Morphological Constrained Feature Enhancement with Adaptive Cepstral Compensation (MCE-ACC) for Speech Recognition in Noise and Lombard Effect

John H. L. Hansen, *Senior Member, IEEE*

Abstract— **The use of present-day speech recognition techniques in many practical applications has demonstrated the need for improved algorithm formulation under varying acoustical environments. This paper describes a low-vocabulary speech recognition algorithm that provides robust performance in noisy environments with particular emphasis on characteristics due to the Lombard effect. A neutral and stressed-based source generator framework is established to achieve improved speech parameter characterization using a morphological constrained enhancement algorithm and stressed source compensation, which is unique for each source generator across a stressed speaking class. The algorithm uses a noise-adaptive boundary detector to obtain a sequence of source generator classes, which is used to direct noise parameter enhancement and stress compensation. This allows the parameter enhancement and stress compensation schemes to adapt to changing speech generator types. A phonetic consistency rule is also employed based on input source generator partitioning. Algorithm performance evaluation is demonstrated for noise-free and nine noisy Lombard speech conditions that include additive white Gaussian noise, slowly varying computer fan noise, and aircraft cockpit noise. System performance is compared with a traditional discrete-observation recognizer with no embellishments. Recognition rates are shown to increase from an average 36.7% for a baseline recognizer to 74.7% for the new algorithm (a 38% improvement). The new algorithm is also shown to be more consistent, as demonstrated by a decrease in standard deviation of recognition from 21.1 to 11.9 and a reduction in confusable word-pairs under noisy, Lombard-effect stressed speaking conditions.**

I. Introduction

AN IMPORTANT problem that has become increasingly evident as speech recognition technology matures is the ability of recognition algorithms to perform reliably under diverse, noisy, stressful conditions. One application for recognition is in military aircraft and helicopter cockpits. Studies have shown that recognition accuracy is severely reduced when speech is uttered in aircraft cockpit environments [34], [45]. Since the majority of past recognition algorithms assume noise-free, tranquil environments, recognition rates have been observed to decrease by as much as 60% under noisy, stressful conditions [12]. Factors that affect such speech entering a recognizer include additive background noise, Lombard effect, and task stress. Since the speaker is able to hear background noise, he will alter his speech characteristics in an effort to increase communication efficiency over the noisy medium (which is known as the Lombard effect [33]). Therefore, reliable recognition in such environments requires more than simply canceling additive acoustic noise.

In earlier studies, Hansen and Clements [20], [12], [13] considered an analysis of vocal tract and speech parameters under stressful conditions, including the Lombard effect. These results show that when a talker experiences the Lombard effect, the following 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, producing increased spectral tilt (especially true for sonorants).

These findings were supported in an independent study by Stanton *et al.* [47]. In [21], it was demonstrated that speech enhancement preprocessing could improve recognition rates of a traditional isolated-word speech recognizer. However, such processing does not address changing speech production effects brought on by the Lombard effect. In [22], a speech recognition algorithm was introduced that incorporates stress compensation and iterative speech enhancement steps for robust recognition. Compensation was performed on formant location and bandwidth over labeled phonemes. Recognition rates increased by 42% for noisy Lombard conditions, demonstrating the usefulness of such vocal tract perturbations during recognition. Since labeling phonemes in noisy recognition scenarios is impractical, alternate algorithms are needed. Other approaches to speech recognition in stress or noise include multistyle training [31], nonlinear spectral subtraction [32], neural network-based stress equalization [8], alternate distance metrics for recognition in noise [35], and others [24], [43], [40], [48], [1], [25], [6], [26].

In another study, Chen [7] proposed a method where each mel-cepstral recognition parameter is assumed to be contaminated by an additive deterministic component, resulting in a constant stress vector for an entire word. Earlier analysis of speech under stress, however, suggests that vocal-tract variation due to the Lombard effect is not uniform over an entire utterance. Significant duration variation also suggests nonuniform spectral variation of isolated words [12]. In studies by Hansen and Bria [17], [4], [18], it was shown that mel-cepstral parameters vary differently over an entire word under Lombard

Manuscript received October 2, 1992; revised April 13, 1994. This work was supported by the National Science Foundation Grant no. NSF-IRI-90-10536.

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

IEEE Log Number 9403970.

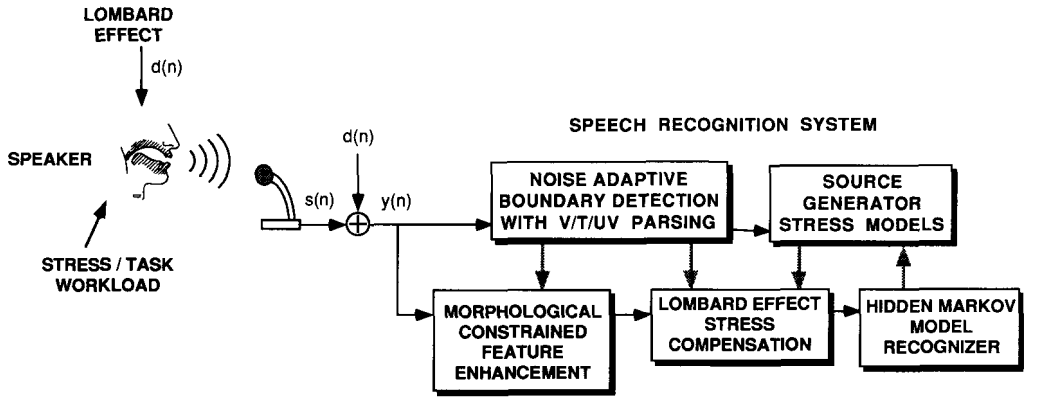


Fig. 1. General framework for the introduction of noise and the Lombard effect and processing employed by the MCE-ACC-HMM speech recognition algorithm.

