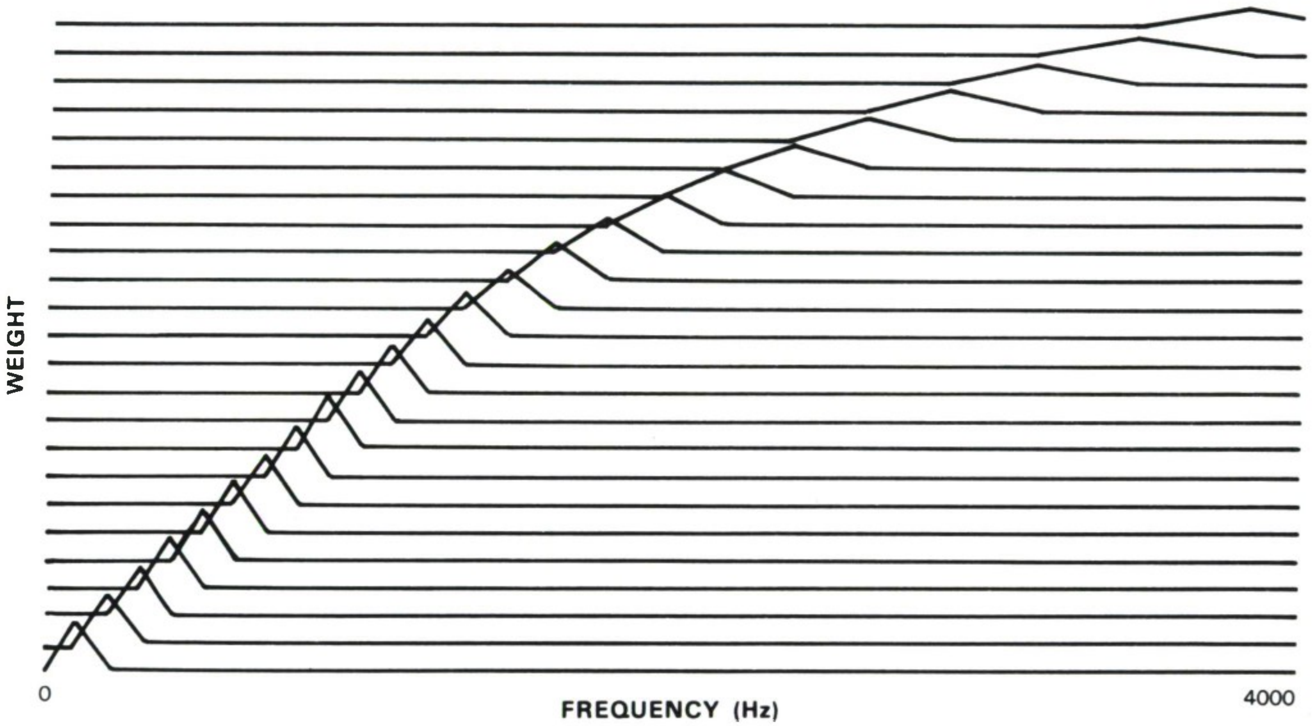

Figure 2. Filter bank used in Equation (1). (Courtesy of D.B. Paul.)

76218-2

matrix **R**. In our model, we assume that all off-diagonal elements of **R** are zero. [This is justified in part by the fact that the result of the cosine transform in Equation (4) is approximately mutually uncorrelated.[18,19] Further justification is provided by the good recognition results obtained by us and by others using this assumption.] With this assumption, the covariance matrix can be reduced to a vector of variances, which we relabeled as **r**.

In summary, the node transition characteristic of a Hidden Markov word model is described by the transition probability matrix **A**; the observation parameter statistics of each node are described by the cepstral mean vector **c** and the cepstral variance vector **r**.

Training of the models uses the forward-backward algorithm,[20] while recognition uses the Viterbi decoding algorithm.[21] Cepstral coefficients are computed once every 10 ms.

Since the recognizer makes a forced decision on each input word, substitution is the only type of error considered here.

4. AN EXPERIMENT ON MULTISTYLE-TRAINED HIDDEN MARKOV WORD MODELS

A number of different training/testing procedures could be used to improve speech recognition performance under stress. Ideally, a recognition system could be trained and tested under the same stress condition. This, however, is often not possible. A second alternative is to use dynamic adaptation of the models based on recent recognition results. This, again, may not be a satisfactory solution because stress conditions are transient. However, it is possible to use multistyle training.[22] Multistyle training requires a talker to train a recognizer using words spoken with different talking styles instead of using words all spoken normally. It has been found to be easy for a talker to change to styles such as fast, slow, loud, and soft, producing changes in speech characteristics that are similar to changes that occur under stress. It remains to be demontrated that multistyle training produces improved recognition under stress.

