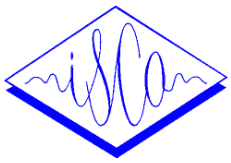

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

Lombard Effect Compensation for Robust Automatic Speech Recognition in Noise*

John H.L. Hansen and Oscar N. Bria
Department of Electrical Engineering
Duke University
Durham, North Carolina 27706

Invited Paper

1 Abstract

This paper addresses the problem of automatic speech recognition under Lombard and noise conditions. The main contributions include the statistical analysis of vocal tract and speech parameters under Lombard effect, and the formulation of a new speech recognition system which employs adaptive noise suppression and Lombard effect compensation front-end processors. The effects on formant location, bandwidth, and mel-cepstral parameters from noise and Lombard effect are presented. These parameters vary greatly, with significant variations across all phonemes for spectral tilt. Approximately half of all mel-cepstral parameters result in statistically significant variation from neutral. The significance of parameter variation between noisefree and noisy Lombard conditions shifts, suggesting the need for an alternate compensation for noisefree and noisy Lombard speech. A new recognition algorithm employing noise adaptive boundary detection, noise suppression, and voiced/unvoiced Lombard compensation is presented. Observed shifts in mean cepstral values from neutral can be modeled using an exponential tilt, as suggested by Chen [3], but that the exponential form appears to differ for each phoneme class. A new Lombard effect compensator is formulated which allows varying degrees of compensation to be placed on voiced/unvoiced speech sections. Preliminary recognition results suggest that separate compensation of voiced and unvoiced speech sections improves recognition performance by as much as 10% over no compensation.

2 Introduction

Current speech recognition systems generally degrade in performance when trained in neutral noise free conditions, and tested in noisy stressful environments. This is because previous recognition studies have largely been directed at issues which reduce speaker restrictions, increase vocabulary size, and transcend the boundaries of isolated to continuous speech. One reason for the limited progress is that past approaches such as dynamic time warping or hidden Markov modeling (HMM) have largely been applied in noise free tranquil environments. One approach to the recognition in noise problem might be to explore the enumerable speech parameterization methods reported in the literature to determine which set of parameters are best suited for an existing recognizer. Such methods cannot predict performance under varying environmental conditions. For example, noise characteristics vary greatly in settings where speech recognition is needed (e.g., pilots in aircraft cockpits, wheelchair control for the disabled, factory use for assembly lines). If robust recognition under such diverse environments is to be achieved, varying speaker and environmental conditions must be incorporated in the algorithm formulation.

The problem considered is the formulation of a robust speaker dependent, recognizer which addresses the problems of changing input noise levels, and varying speech characteristics under such noise. There are three factors which affect speech entering the recognition system. First, background noise will have a degrading effect on the speech signal. Second, since the speaker is able to hear the background noise, he may alter his speech characteristics in an effort to increase communication efficiency over the noisy medium (i.e., the Lombard effect [15]). Lastly, the performance of a secondary task may also affect characteristics of an operator's speech production system. The effects of task stress on recognition performance will not be considered at this time.

2.1 Front-End Processing for Recognition

In the past, the direction we have taken for speech recognition in noise has been to develop enhancement and stress compensation preprocessing front-ends [8, 9, 10, 11]. These preprocessors take advantage of past recognition techniques formulated in noise free tranquil environments by producing speech or recognition features less sensitive to varying factors such as stress and noise. The effect of noise and Lombard effect on recognition rates have been considered in previous studies. For example, when additive noise is introduced, average recognition rates for a discrete observation HMM decrease by 39% for neutral speech, 65% for Lombard speech [9]. Constrained iterative enhancement algorithms [10] which function as front-ends can increase recognition by +34% for neutral speech, +18% for stressed speech. These rates show that noise suppression can improve recognition to acceptable levels for neutral speech, but cannot eliminate all errors found in changing voice characteristics under the Lombard effect. If stress compensation algorithms (Lombard effect was one), based on formant location, bandwidth, and intensity, are combined with enhancement preprocessing, Lombard speech recognition increases by +42%. Generally speaking, enhancement preprocessing resulted in good performance requiring little *a priori* information. Stress compensation however, required extensive *a priori* knowledge. The stress compensators revealed that if average formant location and bandwidth are compensated, two-thirds of the errors resulting from Lombard effect can be eliminated. The next step therefore, is to perform this in a manner which is automatic. In this paper we consider compensation of mel-cepstral parameters in voiced and unvoiced sections prior to recognition. First, we summarize some statistical results of acoustic-phonetic analysis of speech under Lombard effect.