condition. Initial results from an improved compensated mel-cepstral-based recognizer demonstrated improvement in recognition employing a single voiced and unvoiced compensator across an utterance. In this paper, an algorithm for adaptive cepstral compensation with morphological-based constrained enhancement is formulated for isolated-word recognition in noise and the Lombard effect. The general framework for introducing noise and Lombard effect/stress and the basic processing for robust speech recognition is shown in Fig. 1. The recognition approach is based on a source generator framework for stress modeling [16]. In Section II, we discuss the database used for these studies and illustrate the effects of stress and noise on recognition. Section III briefly considers the effects of noise and the Lombard effect on speech recognition parameters, which motivates the proposed algorithm. Section IV presents the algorithm formulation, followed by algorithm performance in Section V.

II. Speech Recognition in Noise and the Lombard Effect

A. SUSAS Stressed Database

The studies conducted in this research were based on data previously collected for analysis and algorithm formulation of speech recognition in noise and stress. This database is called speech under simulated and actual stress (SUSAS) and has been employed extensively in the study of how speech production and recognition varies when speaking during stressed conditions [20], [12], [13], [15], [18]. SUSAS consists of the following five stress domains:

i) psychiatric analysis data (speech under depression, fear, anxiety)
ii) talking styles[1] (slow, fast, soft, loud, angry, clear, question)
iii) computer tracking task or speech produced in noise (Lombard effect)
iv) dual tracking computer task
v) subject motion-fear tasks (*G*-force, Lombard effect, noise, fear).

[1] Approximately half of the SUSAS database consists of style data donated by Lincoln Laboratories [7], [31], [40].

The database offers a unique advantage for analysis and design of speech processing algorithms in that both *simulated* and *actual* stressed speech are available. A common vocabulary set of 35 aircraft communication words make up over 95% of the database. These words consist of mono and multisyllabic words that are highly confusable. Examples include /go-oh-no/, /wide-white/, and /six-fix/. A more complete discussion of SUSAS can be found in the literature.

The subset of data for this study consists of neutral training and test data and speech under the Lombard effect. Speech data under the Lombard effect was produced by having speakers listen to 85 dB SPL pink noise binaurally while uttering test tokens (i.e., all tokens are noise free). Data used in this study consist of three adult male speakers, all sampled at 8 kHz using a 16-bit A/D converter.

B. Effects of Stress and Noise on Recognition

It is known that talkers vary their speech characteristics when speaking in a noisy environment. For example, overall speech level as a function of external noise level has been shown to rise at the rate of 0.3 dB/dB noise to 1.0 dB/dB noise, depending on noise level and the specific task assigned to the speaker [11], [41]. Speakers also tend to vary those factors related to speech clarity when presented with external noise. It has also been shown that auditory fatigue consisting of temporary modifications in hearing can be caused by prolonged exposure to noise [27], [46]. Studies show that even a slight and transient auditory fatigue gives rise to a clear reduction in intelligibility and the rate of correct lexical decision making when speech is transmitted at low levels with masking noise. It is also known that recognition algorithms formulated for noise-free tranquil environments perform poorly when operating in noise. It is suggested that auditory fatigue reduces the ability of the auditory system to properly process speech in noisy environments; whereas it is

TABLE I
RECOGNITION PERFORMANCE OF STRESSFUL SPEECH, NOISY STRESSFUL SPEECH, AND NOISY STRESSFUL SPEECH USING A TANDEM CONSTRAINED ITERATIVE ENHANCEMENT AND STRESS COMPENSATION PREPROCESSING

	STRESSFUL SPEECH RECOGNITION RESULTS †												
Condition	*N*	*Sl*	*F*	*So*	*L*	*A*	*C*	*Q*	*C50*	*C70*	*Lom*	*Avg10*	*StDev10*
Stressful, Noise-free	88%	60%	65%	48%	50%	20%	68%	75%	63%	63%	63%	57.5%	15.35
Stressful, Noisy	49%	45%	28%	33%	18%	15%	40%	28%	35%	33%	28%	30.3%	9.12
FF-LSP:T,Auto:I plus FL/FB/FL+FB	83%	61%	53%	53%	61%	50%	58%	56%	55%	55%	70%	57.2%	5.69

Speaking Styles Key:
N – neutral . F – fast L – loud C – clear C50 – Moderate Task Condition Lom – Lombard effect
Sl – slow So – soft A – angry Q – question C70 – High Task Condition noise condition

suggested that speech recognition systems fail in noise because they are either unable to overcome the statistical variation of speech parameters in noise or unable to extract only those features that reflect noise-free speech production.

To illustrate the effects of stress and noise on recognition performance, a baseline hidden Markov model recognizer (VQ-HMM) was tested using data from the SUSAS database. Recognition[2] rates for 11 stressed speaking styles are shown in Table I. The baseline VQ-HMM recognizer is described later in Section V. Under noise-free stressed speaking conditions, recognition rates decrease by an average of 31% (i.e., from 88% for neutral to AVG10 =57.5% for stressed). When 30-dB additive white Gaussian noise is introduced, the average recognition rate decreases by 58% (i.e., from 88% for noise-free neutral to AVG10 =30.3%). Recognition performance also varies considerably across noise-free and noisy stressed speaking conditions as reflected in the large standard deviation in rate of recognition (STDEV10 =15.35, 9.12 for noise free and noisy stressed conditions). Recognition performance therefore seriously degrades in the presence of noise and/or stress.