An experiment on multistyle-trained Hidden Markov Model word recognition was performed. In this experiment, 11 speech tokens were used to train each word model: 5 tokens from the training data base, and 6 tokens, one per talking style except normal, from the test data base. Recognition tests were conducted on the remaining half of the test data base. The recognition error rates are listed in Table II. For comparison, the error rate of the baseline HMM system is also included.

In comparing experimental results listed in Table II, we see a dramatic improvement in recognition performance. It appears that the HMM word models were able to assimilate the data from the multiple styles and to capture statistically the more invariant features of each word. In the next section we investigate the gross changes of model parameters resulting from multistyle training as well as from style training (as opposed to normal training).

TABLE II

Substitution Rate (Percent): A Comparison of Normal- and Multistyle-Trained HMM Recognizers

Condition	Norm	Fast	Loud	Noise	Soft	Shout	Avg5	Avg6
Baseline HMM*	1.0	6.1	29.1	19.6	13.5	86.4	13.9	25.9
Multistyle†	0.5	5.6	5.1	2.1	5.8	43.6	3.8	10.5

* The baseline system was trained with 5 normally spoken word tokens per talker and tested on 10,080 test tokens.

† The multistyle-trained system was trained on 11 style speech tokens per talker and tested on 5,040 test tokens.

5. CEPSTRAL DOMAIN STRESS COMPENSATION — DRIVEN BY OBSERVATIONS

The success of the multistyle training experiment motivated a comparison of the model parameters trained under various talking styles to determine whether it would be possible to compensate for the cepstral changes through simple transformations on the cepstral means and variances obtained using normal training. Such transformation, if effective, would eliminate the need for asking the user to train the system with multiple styles and for incorporating multiple style data in the Forward-Backward training.

The differences among normally trained, single-style-trained, and multistyle-trained word models are partially reflected in the average shifts of the mean values and in the average scaling of the variances of the cepstral coefficients. To study such differences, seven different sets of word models were examined. Six of the models were trained under six individual conditions (normal training, fast, loud, Lombard, soft, and shout, respectively), while the seventh was trained using a composite of all these conditions (multistyle). The cepstral means and variances, averaged over all words in the TI vocabulary, over all speech nodes in each word, and over all talkers, were computed for each of the models above.

The mean cepstral shifts (i.e., cepstral means of the given model minus the cepstral means of the normal model) for each of the cepstral coefficients are tabulated in Table III(a). Figure 3(a) plots mean cepstral shifts for four cases: soft; shout; average of fast, loud, and Lombard; and multistyle. Figure 3(b) plots the corresponding spectra of these mean shifts, contrasting the effects on spectral tilt of low vocal effort (soft) vs higher vocal effort (fast, loud, Lombard, and shout). Increased vocal effort increases the relative high frequency content, whereas the opposite occurs with low vocal effort.

TABLE III

Mean and Variance Variations in Style-Speech

(a) Mean Shifts in Cepstral Domain*

Coeff	Fast	Loud	Lomb	Soft	Shout	Multi	AVG
1	-0.61	-1.14	-0.84	0.90	-3.08	-0.61	-1.07
2	-0.26	-0.50	-0.59	0.43	-1.94	-0.43	-0.59
3	-0.03	-0.48	-0.34	0.37	-1.28	-0.26	-0.37
4	-0.06	-0.50	-0.36	0.49	-1.09	-0.25	-0.39
5	-0.02	-0.18	-0.08	0.27	-0.53	-0.10	-0.13
6	-0.02	-0.38	-0.29	0.27	-0.70	-0.17	-0.29
7	-0.02	-0.11	-0.05	0.12	-0.19	-0.03	-0.07
8	0.04	-0.18	-0.06	0.05	-0.40	-0.07	-0.09
9	-0.05	-0.16	-0.08	-0.01	-0.32	-0.06	-0.12
10	-0.06	-0.23	-0.10	-0.02	-0.53	-0.13	-0.17
11	-0.04	-0.13	-0.11	-0.03	-0.20	-0.10	-0.13
12	-0.06	-0.15	-0.15	0.04	-0.09	-0.05	-0.14

(b) Ratio of Variance

Coeff	Multi	Coeff	Multi
1	2.07	7	1.55
2	1.84	8	1.70
3	1.75	9	1.84
4	1.71	10	1.77
5	1.47	11	1.83
6	1.62	12	2.10

* Because the differences in the fast, loud, and Lombard conditions are reasonably similar, their averages are listed in the last column.