2.2 Vocal Tract and the Lombard Effect

To understand how stress effects speech production, a speech under stress data base was collected (32 speakers were employed to generate in excess of 16,000 utterances) [9]. A four year study which

Phoneme	Style	Average Formant Frequencies				Average Formant Bandwidths			
		F1	F2	F3	F4	B1	B2	B3	B4
IY	Neutral	411	1970	2607	3368	52	222	496	366
	Lombard	412	2006 ⋆	2644 ⋆	3376	73 ⋆	139 ⋆	250 ⋆	185 ⋆
I	Neutral	296	1668	2417	3329	160	147	408	442
	Lombard	361 ⋆	1710 ⋆	2473 ⋆	3220 ⋆	359 ⋆	156	279 ⋆	251 ⋆
E	Neutral	573	1260	2576	3239	136	233	788	849
	Lombard	675 ⋆	1509 ⋆	2638	3282	68 ⋆	192	474 ⋆	476 ⋆
oU	Neutral	355	955	2422	3314	225	317	623	552
	Lombard	526 ⋆	959	2344 ⋆	3236 ⋆	122 ⋆	209 ⋆	443 ⋆	287 ⋆
N	Neutral	241	1480	2515	3387	163	981	651	702
	Lombard	309 ⋆	1087 ⋆	2177 ⋆	3033 ⋆	497 ⋆	984	359 ⋆	581
R	Neutral	457	1380	1824	3197	99	467	562	318
	Lombard	447	1494 ⋆	2000 ⋆	3158	95	190 ⋆	333 ⋆	115 ⋆

Table 1: Average formant frequency and formant bandwidth for six phonemes under neutral and Lombard effect (⋆ indicates a statistically significant variation).

consisted of analysis of pitch, glottal source, duration, intensity, and vocal-tract shaping (approximately 200 speech parameters) gave insight into how talkers' vary their production systems [7, 9]. The significance of the variation of each parameter in mean, variance and distribution was considered. Of interest here, are results from Lombard effect. To illustrate how vocal tract characteristics vary under Lombard effect, we summarize average formant locations and bandwidths in Table 1. The results show that when a talker experiences Lombard effect, the following generally occur, i) average bandwidths decrease for most phonemes, ii) formant locations for vowels increase, iii) first formant locations increase for most phonemes, iv) formant amplitudes increase, giving rise to a shift in spectral energy from low to high frequency (this was especially true for sonorants). Next, we identify how speech parameters used for recognition are affected by noise.

2.3 Speech Parameters and Lombard Effect

An analysis was performed over the same speech under stress data base to determine statistical variation of linear predictive coding (LPC) parameters and mel-frequency Cepstral coefficients. Analysis of mean, variance, and distribution of the first 10 PARCOR coefficients and 9 mel-Cepstral coefficients was performed. Analysis was conducted on four data sets; i) noisefree neutral, ii) noisefree Lombard, iii) noisy neutral, and iv) noisy Lombard (Lombard effect speech was simulated by having talkers' speak with 85dB SPL noise played through headphones). Table 2 summarizes average mel-cepstral coefficients under the four conditions. Results show that approximately half of all average mel-cepstral parameters resulted in statistically significant shifts from neutral (as measured by pairwise Student T tests). When 6dB of additive white Gaussian noise is introduced, the first mel-cepstral coefficient is always significantly different from neutral. This confirms the change in spectral tilt between neutral and Lombard effect. An important point here, is that mel-cepstral parameters resulting in significant shifts under noise free settings, may not vary significantly when noise is introduced. This implies that different stress compensation must be used if no noise suppression front-end is used. Another important result is that average mel-cepstral parameters behave differently between phonemes.