III. PARAMETER ANALYSIS IN NOISE AND LOMBARD EFFECT

Let $\vec{s}$ be a sample vector of noise-free neutral speech in a sample space Υ_s. Let the sample space Υ_s consist of J independent and mutually exclusive random speech type sources,

$$\vec{s} \in \Upsilon_s : \{\gamma_j;\ j = 1, 2, \ldots, J\}. \quad (1)$$

Here, the collection of generators $\vec{\gamma}$ span the entire source generator space and could represent isolated phonemes, diphone pairs, or a temporal partition of detected speech sections. Let $\vec{y}$ be a sample vector from some source generator γ_j, which is corrupted by an additive noise vector $\vec{d}$

$$\vec{y}_{\gamma_j} = \vec{s}_{\gamma_j} + \vec{d}. \quad (2)$$

Here, the effect of additive noise on characteristics that contribute to speech quality or intelligibility will depend on the specific source generator γ_j. Next, we consider the effects of stress on speech production for the observation vector $\vec{s}$. It is known that the presence of stress will cause changes in phoneme production with respect to intensity, duration, and spectral shape. Let this change be represented by a change in the speech source generator from γ_j to $\Psi[\gamma_j]_i$, corresponding to the ith-type stress generator class for speech type γ_j[3]

$$\vec{s} \in \Upsilon_s : \{\Psi[\gamma_j]_i;\ j = 1, 2, \ldots, J;\ i = 1, 2, \ldots, I\} \quad (3)$$

where $j = 1, \ldots, J$ spans the number of possible source generators, and $i = 1, \ldots, I$ spans the domain of stressed speech classes. The resulting noise corrupted stressed speech vector is

$$\vec{y}_{\Psi[\gamma_j]_i} = \vec{s}_{\Psi[\gamma_j]_i} + \vec{d} \quad (4)$$

where the level and type of noise $\vec{d}$ will effect specific variation in speech production under a given stress condition.

Next, consider a speech parameterization of each vector $\vec{s}_{\gamma_j}$, $\vec{y}_{\gamma_j}$, $\vec{s}_{\Psi[\gamma_j]_i}$, and $\vec{y}_{\Psi[\gamma_j]_i}$, corresponding to noise-free neutral, noisy neutral, noise-free stressful, and noisy stressful speech. At this point, we consider a statistical parameter analysis of these vectors to determine how noise and/or stress influences the set of speech source generators. The first 10 mel-cepstral parameters c_k, $k = 0, 1, \ldots, 9$ were estimated in a manner similar to those in [9]. Each speech vector $\vec{s}_{\gamma_j}$ is obtained using a 32–ms Hamming window with subsequent vector frames overlapping by 16 ms. Nineteen triangular band-pass filters are formed, centered at the following mel-scale frequencies: $m_i = 2595 \cdot \log_{10}\ [1 + \frac{f}{700}]$. The output log energy for each is obtained as X_j, $j = 1, 2, \ldots, 19$, and 10 mel-cepstral parameters c_k are computed as the symmetric cosine transform of these energy values

$$c_k = \sum_{j=1}^{19} X_j \cos\left[k\frac{\pi}{19}\left(j - \frac{1}{2}\right)\right],\ k = 0, 1, 2, \ldots, 9. \quad (5)$$

A representative set of source generators were selected for statistical parameter analysis. Although the notion of stressed induced source generators is general, the present study will focus on stress associated with the Lombard speaking condition. Fig. 2 illustrates how the Lombard effect causes changes in spectral content, duration, and intensity on recognition parameters for the word "degree." The figure shows time evolution of the relative magnitudes of mel-cepstral parameters

[2]The term 'recognition' rate is used in this context with the understanding that no rejection or deletion was allowed by any of the recognizers. Other studies may refer to this as a 'substitution' rate.

[3]Here, the 11 stressed speaking styles from Table I are considered to represent a finite set of stressed speech source generator classes.