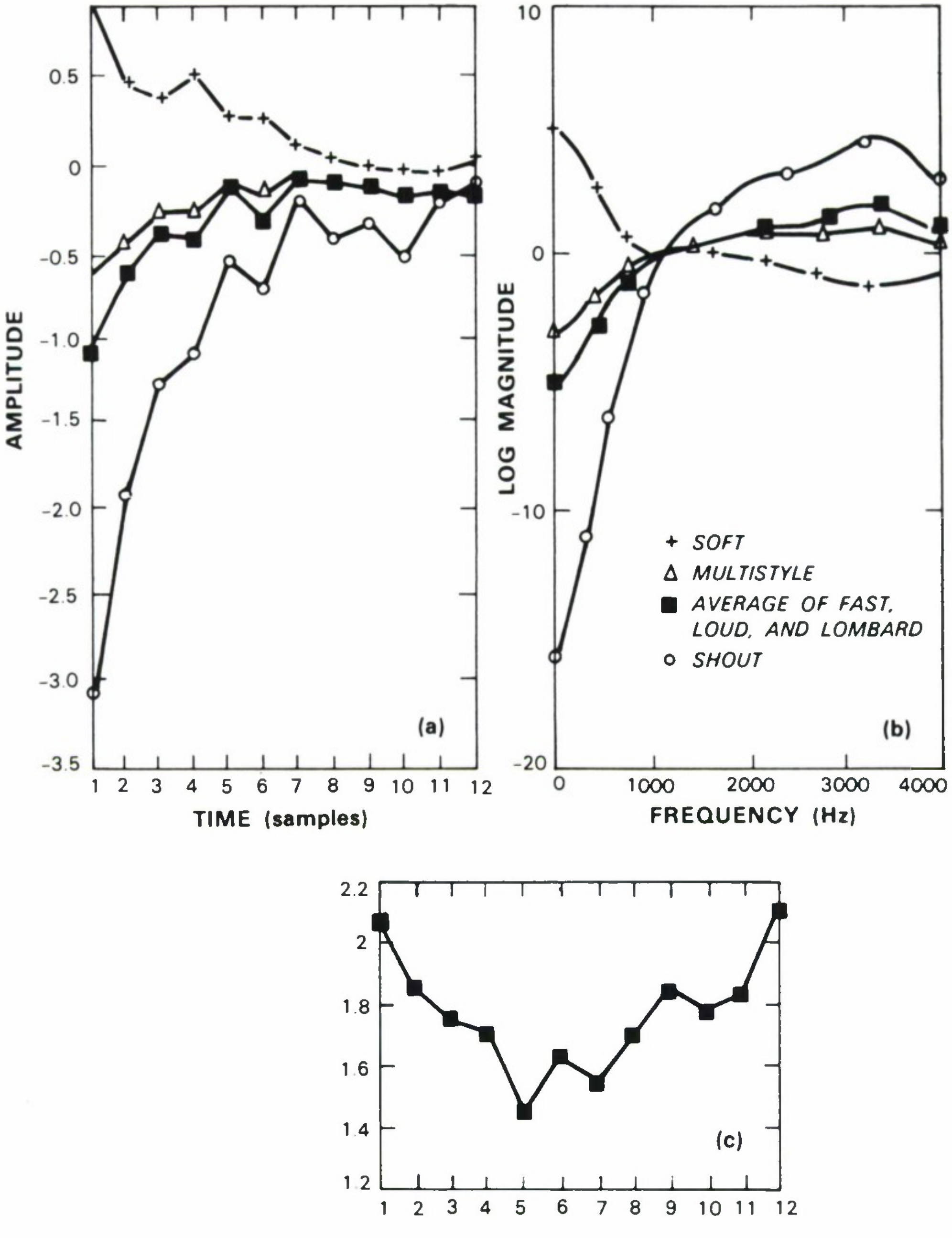

Figure 3. *Variations of cepstral coefficient compared to normally spoken words. (a) Difference of mean (style minus normal), (b) spectra of differences of mean, and (c) ratio of variance (multistyle normal).*

It is well known that spectral tilt exhibits large variation when a talker speaks under stress. Such variation usually contaminates the distance measure and is one of the most significant causes of recognition performance degradation. It appears that the effect of spectral tilt could be compensated, to some extent, by applying the appropriate cepstral compensation to normally trained word models.

Because variance estimation is less reliable than mean estimation, we have only compared cepstral variances of multistyle-trained models which used 11 training tokens with the normally trained models. Their ratios (multistyle/normal) are listed in Table III(b) and plotted in Figure 3(c). It appears that the major style-induced variations occur in the most slowly varying spectral components (corresponding to lower order cepstral coefficients), and in the most rapidly varying spectral components (corresponding to the higher order coefficients).

The following cepstral compensation experiments were performed, in which new word models were generated by modifying normally trained Hidden Markov word models by one or more sets of cepstral differences. The word models were talker-dependent, but the modifications were the same for all words and all talkers.

(a) *Single Model Compensation:—* The set of cepstral mean differences and variance ratios observed in multistyle-trained models [represented by filled squares in Figure 3(a) and (c)] was applied as compensation in recognition tests on all styles.

(b) *Multimodel Compensation:—* Three sets of cepstral mean compensations corresponding to the soft, the loud, and the shout-trained models, as well as cepstral variance ratios for multistyle-trained models, were applied to generate three new word models; together with the original normally trained word model, they were used in recognition tests on all styles. In recognition, the four models were used for each vocabulary word, and were treated independently and equally; in effect, the computation for HMM recognition was quadrupled.