3 HMM Recognition Formulation

Figure 1 illustrates a block diagram of a new HMM-based recognition system. The system employs a front-end noise adaptive word-boundary detector which provides initial boundary information to an adaptive spectral subtraction based noise suppression task. Once noise suppression has been performed, the boundary detector is applied a second time to obtain better estimates for subsequent HMM recognition. During this second application, speech activity is also partitioned into voiced/transition/unvoiced sections for

Phoneme	Style	Average mel-Cepstral Values								
		C1	C2	C3	C4	C5	C6	C7	C8	C9
IY	Neutral	25.	-8.5	-5.3	1.8	0.9	-0.8	-0.9	-0.2	-0.1
	Lombard	26.	-10. ⋆	-5.3	2.0	0.5	-0.2 ⋆	-0.1 ⋆	0.1	-0.2
I	Neutral	5.2	-12.	0.9	0.2	-1.5	1.3	-0.5	-1.2	-0.2
	Lombard	3.6	-11.	1.4	-1.6 ⋆	-2.4 ⋆	1.0	-0.4	-0.8	-0.1
E	Neutral	-2.8	-6.1	1.5	-2.0	-3.1	0.6	-0.1	-0.8	-0.2
	Lombard	-4.5	-6.0	2.7	-4.5 ⋆	-2.3	1.3 ⋆	-0.5	-0.5	0.0 ⋆
oU	Neutral	1.2	-0.5	-5.5	1.3	-0.2	-0.8	-0.9	-0.6	-0.1
	Lombard	-5.1 ⋆	-0.9	-5.4	2.4 ⋆	-1.4 ⋆	-1.2	-1.1	-0.8	0.1 ⋆
N	Neutral	17.	-2.8	-2.8	0.5	-0.0	-0.1	-0.2	-0.1	-0.1
	Lombard	12. ⋆	-5.1 ⋆	-5.0 ⋆	-1.1 ⋆	0.7 ⋆	1.2 ⋆	0.2 ⋆	-0.3 ⋆	-0.3 ⋆
R	Neutral	-4.8	-7.6	3.2	-0.9	-2.9	-3.4	-0.8	-0.3	-0.6
	Lombard	-2.4	-7.7	1.4 ⋆	-2.3 ⋆	-3.8	-2.0 ⋆	-0.4	-0.8	-0.3
		Average mel-Cepstral Values in Noise								
IY	Neutral	12.	-2.5	-2.8	-0.5	0.9	0.7	0.1	0.0	-0.1
	Lombard	10. ⋆	-3.1	-3.7 ⋆	-0.4	0.8	0.6	0.2	0.1	-0.1
I	Neutral	11.	-8.2	-0.1	0.1	-0.2	0.1	-0.7	0.0	0.1
	Lombard	6.2 ⋆	-9.0	1.6 ⋆	-0.9 ⋆	-1.2 ⋆	1.2 ⋆	-0.4 ⋆	-0.4 ⋆	0.0
E	Neutral	6.9	-5.8	1.5	-0.6	-0.9	0.1	-0.4	-0.1	-0.0
	Lombard	0.7 ⋆	-5.1	3.3 ⋆	-2.1 ⋆	-0.6	1.3 ⋆	-0.6	-0.2	0.1
oU	Neutral	7.0	-3.4	-1.3	0.9	0.4	0.0	-0.2	0.0	-0.0
	Lombard	0.9 ⋆	-3.1	-1.0	1.7 ⋆	-0.1 ⋆	-0.0	-0.0 ⋆	0.1	-0.0
N	Neutral	8.8	0.3	-0.7	-0.3	-0.1	0.1	-0.1	0.1	-0.0
	Lombard	5.5 ⋆	-0.2	-1.5 ⋆	-0.1	-0.2	0.2	-0.2	0.2	0.1 ⋆
R	Neutral	7.2	-5.5	-0.6	0.1	1.1	-0.1	-0.8	0.2	0.1
	Lombard	3.7 ⋆	-4.9	-0.7	0.3	0.6	0.0	-0.3 ⋆	0.1	0.0