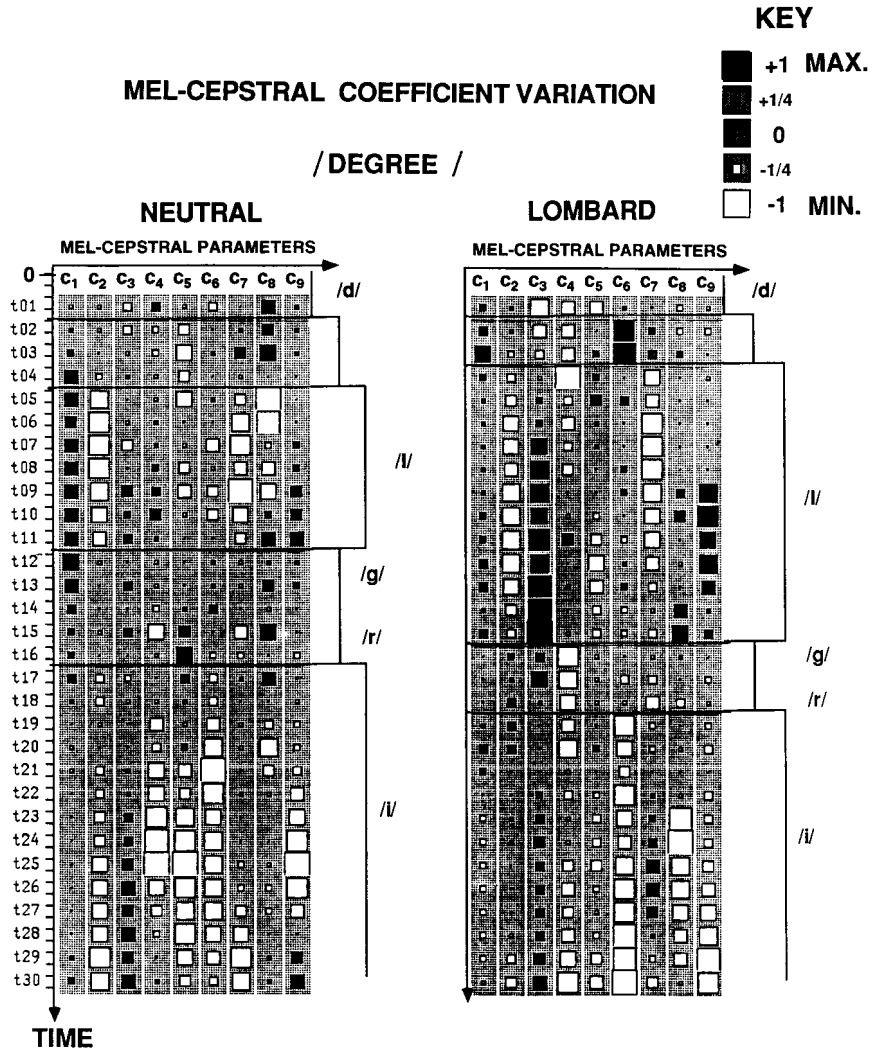


Fig. 2. Time evolution of normalized mel-cepstral coefficient variation for the word "degree" under noise-free neutral and Lombard effect speaking conditions. The key reflects the relative change in mel-cepstral parameter across time, where +1 and -1 represent the maximum positive and negative parameter values across time.

for each speaking style.[4] It is clear that individual phoneme duration, as well as changes in the spectral magnitude, vary considerably across the isolated word. These results are further supported by earlier statistical studies on the variation of speech production under stressed speaking conditions [12]. This suggests that a given stress class, $\Psi[\cdot]_i$ will nonuniformly influence the sequence of speech-type source generators $\gamma_j(t)$ needed to produce an isolated word.

[4]The magnitude of each mel-cepstral parameter has been normalized with respect to the maximum and minimum value of coefficient c_k across the utterance.

To verify this, a statistical parameter analysis was performed on mel-cepstral parameters for long duration phonemes (i.e., vowels, nasals, fricatives, etc.). The analysis was conducted on four data sets:

i) noise-free neutral $\vec{s}_{\gamma_j}$
ii) noise-free Lombard $\vec{s}_{\Psi[\gamma_j]_i}$
iii) noisy neutral $\vec{y}_{\gamma_j}$
iv) noisy Lombard $\vec{y}_{\Psi[\gamma_j]_i}$.

Additive white Gaussian noise at a SNR of 6 dB is introduced into both neutral and Lombard effect speech to obtain noisy vectors. All isolated phoneme speech types were hand labeled for statistical data analysis. Vectors of mel-cepstral parameters

TABLE II
STATISTICAL CHARACTERIZATION OF THREE SPEECH SOURCE GENERATORS γ_j BETWEEN NEUTRAL AND LOMBARD EFFECT STRESS CLASSES IN NOISE-FREE AND NOISY CONDITIONS (6 dB ADDITIVE WHITE GAUSSIAN NOISE). SOURCE GENERATORS ARE PARAMETERIZED USING MEL-CEPSTRAL PARAMETERS. THE ESTIMATED MEAN AND VARIANCE RATIOS, STUDENT'S T TEST, F TEST, AND KOLMOGOROV-SMIRNOV TESTS ARE SHOWN. ◇ INDICATES A STATISTICALLY SIGNIFICANT VARIATION FROM NOISE-FREE OR NOISY NEUTRAL.

STATISTICAL CHARACTERIZATION OF NOISE FREE SOURCE GENERATORS

/IY/						/N/						/OU/					
C_i	ρ_m	ρ_{σ^2}	T	F	KS	C_i	ρ_m	ρ_{σ^2}	T	F	KS	C_i	ρ_m	ρ_{σ^2}	T	F	KS
C_1	1.02	0.71				C_1	0.72	1.08	◇		◇	C_1	-4.29	0.53	◇	◇	◇
C_2	1.56	0.42	◇	◇	◇	C_2	1.82	0.47	◇	◇	◇	C_2	1.73	0.98			
C_3	1.00	0.78				C_3	1.74	1.01	◇		◇	C_3	0.99	0.81			
C_4	1.09	0.78				C_4	-2.09	1.94	◇	◇	◇	C_4	1.78	1.51	◇		◇
C_5	0.61	1.85		◇		C_5	-20.5	0.96	◇		◇	C_5	6.44	1.31	◇		◇
C_6	0.29	0.90	◇		◇	C_6	-10.7	1.48	◇		◇	C_6	-1.69	1.72		◇	◇
C_7	0.08	0.49	◇	◇	◇	C_7	-1.09	1.70	◇			C_7	-0.77	1.09			
C_8	0.34	0.86				C_8	5.20	1.70	◇		◇	C_8	-1.10	1.91		◇	◇
C_9	1.46	0.45		◇		C_9	1.93	0.42	◇	◇	◇	C_9	-2.57	3.16	◇	◇	◇