The recognition error rates of these experiments are listed in Table IV along with the error rates of the baseline system and the multistyle-trained system for comparison. The error rate reductions relative to the baseline system seem quite promising given the simplicity of the compensation technique.

It is not clear how many different styles it would be useful to add in similar experiments before recognition performance would start to decline. However, evidence recently gathered[23] indicates that a small number of well-selected styles might be sufficient. The next section discusses a variation of the above technique — hypothesis-driven stress compensation.

TABLE IV

**Substitution Rate (Percent):
A Comparison of Fixed Stress Compensation**

Condition	Norm	Fast	Loud	Noise	Soft	Shout	Avg5	Avg6
Baseline HMM	1.0	6.1	29.1	19.6	13.5	86.4	13.9	25.9
Multistyle	0.5	5.6	5.1	2.1	5.8	43.6	3.8	10.5
Single Model	1.2	4.6	15.2	12.2	15.4	79.5	9.7	21.4
Multimodel	1.0	4.2	12.1	6.7	5.5	68.7	5.9	16.4

6. CEPSTRAL DOMAIN STRESS COMPENSATION — A HYPOTHESIS-DRIVEN APPROACH

It is the high cost of increased computation and the uncertainty about training-style sufficiency and efficiency that prompted us to search for alternatives. As a result of this effort, the hypothesis-driven cepstral mean compensation technique, which adapts to the input speech and to the hypothesized reference word, was developed. Fixed multistyle variance compensation has been found beneficial for all styles and will be used in conjunction with the adaptive mean compensation, unless stated otherwise.

As depicted in Figure 4, a talker is modeled as an information source that puts out a sequence of deterministic cepstral vectors $\{\mathbf{c}_t\}$.* Before the vectors are received by the decoder, we assume that they undergo two stages of contamination.

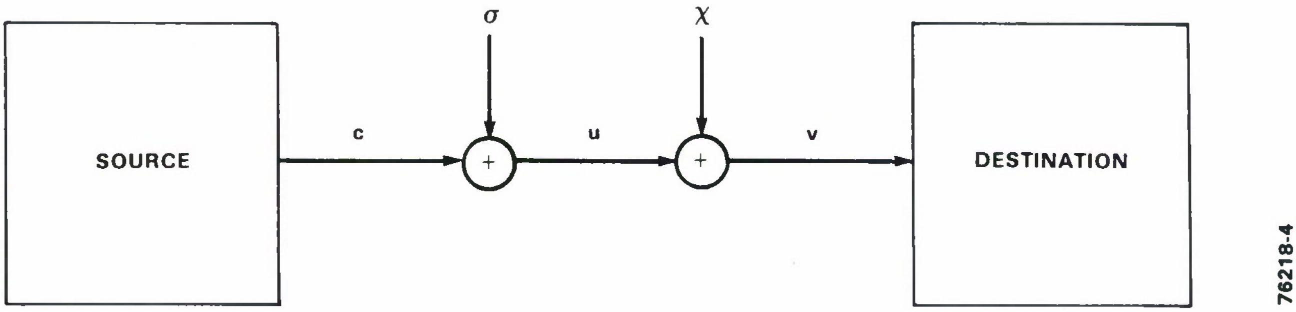

Figure 4. Model of the contamination of cepstral coefficients, where σ is a random noise factor and χ is a deterministic, but unknown, stress factor.

* The subscript t is an index of time.

Stage 1

A sequence of independent identically distributed (i.i.d.) random vectors $\{\delta_t\}$ is added to the cepstral sequence $\{\mathbf{c}_t\}$ to create a new sequence $\{\mathbf{u}_t\}$:

$$\mathbf{u}_t = \mathbf{c}_t + \delta_t \quad . \tag{4}$$

The sequence $\{\delta_t\}$ models the randomness of speech cepstral parameter outputs; its elements are assumed to be normally distributed with zero mean vector and diagonal covariance matrix (see discussion in Section 3).

Stage 2

A deterministic but unknown vector χ is added to the sequence $\{\mathbf{u}_t\}$ to create the observation sequence $\{\mathbf{v}_t\}$, i.e.,

$$\mathbf{v}_t = \mathbf{u}_t + \chi \quad . \tag{5}$$

The vector χ is the additive "stress" component. It is assumed to have the functional form [see Figure 3(a)]:

$$\chi_i = a\ e^{-b(i-1)} \quad , \tag{6}$$

and is further assumed to remain unchanged within a word interval.

Given a sequence of observations $\mathbf{v}_t$, t = 1,2,...,T we have developed a procedure for estimation, based on maximum likelihood principles, of the parameters a and b in Equation (6). The remaining part of this section deals with the derivation of this estimation procedure; readers who are only interested in the experimental results may skip to the next section without loss of continuity.