Table 2: Average mel-frequency Cepstral parameters for six phonemes under neutral and Lombard effect, with and without 6dB additive White Gaussian noise (⋆ indicates a statistically significant variation).

Lombard stress compensation. The Lombard effect compensation is performed on estimated mel-frequency cepstral coefficients. Once noise suppression and stress compensation has been performed, a vector quantizer is used to compress input data prior to the HMM recognition task.

3.1 Noise Adaptive Boundary Detection

A noise adaptive boundary detector was formulated to improve performance of noise suppression and Lombard effect compensation. The detection method is similar in principle to many energy thresholding methods such as the hybrid technique proposed by Lamel, Rabiner, Rosenberg, and Wilpon [13]. The present approach is different in the sense that all thresholds (in both time and duration), adapt to varying background noise levels. Adaptive thresholds are needed, since it has been shown that word duration and intensity vary significantly under Lombard effect [9]. For example, word duration increases by 20%, and word intensity by 8% under Lombard effect. Of particular concern is that vowels and semivowels show significant increases in duration, 24% and 63% respectively. An example of boundary detector performance is shown in Figure 2. An isolated word, corrupted with additive noise is shown *(a)*. Normalized log frame energy b is used in the detection process. The process determines begin and endpoints *(c)*, and classifies them into primary and secondary depending on several a priori conditions (i.e., word duration limits, stop gap limits, etc). Once speech enhancement has been performed, the boundary detector is applied again. Primary and secondary endpoints are estimated and a subsequent voiced/transition/unvoiced detection procedure performed, *(d)*.

3.2 Noise Suppression Prefiltering

A noise suppression spectral estimator is obtained by subtracting an estimated noise bias found during non-speech activity. The noise power spectrum is determined from data outside word boundary values. If we assume the speech signal is short-time stationary, corrupted with noise which is additive and uncorrelated with speech, the resulting spectral subtraction relation is,

$$\left|\hat{\mathrm{S}}_w(j\omega)\right|^2 = |\mathrm{Y}_w(j\omega)|^2 - E[|\mathrm{D}_w(j\omega)|^2]. \qquad (1)$$

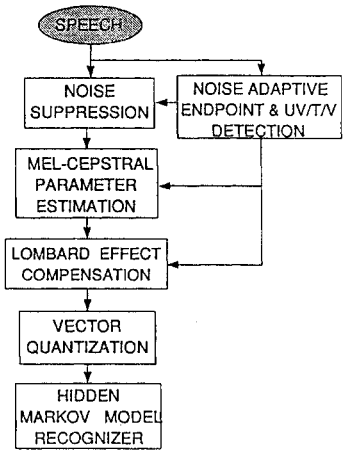

Figure 1: Block diagram of HMM-based recognition system in noise and Lombard effect.