STATISTICAL CHARACTERIZATION OF NOISE CORRUPTED SOURCE GENERATORS

/IY/						/N/						/OU/					
C_i	ρ_m	ρ_{σ^2}	T	F	KS	C_i	ρ_m	ρ_{σ^2}	T	F	KS	C_i	ρ_m	ρ_{σ^2}	T	F	KS
C_1	0.84	1.51	◇		◇	C_1	0.62	0.51	◇	◇	◇	C_1	0.13	1.78	◇	◇	◇
C_2	1.26	0.76				C_2	-0.73	1.44				C_2	0.92	0.79			
C_3	1.32	0.87	◇			C_3	2.16	0.90			◇	C_3	0.74	1.17			◇
C_4	0.89	0.88				C_4	0.36	0.97				C_4	1.89	0.99	◇		◇
C_5	0.92	1.09				C_5	1.69	1.19				C_5	-0.14	1.58	◇	◇	◇
C_6	0.95	0.57		◇		C_6	2.76	0.92				C_6	-0.95	0.63		◇	
C_7	1.58	1.13				C_7	-2.81	1.09				C_7	0.15	1.07	◇		
C_8	3.71	1.10				C_8	1.54	0.84				C_8	1.68	0.99			
C_9	0.82	0.66				C_9	-5.75	0.77	◇			C_9	0.07	1.00			

were extracted over labeled sections from the the SUSAS database corresponding to individual source generators. The mean $m_{i,j,k}$, variance $\sigma^2_{i,j,k}$, and distribution $f_{x(i,j)}(x_{i,j,k})$ of each source generator were estimated corresponding to stress class i, speech generator j, and model parameter k. The ratio of source generator means and variances between neutral and Lombard effect speaking style in both noise-free and noisy conditions were found for each mel-cepstral parameter c_k

$$\rho_m(j,k) = \frac{E[C_k \mid \text{stress class} = \text{Lombard}, \gamma_j]}{E[C_k \mid \text{stress class} = \text{Neutral}, \gamma_j]} = \frac{\hat{m}_{C_k}(i = \text{Lom}, j)}{\hat{m}_{C_k}(i = \text{Neu}, j)} \tag{6}$$

$$\rho_{\sigma^2}(j,k) = \frac{\text{VAR}[C_k \mid \text{stress class} = \text{Lombard}, \gamma_j]}{\text{VAR}[C_k \mid \text{stress class} = \text{Neutral}, \gamma_j]} = \frac{\hat{\sigma}^2_{C_k}(i = \text{Lom}, j)}{\hat{\sigma}^2_{C_k}(i = \text{Neu}, j)}. \tag{7}$$

In (6), $\hat{m}_{C_k}$ $(i = \text{Lom}, j)$ corresponds to the estimated mean of mel-cepstral parameter k for speech source j under the Lombard effect, and $\hat{m}_{C_k}$ $(i = \text{Neu}, j)$ corresponds to estimated mean for neutral speaking conditions. Here, unknown parameters are assumed to be Gaussian distributed. The estimated variance of mel-cepstral parameter k for speech source j under neutral and Lombard styles is $\sigma^2_{C_k}(i = \text{Neu}, j)$ and $\sigma^2_{C_k}(i = \text{Lom}, j)$.

Table II summarizes ratios for three source generators in noise-free and noisy conditions. The Student's T test, F test, and Kolmogorov–Smirnov tests were used to analyze changes in mean, variance, and distribution. In the table, ◇ is used to indicate that a statistically significant change was observed between characteristics of the two source generators (a 95% confidence interval was used). For noise-free Lombard speech, approximately half of all average mel-cepstral parameters resulted in statistically significant shifts from neutral. However, the major stress-induced variations for vowels occurred in the moderate to most rapidly varying spectral components, whereas for liquids, glides, and diphthongs, they occur in the most slowly to moderately varying spectral components. Parameter variance generally decreases under the Lombard effect for vowels, remains unchanged for slowly varying spectral components in nasals, and increases for rapidly varying spectral components in nasals, diphthongs, and glides.

When 6 dB of additive white Gaussian noise is introduced, the first mel-cepstral coefficient is always significantly different from neutral. This shows that when speech is produced under the Lombard effect in noise-free or noisy conditions, a change occurs in the spectral tilt. However, when speech under the Lombard effect is corrupted with broad-band additive noise, the variation in spectral tilt is found to be present across more phoneme classes than for noise-free conditions. This occurs since high levels of broad-band background noise can have a diminishing effect on the changes in spectral tilt for low-energy phonemes. Some source generator types such as nasals displayed less of a change in stress-induced spectral variation when noise is added. Other types, such as glides and liquids, were more consistent in their statistically

significant variation between noise-free and noisy conditions. These results suggest that the stress-induced Lombard effect causes spectral content of speech source generators to vary differently across phonemes. Therefore, compensation of speech modeling parameters for recognition cannot be fully characterized across isolated words by a fixed vector of means.

This result is further supported by previous studies using a tandem constrained iterative speech-enhancement algorithm with stress compensation based on formant location and/or bandwidth[5] [19], [12], [22]. Table I shows results using this tandem processor to enhance speech and compensate for stress over the ten speaking styles of the SUSAS database. Enhancement and compensation of average formant location and/or bandwidth increases recognition by +26.9% with a decrease in recognition variance from 9.12 to 5.69, indicating a more consistent level of performance over noisy stressed speaking styles. The algorithm proposed in the next section, however, does not require *a priori* knowledge of phoneme boundaries, is more computational efficient, and achieves higher and more consistent levels of recognition over noisy Lombard effect conditions.

IV. MCE-ACC-HMM Algorithm Formulation

Employing the stressed speech source generator framework from Section III, a new recognition algorithm is proposed that performs improved speech parameterization using morphological constrained enhancement and Lombard effect compensation across an estimated source generator sequence. Fig. 3 illustrates a block diagram of the new algorithm entitled morphological constrained feature enhancement with adaptive mel-cepstral compensation based hidden Markov model recognition (MCE-ACC-HMM). The algorithm uses a noise-adaptive boundary detector and voiced/transitional/unvoiced classifier to partition input speech into sequences of source generator vectors. Nonlinear frequency domain feature enhancement employing morphological operator theory is used to suppress additive noise distortion in the source generator observation sequence. Next, Lombard effect stress compensation is performed for each detected generator across the input utterance. Using a statistical model of generator duration for each word model, a phonetic consistency rule is applied that partitions utterances into single and multisyllabic classes prior to hidden Markov model recognition. Each processing step is discussed in the following subsections.