Our parameter estimation procedure is divided into two steps, the estimation of χ_i and the smoothing:

Step 1 (Estimating χ_i)

The probability density function of the observation v_i is given by

$$f(v_i) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[- \frac{(v_i - c_i - \chi_i)^2}{2\sigma_i^2} \right] \quad . \tag{7}$$

The likelihood function of the i^{th} observation variable, v_{it} for t = 1,...,T, conditioned on both the sequence of the i^{th} cepstral coefficient c_{it}, t = 1,...,T and the i^{th} "stress" component χ_i is given by

$$l(v_{i1},\ldots,v_{iT} | c_{i1},\ldots,c_{iT},\chi_i) = \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[- \frac{(v_{it} - c_{it} - \chi_i)^2}{2\sigma_i^2} \right] \quad . \tag{8}$$

Taking logarithms of both sides, we obtain the log likelihood function

$$L(v_{i1},\ldots,v_{iT}|\, c_{i1},\ldots,c_{iT},\chi_i) = -T \log\sqrt{2\pi\sigma_i} - \sum_{t=1}^{T} \frac{(v_{it} - c_{it} - \chi_i)^2}{2\sigma_i^2} \quad . \tag{9}$$

The $\bar{\chi}_i$ that maximize Equation (9) is the classical maximum likelihood estimate of χ_i, it is given by

$$\begin{aligned}\bar{\chi}_i &= \frac{1}{T} \sum_{t=1}^{T} (v_{it} - c_{it}) \\ &= \frac{1}{T} \sum_{t=1}^{T} v_{it} - \frac{1}{T} \sum_{t=1}^{T} c_{it} \quad .\end{aligned} \tag{10}$$

We replace the sample average of c_i, which is not observable, by the expected average value, drived from the word hypothesis:

$$\begin{aligned}\bar{c}_i &= E\left[\frac{\sum_i \tau_i c_i}{\sum_j \tau_j}\right] \\ &= \sum_i E\left[\frac{\tau_i}{\sum_j \tau_j}\right] c_i \quad ,\end{aligned} \tag{11}$$

where the τ_i's are a set of mutually independent discrete random variables whose values represent the dwell time in each of the i nodes, and the summations are over all speech nodes.

In Equation (11), τ_i is known to have geometric probability mass function

$$P_{\tau_i}(k) = p_i(1 - p_i)^{k-1} \quad , \quad k = 1,2,\ldots, \tag{12}$$

where $1 - p_i$ is the self-looping probability of node i. If the Hidden Markov Model has N speech nodes and we let the random variable $\gamma = \sum_{i=1}^{N} \tau_i$, it can be shown that if $P_1 \neq P_2 \neq \ldots \neq P_N$ the probability mass function of γ is the mixture distribution

$$P_\gamma(k) = \sum_{i=1}^{N} w_i P_{\tau_i}(k) \quad , \quad k = N, N + 1,\ldots, \tag{13}$$

where

$$w_i = \prod_{\substack{j=1 \\ j\neq i}}^{N} \frac{p_j(1 - p_i)}{p_j - p_i} \quad . \tag{14}$$

Since a closed form formula for $E\left[\frac{\tau_i}{\Sigma\, \tau_j}\right]$ has not been found, we use an approximation using up to the second-order moments. Let

$$\gamma_i = \sum_{\substack{j=1 \\ j\neq i}}^{N} \tau_j \quad , \tag{15}$$

$$g(\tau_i,\gamma_i) = \frac{\tau_i}{\tau_i + \gamma_i} \quad , \tag{16}$$

then τ_i and γ_i are independent and the expectation can be approximated by

$$\begin{aligned} E[g(\tau_i,\gamma_i)] &\approx g(\bar{\tau}_i,\bar{\gamma}_i) + \frac{1}{2}\left(\sigma^2_{\tau_i} \frac{\partial^2 g}{\partial \tau_i^2} + \sigma^2_{\gamma_i} \frac{\partial^2 g}{\partial \gamma_i^2}\right) \\ &= \frac{\bar{\tau}_i}{\bar{\tau}_i + \bar{\gamma}_i} + \frac{\bar{\tau}_i\sigma^2_{\gamma_i} - \bar{\gamma}_i\sigma^2_{\tau_i}}{(\bar{\tau}_i + \bar{\gamma}_i)^3} \quad , \end{aligned} \tag{17}$$

with the means and variances given by

$$\left\{\begin{aligned} \bar{\tau}_i &= \frac{1}{p_i} \quad , \\ \bar{\gamma}_i &= \sum_{\substack{j=1 \\ j\neq i}}^{N} \frac{1}{p_j} \quad , \\ \sigma^2_{\tau_i} &= \frac{(1 - p_i)}{p_i^2} \quad , \\ \sigma^2_{\gamma_i} &= \sum_{\substack{j=1 \\ j\neq i}}^{N} \frac{(1 - p_j)}{p_j^2} \quad . \end{aligned}\right. \tag{18}$$

The estimation formula (10) becomes

$$\bar{\chi}_i = \frac{1}{T} \sum_{t=1}^{T} v_{it} - \sum_{i=1}^{N} E[g(\tau_i,\gamma_i)]\ c_i \quad . \tag{19}$$

In Equation (19) the first sum is over the observed cepstral coefficient sequence, and the second sum is over the parameters of the hypothesized model. Therefore, we refer to this technique as a hypothesis-driven technique.