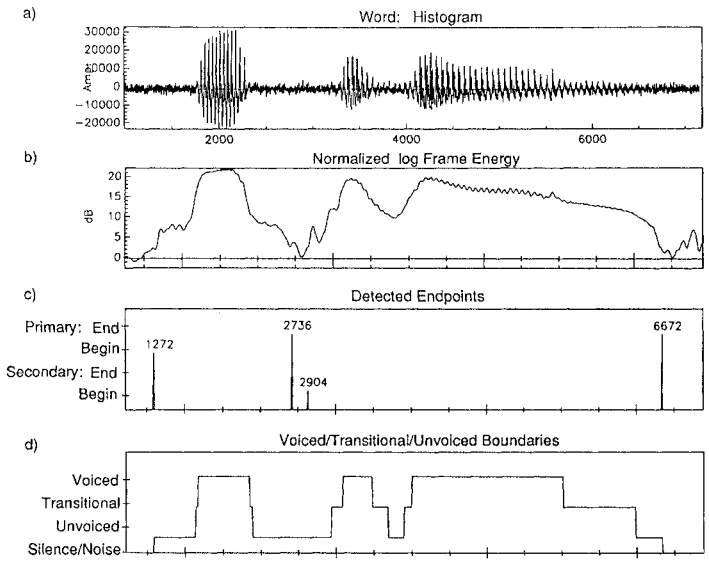

Figure 2: Example of the noise adaptive boundary detection process.

A half-wave rectification step is performed to eliminate errors in over estimating noise bias. Magnitude averaging as proposed by Boll [2] is applied to reduce residual tone artifacts. The resulting estimator, employing phase information from the original noisy speech $\Theta_y(j\omega)$ is shown below.

$$\hat{S}_{MA}(j\omega) = \left\{ \frac{1}{M_2 + M_1 + 1} \sum_{i=-M_1}^{M_2} |S_i(j\omega)|^2 - E\left[|D_w(j\omega)|^2\right] \right\}^{\frac{1}{2}} e^{j\Theta_y(j\omega)}$$

A soft decision is also made between speech and non-speech activity using the estimated boundary values from the previous task. This allows for additional noise suppression during periods of silence which improves the boundary detector's ability to distinguish voiced/transitional/unvoiced speech activity.

3.3 Parameter Estimation

The parameter representation used in this study are mel-frequency cepstral coefficients estimated from enhanced speech. The coefficients are derived in a manner similar to those used in Davis and Mermelstein [4]. Noise suppressed speech is windowed using a 256 sample Hamming window, with subsequent frames overlapping by 128 samples. Nineteen triangular bandpass filters were formed, centered at the following mel-scale frequencies, $m_i = 2595 \cdot \log_{10}\left[1 + \frac{f}{700}\right]$. The output log energy for each is obtained as $X_j, j = 1, 2, \ldots, 19$, and ten mel-cepstral parameters c_n are computed as the symmetric cosine transform of these energy values.

$$c_n = \sum_{j=1}^{19} X_j \cos\left[n\frac{\pi}{9}\left(j - \frac{1}{2}\right)\right], n = 0, 1, 2, \ldots, 9. \qquad (3)$$

3.4 mel-Cepstral Compensation

For mel-cepstral compensation, it is assumed that each coefficient is contaminated by an additive deterministic stress component. In the study by Chen [3], it is assumed that the stress effect remains unchanged within a word interval, resulting in a constant stress vector for an entire word. From our analysis on formant location, bandwidth, PARCOR coefficients, and mel-cepstral coefficients over phonemes, speech parameters vary differently over an entire word under the Lombard condition. Therefore, the present stress compensator asserts that each word will have as many stress vectors as phoneme classes within the word. The compensator consists of two phases, training and in-line recognition use. For training, a token under simulated Lombard conditions is compared to training tokens used for the HMM recognizer. With the aid of the adaptive boundary detector, all tokens are split into mel-cepstral sequences within phoneme classes. For the present study, we consider only voiced, transitional, and unvoiced phoneme classes. Once the mel-cepstral parameters are split into classes, sample averages are calculated, and Lombard effect difference vectors obtained. During recognition, Lombard effect compensation vectors for each class are used on boundary labeled test tokens. Therefore, given a test and training token for each HMM, the following steps are performed on the test token, i) compute one stress vector for each phoneme class, ii) smooth the stress vectors by fitting an exponential function to the vector as proposed by Chen [3], iii) we subtract the exponential function from the cepstral vectors of the test token across the appropriate phoneme class, finally iv) we use the compensated test tokens as the observation sequence for the HMM recognizer.