A. Noise Adaptive Boundary Detection

An adaptive boundary detector proposed in [14] is used to provide the necessary source generator information to subsequent processing sections. The detection method is similar to many energy thresholding methods such as the hybrid technique proposed in [28] but differs in that thresholds are adapted based on background noise levels. The process begins by obtaining a sequence of frame energy measurements $e(i)$ over analysis windows of 87.5 ms every 1 ms. A sequence of potential begin/end points $(p_{b_1}, p_{b_2}, \ldots)$ and $(p_{e_1}, p_{e_2}, \ldots)$ are detected using frame-to-frame energy $e(i)$ amplitude, curvature, and duration. Seven noise adaptive thresholds $(\alpha_1, \ldots, \alpha_7)$ are defined as follows: α_1, α_2, and α_3 are used for rise and fall-time begin/end detection, α_4 is the peak frame energy that must exist for detection, and α_5 is the maximum frame energy for isolated words in a given background noise. The last two thresholds (α_6 and α_7) are used to distinguish unvoiced, transitional, and voiced speech frames. Each threshold adapts based on peak and RMS signal energy across isolated words. Duration and energy curvature rules are employed to obtain the sequence of possible begin/end point pairs. Adaptive thresholds are needed since it has been shown (see [12]) that word duration and intensity vary significantly when speech is spoken in Lombard effect and noise. For example, word duration increases by 20% and word intensity by 8% under the Lombard effect. Of particular concern is that vowels and semivowels show significant increases in duration of 24 and 63% respectively. Next, the (p_{b_i}, p_{e_j}) begin-end point sequence is examined pairwise to determine the likelihood of syllable count within the word. The pair sequence is also rank ordered into primary and secondary boundary pairs. For recognition studies, preference was always given to higher duration boundary pairs during training and testing.

Since the Lombard effect and background noise influence speech production differently across phonemes, a voiced/transitional/unvoiced $(v/t/uv)$ detection procedure is also performed. Although it is desirable to obtain true subword unit partioning (demisyllable, diphone, phoneme, etc.), this is difficult to achieve reliably in noisy environments. Instead, a $v/t/uv$ detection approach, which was previously shown to be successful for constrained speech enhancement [14], is used. The $v/t/uv$ detection produces a sequence of phoneme-like source generator boundaries $\vec{b}_{\text{word}_i} = (b_1, b_2, \ldots, b_L)$, which are used in subsequent processing for noise and Lombard effect. An example of begin/end and $v/t/uv$ boundary detection is shown in Fig. 4.

B. Feature Enhancement

A well-known speech enhancement technique originally developed by Boll [3] solves for an estimated speech spectrum by subtracting a spectral noise bias obtained during nonspeech activity. Although this technique was shown to increase overall speech quality, unnatural sounding artifacts result. Magnitude averaging can reduce errors in noise bias estimation, although some "musical tone" artifacts persist. McAulay and Malpass [37] later formulated a procedure using a soft-decision noise suppression filter. The idea was to subtract a larger noise bias if the probability of speech activity was low [2], thus reducing tone artifacts during silent periods between words. Here, a spectral subtraction-based algorithm is formulated that employs morphological-based spectral constraints for the purpose of recognition parameter enhancement in noise. These constraints are applied based on temporal information provided by a noise-adaptive endpoint detector, thereby adapting the

[5] The front-end processor uses iterative speech enhancement with interframe constraints applied to the line-spectral-pair (LSP) parameters and intraframe constraints on the autocorrelation lags, followed by various forms of formant bandwidth compensation (FF-LSP:T, Auto:I, plus FL/FB/FL+FB).