Step 2 (Smoothing χ_i)

After $\chi_1,\ldots,\chi_{12}$ are estimated, we fit Equation (6) to them. A least-mean-square fit requires solving the following equations:

$$a = \frac{\sum \chi_i\ e^{-b(i-1)}}{\sum e^{-2b(i-1)}} \tag{20a}$$

and

$$\frac{\sum \chi_i\ e^{-b(i-1)}}{\sum e^{-2b(i-1)}} = \frac{\sum i\ \chi_i\ e^{-b(i-1)}}{\sum i\ e^{-2b(i-1)}} \quad , \tag{20b}$$

where Equation (20b) can be solved numerically. A less computationally intensive and yet more robust fit (i.e., one which is less susceptible to the effect of outlying data) is given by fitting exponential functions to all pairs $\{\chi_i,\chi_j\}$, $i \neq j$, or a subset of these paris, and then by averaging magnitudes and time constants of the fits. We have chosen to fit the pairs that contain χ_1 and one of χ_2,χ_3,χ_4 and χ_5, namely, $\{\chi_1,\chi_j\}$, $j = 2,3,4,5$. Therefore,

$$b_j = \begin{cases} -\ln \dfrac{\chi_j}{\chi_1} \quad , & \chi_1\ \chi_j > 0 \ \text{ and } \ |\chi_1| > |\chi_j| \\ 0 \quad , & \text{otherwise} \quad . \end{cases}$$

$$a_j = \begin{cases} \chi_1 \quad , & b \neq 0 \\ 0 \quad , & \text{otherwise} \end{cases} \tag{21}$$

and a and b are the average of nonzero a_j's and b_j's.

7. SUMMARY OF THE PROCEDURE FOR CEPSTRAL COMPENSATION* AND THE EXPERIMENTAL RESULTS

Given the cepstral vectors of a test token and the Hidden Markov word model for a reference (the procedure is done for every reference word), the procedure for the adaptive cepstral compensation and recognition is described as follows:

Step 1: Compute a set of stress components [c.f. Equation (19)].

Step 2: Smooth the stress components by fitting an exponential function to them [c.f. Equations (6) and (21)].

Step 3: Subtract the values of the exponential function from the cepstral vectors of the test token.

Step 4: In recognition, perform likelihood tests using the compensated test tokens.

In Table V we summarize the recognition error rates when the hypothesis-driven stress compensation is applied to the "simulated stress" data base. For comparison the error rates of the baseline and of multimodel compensation are also included. This technique has also been applied to a more advanced 14-node, fixed-variance HMM system[16] whose parameters contain cepstral coefficients as well as differential cepstral coefficients. Because cepstral variances are fixed in this recognizer, no variance scaling is performed. The recognition results, with and without cepstral compensations, are listed in Table VI.

TABLE V

Substitution Rate (Percent): A Comparison of Multimodel Fixed Stress Compensation with Hypothesis-Driven Stress Compensation

Condition	Norm	Fast	Loud	Noise	Soft	Shout	Avg5	Avg6
Baseline HMM	1.0	6.1	29.1	19.6	13.5	86.4	13.9	25.9
Multimodel	1.0	4.2	12.1	6.7	5.5	68.7	5.9	16.4
Hypothesis-Driven	0.9	4.7	12.7	7.0	5.7	72.4	6.2	17.2

* The compensation technique is not restricted to a HMM baseline system, similar estimation formula can be derived for DTW (dynamic time warping) based recognition systems.

TABLE VI

Substitution Rate (Percent):
An Advanced HMM Recognizer

Condition	Norm	Fast	Loud	Noise	Soft	Shout	Avg5	Avg6
Without Compensation	0.4	1.7	3.4	2.9	4.4	49.8	2.5	10.4
With Compensation	0.4	1.7	3.4	1.4	2.4	45.3	1.9	9.0

8. CONFIDENCE INTERVAL ANALYSIS OF EXPERIMENTAL RESULTS

We wish to demonstrate that the reductions achieved in substitution rate are statistically significant.