3.5 Vector Quantization

The implementation of the HMM recognizer in section 3.6 requires the model inputs to be sequences of discrete symbols chosen from a finite alphabet. These discrete symbols are obtained via vector quantization of the compensated mel-cepstral coefficients. A 64 state vector quantizer was used. Training was performed using a binary-split procedure similar to the Lloyd algorithm (see Gray [6]). The distance measure used is Euclidean based, with the first mel-cepstral coefficient used to normalize overall gain. The resulting distance equation is given below,

$$d(\vec{c}_r, \vec{c}_t) = \sqrt{\sum_{j=1}^{9} \left[\frac{c_{rj}}{c_{r_0}} - \frac{c_{tj}}{c_{t_0}}\right]^2} \qquad (4)$$

where $\vec{c}_r$ is the reference vector of mel-cepstral coefficients (e.g., from the VQ codebook), and $\vec{c}_t$ is the test coefficient vector.

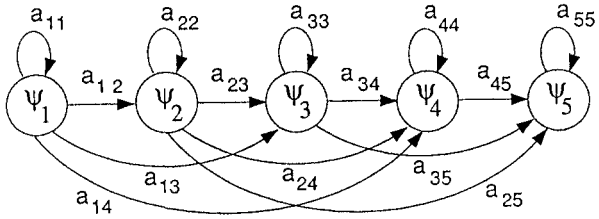

Figure 3: A state diagram for a 5-state Markov model.

3.6 HMM Recognizer

The theory of hidden Markov models and their application to speech recognition have been reported in a number of papers [12, 14, 16]. In this study, a speaker dependent, isolated word, discrete-observation hidden Markov model recognizer was formulated. Figure 3 illustrates the type of Markov model used. A separate HMM is obtained for each word. Each HMM is a five state left-to-right model, beginning in state 1. For training, each model was initiated with essentially random choices for non-zero elements and then iteratively adjusted to increase $P(\vec{\Phi}|\underline{M})$, the probability of the observation sequence $\vec{\Phi}$ having been generated by model $\underline{M}$. The training algorithm was based on the Baum-Welch forward-backward reestimation algorithm [1]. For recognition, the probability $P(\vec{\Phi}|\underline{M})$ is computed using the following relations. The recursion relation for $\alpha_{t+1}(j)$ is due to Ferguson [5].

$$P(\Phi|\underline{M}) = \sum_{i_1,i_2,\ldots,i_K} P_{i_1} b_{i_1}(\Phi_1) a_{i_1 i_2} \cdots b_{i_K}(\Phi_K) a_{i_{K-1} i_K} \quad (5)$$

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(\Phi_{t+1}) \quad t = 1, 2, \ldots, K-1. \quad (6)$$

$$P(\Phi|\underline{M}) = \sum_{j=1}^{N} \alpha_K(j). \quad (7)$$

3.7 Recognition Evaluation

Recognition evaluation was performed using a dictionary of twenty highly confusable words from our previous speech under stress data base. These words are also used by Texas Instruments and Lincoln Labs to evaluate recognition systems. Subsets include {go, oh, no}, {six, fix}, and {wide, white}. From the analysis of vocal tract and mel-cepstral parameters, it was shown that Lombard effect causes parameters to behave differently between phoneme classes. At this point, we consider different compensation between voiced and unvoiced speech sections. Fourteen examples of each word, for each speaker, were used in recognition evaluation, six neutral examples for training, six neutral examples for recognition, and two Lombard examples. All tests were fully open employing a neutral trained HMM system. Preliminary recognition results are as follows; i) 88% for noise free neutral speech, ii) 65% for noise free Lombard speech, iii) 70% with compensation for only voiced sections, iv) 70% with compensation of only unvoiced sections, and v) with compensation of voiced and unvoiced sections, recognition increased to 75%. For no compensation, voiced only, and unvoiced only compensation, 50% of errors were caused by confusable word-pairs. For combined voiced and unvoiced compensation, only 20% of the errors were due to confusable pairs. This implies that if a larger training set were used for HMM training, recognition rates may increase even further. These