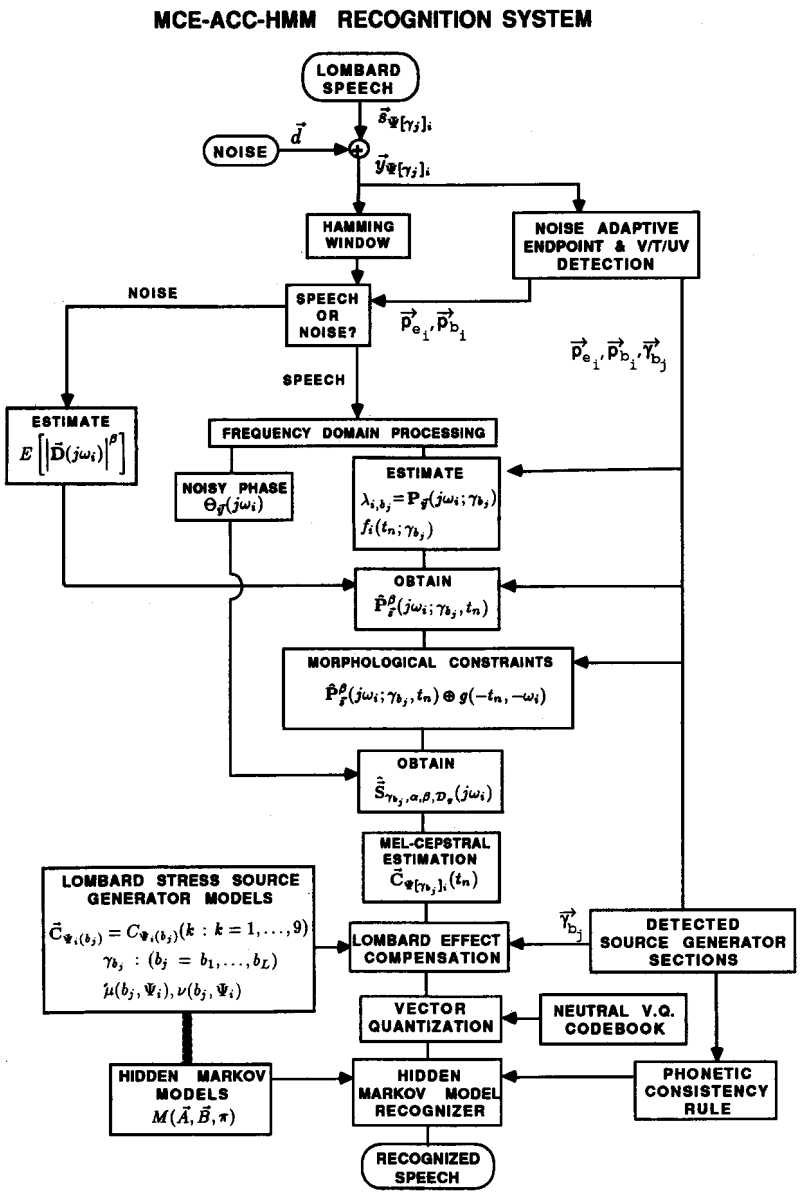


Fig. 3. Flow diagram of the MCE-ACC-HMM speech recognition algorithm in noise and Lombard effect.

enhancement procedure as speech characteristics change on a frame-by-frame basis.

1) Frequency Domain Processing: Consider a zero mean random process $s(t)$ that is degraded by additive, uncorrelated random noise $d(t)$. It is assumed that $d(t)$ can be characterized during nonspeech activity. To ensure short-time stationarity, a Hamming window is applied over $0 \leq t_n \leq T$ to obtain a sequence of sample speech vectors. It is assumed that the random process $s(t)$ is a sample function from a known or detected generator section b_j as determined in Section IV-A, resulting in the following degraded speech signal:

$$\vec{y}_{\gamma_{b_j}}(t_n) = \vec{s}_{\gamma_{b_j}}(t_n) + \vec{d}(t_n). \tag{8}$$

Such an interpretation is made since the addition of broad-band noise $\vec{d}(t_n)$ will effect each speech generator differently based on its perceptual importance across a given isolated word. Therefore, the input utterance will consist of a sequence of

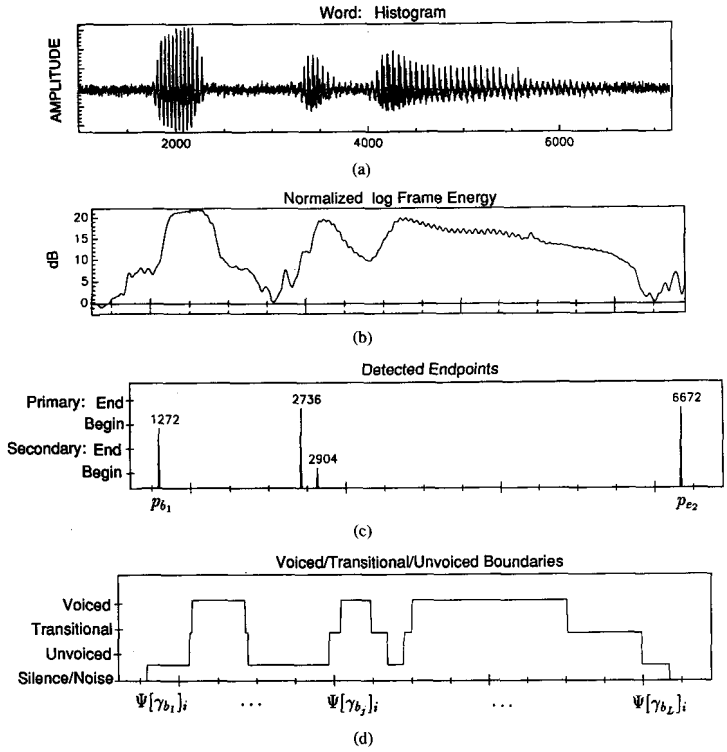


Fig. 4. Example of the noise adaptive boundary detection process: (a) Isolated word, corrupted with additive noise; (b) normalized log frame energies $\vec{e}_i$:$(i = 1, \ldots, K)$; (c) detected series of begin and endpoints $(p_{b_i}, p_{e_i}, \ldots)$, classified as primary and secondary; (d) results from voiced/transitional/unvoiced boundary detection.

noisy sample vectors, grouped into source generator sections

$$\{\vec{y}(t_n : n = 1, \ldots, N)\} = \left\{\vec{y}_{\gamma_{b_1}}(t_{b_1} = 1, \ldots, N_{b_1}), \ldots \vec{y}_{\gamma_{b_L}}(t_{b_L} = 1, \ldots, N_{b_L})\right\}, \quad (9)$$

where $N = \sum_{j=1}^{L} N_{b_j}$ and where N_{b_j} corresponds to the frame count for source generator b_j. If the autocorrelation function of $\vec{y}_{\gamma_{b_j}}(t_n)$ is $\Phi_{yy}(t_n, \tau_n;\ \gamma_{b_j})$, then $\vec{y}_{\gamma_{b_j}}(t_n)$ can be expanded into a set of orthonormal functions $f_i(t_n;\ \gamma_{b_j})$ on the interval $(0, T)$,

$$\vec{y}_{\gamma_{b_j}}(t_n) = \lim_{I \to \infty} \sum_{i=1}^{I} y_{i,b_j} f_i(t_n; \gamma_{b_j}) \quad (10)$$

$$\int_0^T f_k(t_n; \gamma_{b_j}) f_i^*(t_n; \gamma_{b_j}) dt_n = \delta_{ki,b_j}. \quad (11)$$