Suppose that the probability of a substitution error is p; then the probability of committing k errors in a test data bse of n tokens is given by the binomial function

$$\Pr(k \text{ errors in n tokens}) = \binom{n}{k} p^k(1-p)^{n-k} \quad . \tag{22}$$

The mean and variance of the number of errors, k, are given by

$$\mu = np$$
$$\sigma^2 = np(1-p) \quad . \tag{23}$$

For large n we approximate the binomial function by the Gaussian probability density N[np,np(1 – p)]. The confidence interval of 95% is given by (μ – 1.96 σ, μ + 1.96 σ). We define the parameter X as the ratio:

$$X = \frac{1.96\ \sigma}{\mu} \quad , \tag{24}$$

so that the 95% confidence interval becomes (μ – Xμ, μ + Xμ). Substituting (23) into (24), assuming p $<<$ 1,

$$X \simeq \frac{1.96}{\sqrt{np}} = \frac{1.96}{\sqrt{\mu}} \quad . \tag{25}$$

For n = 8400, p = 13.9% (5-avg, baseline HMM) and p = 2.5% (5-avg, advanced HMM), we have X = 5.7% and X = 13.5%, respectively. The 95% confidence itnervals, corresponding to p = 13.9% and p =2.5%, are roughly (13.1%, 14.7%) and (2.16%, 2.84%), respectively. In Table V, the substitution error rate p = 6.2% (5-avg, hypothesis-driven compensation) lies well outside the interval (13.1%, 14.7%); similarly in Table VI the error rate p = 1.9% (5-avg, compensation) lies well outside the interval (2.16%, 2.84%). Hence the improvements obtained using this type of compensation are statistically significant.

9. CONCLUSION

Spectral tilt has been found to vary significantly for speech spoken in stressful talking environments. We studied the statistical variations of cepstral coefficients embedded in the framework of Markov models and found that the observed changes in cepstral mean values, from normal speech trained models to simulated-stress trained models, corresponded approximately to an exponential type of spectral tilt. A simple and efficient compensation technique, the hypothesis-driven cepstral compensation, has been formulated. Using this simple compensation technique, recognition experiments yielded significant reduction in error rate.

It is likely that further improvement may be achieved via reliable silence/voiced/unvoiced separation before the application of cepstral coefficient compensation, with compensation for spectral tilt only in voiced segments. A Bayes estimate, that incorporates and updates *a priori* knowledge of the distributions of the stress component χ_i and of the parameters a and b in the smoothing process, may also be superior to our estimate. Other improvements may be achieved through more detailed understanding and modeling of speech variations in stress.

ACKNOWLEDGMENT

The author wishes to acknowledge that the "simulated stress" data base was designed and collected by Dr. G.R. Doddington and his group in the Central Research Laboratories of Texas Instruments, Inc.; that the baseline HMM speech recognizer used in algorithm development was developed by Dr. D.B. Paul, who also conducted the experiments reported in Table II and on the first line of Table VI; and that multistyle training techniques discussed here were developed and tested by Dr. R.P. Lippmann and Dr. D.B. Paul.

This author has benefited from the work done by these individuals.

REFERENCES

1. C.E. Williams and K.N. Stevens, "Emotions and Speech: Some Acoustic Correlates," J. Acoust. Soc. Am. **52**, 1238-1250 (1972).
2. P.V. Simonov and M.V. Frolov, "Analysis of the Human Voice as a Method of Controlling Emotional State: Achievements and Goals," Aviation, Space, and Environmental Medicine **48**, 23-25 (January 1977).
3. L.A. Streeter *et al.*, "Acoustic and Perceptual Indicators of Emotional Stress," J. Acoust. Soc. Am. **73**, 1354-1360 (1983).
4. G.K. Poock, "Experimental Evidence About What Factors Influence Recognizer Performances," in Conf. Rec. of "Towards Robustness in Speech Recognition" Conference, November 2-4, Santa Barbara, California (1983).
5. E. Lombard, "Le Signe de l'Elevation de la Voix," Ann. Maladiers Oreille, Larynx, Nez, Pharynx **37** (1911).
6. H. Lane and B. Tranel, "The Lombard Sign and the Role of Hearing in Speech," J. Speech and Hearing Research **14**, 677-709 (1971).
7. D. Pisoni, R.H. Bernacki, H.C. Nusbaum, and M. Yuchtman, "Some Acoustic Phonetic Correlates of Speech Produced in Noise," Proc. of Intl Conf. on Acoustics, Speech and Signal Processing, pp. 1581-1584 (1985).
8. G.R. Doddington and T.B. Schalk, "Speech Recognition: Turning Theory to Practice," IEEE Spectrum, pp. 26-32 (September 1981).
9. J.L. Hieronymus and H.J. Enea, "Evaluating the Performance of Commercial Speech Recognizers," Proc. of Spring COMPCON'82, pp. 207-211 (1982).
10. P.K. Rajasekaran, G.R. Doddington, and J.W. Picone, "Recognition of Speech Under Stress and in Noise," Proc. of Intl Conf. on Acoustics, Speech and Signal Processing, 1986.
11. J.K. Baker, "The Dragon System — An Overview," IEEE Trans. Acoustics, Speech and Signal Processing **ASSP-23**, No. 1, 24-29 (February 1975).
12. F. Jelinek, "Continuous Speech Recognition by Statistical Methods," Proc. IEEE **64**, 532-556 (April 1976).
13. A.B. Poritz, "Linear Predictive Hidden Markov Models and the Speech Signal," Proc. Intl Conf. on Acoustics, Speech and Signal Processing, pp. 1291-1294 (1982).
14. S.E. Levinson, L.R. Rabiner, and M.M. Sondhi, "An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition," Bell Sys. Tech. J. **62**, No. 4, Pt. 1, 1035-1074 (April 1983).