4 Conclusions

This paper has investigated the effect of Lombard condition on vocal tract parameters (formant location and bandwidth) and mel-cepstral parameters. We have seen that mel-cepstral parameters vary greatly under Lombard condition, and that spectral tilt varies significantly across all phonemes. A statistical analysis of these parameters in terms of mean, variance and distribution was performed. Approximately half the mel-cepstral parameters resulted in significant variations from neutral. Also, parameters which vary significantly under noisefree Lombard condition, may not when both neutral and Lombard speech are corrupted by additive noise. Observed shifts in mean cepstral values from neutral can be modeled using an exponential tilt, as outlined by Chen [3], but that the exponential form appears to be different for each phoneme class. Finally, a new recognition algorithm employing noise adaptive boundary detection, noise suppression, and voiced/unvoiced Lombard effect compensation was presented. Preliminary recognition results suggest separate compensation of voiced/unvoiced speech sections give improved performance over no compensation.

*This work sponsored in part by grants from the National Science Foundation Grant No. IRI-9010536, and the I.B.M. Corp. Grant No. N-UN-225-00.

References

[1] L.E. Baum, T. Petrie, G. Soules, N. Weiss, " A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains," *Annals of Mathematical Statistics*, Vol. 41, pp. 164-171, 1970.

[2] S.F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Trans. on ASSP*, April 1979.

[3] Y. Chen, "Cepstral Domain Talker Stress Compensation for Robust Speech Recognition," *IEEE Trans. on ASSP*, April 1988.

[4] S.B. Davis, P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Trans. on ASSP*, pp. 357-366, August 1980.

[5] J. Ferguson, Ed., *Hidden Markov Models for Speech*, Institute For Defense Analysis: Communications Research Division, October 1980.

[6] R.M. Gray, "Vector Quantization," *IEEE ASSP Magazine*, April 1984.

[7] J.H.L. Hansen, "Evaluation of Acoustic Correlates of Speech Under Stress for Robust Speech Recognition." invited paper *Proc. Northeast Bioengineering Conference*, Boston, Mass., March 1989.

[8] J.H.L. Hansen, M.A. Clements, "Constrained Iterative Speech Enhancement with Application to Automatic Speech Recognition," *Proc. 1988 IEEE ICASSP*, New York, NY, April 1988.

[9] J.H.L. Hansen, " Analysis and Compensation of Stressed and Noisy Speech With Application to Robust Automatic Recognition," Ph.D. Thesis, Georgia Institute of Technology, July 1988.

[10] J.H.L. Hansen, M.A. Clements, " Constrained Iterative Speech Enhancement, " accepted for *IEEE Trans. on ASSP*, April 1991.

[11] J.H.L. Hansen, M.A. Clements, " Stress and Noise Compensation Algorithms for Robust Automatic Speech Recognition," *Proc. 1989 IEEE ICASSP*, Glasgow, Scotland, U.K., May 1989.

[12] F. Jelinek, "Continuous Speech Recognition by Statistical Methods," *Proceedings of the IEEE*, April 1976.

[13] L.F. Lamel, L.R. Rabiner, A.E. Rosenberg, J.G. Wilpon, "An Improved Endpoint Detector for Isolated Word Recognition," *IEEE Trans. on ASSP*, August 1981.

[14] S.E. Levinson, L.R. Rabiner, M.M. Sondhi, " An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Porcess to Automatic Speech Recognition," *Bell System Technical Journal*, April 1983.

[15] E. Lombard, " Le Signe de l'Elevation de la Voix, " *Ann. Maladies Oreille, Larynx, Nez, Pharynx*, vol. 37, pp. 101-119, 1911.

[16] A.B. Poritz, A.G. Richter, " On Hidden Markov Models in Isolated Word Recognition," *Proc. 1986 IEEE ICASSP*, Tokyo, Japan, April 1986.