15. L.R. Rabiner and B.H. Juang, "An Introduction to Hidden Markov Models," IEEE ASSP Mag. **3**, No. 1, 4-16 (January 1986).

16. D.B. Paul, "A Speaker-Stress Robust HMM Isolated Word Recognizer," submitted to the 1986 Digital Signal Processing Workshop, Chatham, Massachusetts.

17. S.B. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," IEEE Trans. Acoustics, Speech and Signal Processing **ASSP-28**, No. 4 (August 1980).

18. N. Ahmed, T. Natarajan, and K.R. Rao, "Discrete Cosine Transform," IEEE Trans. Computers (January 1974), pp. 90-93.

19. M. Hamidi and J. Pearl, "Comparison of the Cosine and Fourier Transforms of Markov-1 Signals," IEEE Trans. Acoustics, Speech and Signal Processing **ASSP-24**, 428-429 (1976).

20. L.E. Baum, T. Petrie, G. Soules, and N. Weiss, "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains," Ann. Math. Statistics **41**, 164-171 (1970).

21. G.D. Forney, Jr., "The Viterbi Algorithm," Proc. IEEE **61**, 268-278.

22. R.P. Lippmann, M.A. Mack, and D.B. Paul, "Multi-style Training for Robust Speech Recognition Under Stress," J. Acoust. Soc. Am. Supplement 1, **79**, 595 (1986).

23. R.P. Lippmann, private communication.

APPENDIX
LIST OF WORDS IN THE "SIMULATED STRESS" VOCABULARY

zero
ten
one
twenty
two
thirty
three
forty
fifty
five
sixty
six
seventy
seven
eighty
eight
ninety
nine
advise
affirmative
airspeed
air
alpha
altitude
auto
azimuth
back
bar
bravo
break
change
Charlie
combat
comm
confirm
control
cursor
degrees
delta
destination
east
echo
elevation
negative
erase
fix
freeze
fuel
go
ground
hello
help
history
hot
hundred
inventory
lock
map
mark
medium
mode
narrow
nav
north
north
no
off
oh
out
point
profile
quiet
radar
range
recall
release
repeat
return
rubout
select
sensor
south
standby
start
status
steerpoint
step
stop
synthesis
target
thousand
threat
tracker
train
voice
weapon
west
white
wide
yes

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS
BEFORE COMPLETING FORM

1. REPORT NUMBER
ESD-TR-86-098

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE *(and Subtitle)*
Cepstral Domain Talker Stress Compensation
for Robust Speech Recognition

5. TYPE OF REPORT & PERIOD COVERED
Technical Report

6. PERFORMING ORG. REPORT NUMBER
Technical Report 753

7. AUTHOR*(s)*
Yeunung Chen

8. CONTRACT OR GRANT NUMBER*(s)*
F19628-85-C-0002

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Lincoln Laboratory, MIT
P.O. Box 73
Lexington, MA 02173-0073

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
Program Element No. 62301E
Project No. 5E20
ARPA Order No. 5328

11. CONTROLLING OFFICE NAME AND ADDRESS
Defense Advanced Research Projects Agency
1400 Wilson Boulevard
Arlington, VA 22209

12. REPORT DATE
10 November 1986

13. NUMBER OF PAGES
32

14. MONITORING AGENCY NAME & ADDRESS *(if different from Controlling Office)*
Electronic Systems Division
Hanscom AFB, MA 01731

15. SECURITY CLASS. *(of this Report)*
Unclassified

15a. DECLASSIFICATION DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

None

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

robust speech recognition
talker stress compensation
hypothesis-driven stress compensation
simulated stress
HMM word recognizer
multistyle training
Hidden Markov Model
speech recognition

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

Automatic speech recognition algorithms generally rely on the assumption that for the distance measure used, intraword variabilities are smaller than interword variabilities so that appropriate separation in the measurement space is possible. As evidenced by degradation of recognition performance, the validity of such an assumption decreases from simple tasks to complex tasks, from cooperative talkers to casual talkers, and from laboratory talking environments to practical talking environments.

This report presents a study of talker-stress-induced intraword variability, and an algorithm that compensates for the systematic changes observed. The study is based on Hidden Markov Models trained by speech tokens spoken in various talking styles. The talking styles include normal speech, fast speech, loud speech, soft speech, and talking with noise injected through earphones; the styles are designed to simulate speech produced under real stressful conditions.

Cepstral coefficients are used as the parameters in the Hidden Markov Models. The stress compensation algorithm compensates for the variations in the cepstral coefficients in a hypothesis-driven manner. The functional form of the compensation is shown to correspond to the equalization of spectral tilts.

Preliminary experiments indicate that a substantial reduction in recognition error rate can be achieved with relatively little increase in computation and storage requirements.