Here, y_{i,b_j} are the coefficients in the expansion for each detected speech generator section b_j. Since $f_i(t_n;\ \gamma_{b_j})$ are orthonormal, the coefficients may be expressed as

$$y_{i,b_j} = \int_0^T y_{\gamma_{b_j}}(t_n) f_i^*(t_n, \gamma_{b_j}) dt_n. \quad (12)$$

If there is the additional condition that the coefficients y_{i,b_j} be mutually uncorrelated, then the orthonormal functions $f_i(t_n; \gamma_{b_j})$ are the eigenfunctions of the integral equation

$$\int_0^T \Phi_{yy}(t_n, \tau_n; \gamma_{b_j}) f_i(\tau_n; \gamma_{b_j}) d\tau_n = \lambda_{i,b_j} f_i(t; \gamma_{b_j}) \quad (13)$$

where λ_{i,b_j} is the eigenvalue corresponding to the eigenfunction $f_i(t; \gamma_{b_j})$ for the detected generator type b_j. The eigenvalues and eigenfunctions can then be written as

$$\lambda_{i,b_j} = \int_0^T \Phi_{yy}(\tau_n; \gamma_{b_j}) e^{-j\frac{2\pi i \tau_n}{T}} d\tau_n = \mathbf{P}_{\vec{y}}(j\omega_i; \gamma_{b_j}) \quad (14)$$

$$f_i(t_n; \gamma_{b_j}) = \frac{1}{\sqrt{T}} e^{j\frac{2\pi i t_n}{T}}. \quad (15)$$

It can be seen that the eigenvalue λ_{i,b_j} is simply a discrete sample of the power spectrum $\mathbf{P}_{\vec{y}}(j\omega_i; \gamma_{b_j})$ at frequency $\omega_i = \frac{2\pi i}{T}$ for the detected generator type b_j. Since the human auditory system is relatively insensitive to phase distortion, a reasonable approach for estimating the random signal $\vec{s}_{\gamma_{b_j}}(t_n)$ is to estimate the magnitude of its spectral component, incorporate the noisy phase from $\vec{y}_{\gamma_{b_j}}(t_n)$, and perform an inverse transform. The Fourier transform of the vector $\vec{s}_{\gamma_{b_j}}(t_n)$ can then be represented as $\vec{\mathbf{S}}_{\gamma_{b_j}}(j\omega_i)$. With further expansion, it is noted that two cross estimation terms

result: $E[\vec{\mathbf{S}}_{\gamma_{b_j}}(j\omega_i)\vec{\mathbf{D}}^*(j\omega_i)]$ and $E[\vec{\mathbf{S}}^*_{\gamma_{b_j}}(j\omega_i)\,\vec{\mathbf{D}}(j\omega_i)]$. However, since $\vec{d}(t_n)$ is zero mean, and $\vec{s}_{\gamma_{b_j}}(t_n)$ and $\vec{d}(t_n)$ are uncorrelated, both terms drop out. The effect of additive noise at frequency ω_i will depend on the particular source generator γ_{b_j}. Therefore, estimation of the spectral line component $\mathbf{P}_{\vec{s}}(j\omega_i;\gamma_{b_j})$ will be performed in a power domain $\beta(b_j)$ instead of the normal squared magnitude domain. When a smoothed noise spectral line is subtracted from a noisy speech spectral component, remaining peaks are perceived as musical tones, causing errors in source generator parameterization for recognition. To reduce these effects, magnitude averaging across time as proposed in [3] is applied across each spectral line component. Since the level of perceived interference varies for an utterance across the sequence of noisy generator vectors, a weighting subtraction coefficient $\alpha(b_j)$ is used. The resulting spectral line estimator for source generator γ_{b_j} employing phase information from the original noisy speech vector $\vec{y}_{\gamma_{b_j}}(t_n)$ as $\Theta_{\vec{y}}(j\omega_i)$ is abovedisplayskip10pt

$$\hat{\vec{\mathbf{S}}}_{\gamma_{b_j},\alpha,\beta}(j\omega_i) = \left\{\frac{1}{3}\sum_{i=-1}^{1}\left|\vec{\mathbf{Y}}_{\gamma_{b_j}}(j\omega_i;t_{n+i})\right|^{\beta} - \alpha E\left[\left|\vec{\mathbf{D}}(j\omega_i)\right|^{\beta}\right]\right\}^{\frac{1}{\beta}} e^{j\Theta_{\vec{y}}(j\omega_i)}. \quad (16)$$

For general feature enhancement, the number of spectral lines used for smoothing, power domain β, and weighting term α can all be determined for a given noise degradation. Since the spectral magnitude must be positive, half-wave rectification is performed by setting all negative estimated harmonics to zero. This eliminates errors in over estimating noise bias.

2) Spectral Constraints and Morphological Processing: The spectral line estimator in (16) requires estimates of the power exponent β and weighting coefficient α. These parameters effect the tradeoff between remaining spectral floor of the original broadband noise $\vec{d}(t_n)$ and the residual noise peaks at frequency ω_i for source generator γ_{b_j}. An increase in α results in further broad-band noise reduction but with a corresponding increase in musical tone artifacts. Conversely, a decrease in α causes a decrease in musical tones but with a raised broadband spectral noise floor. Tradeoffs in the choice of α and β are discussed in greater detail in [14].

In order to minimize musical tone effects, morphological set operators are applied in the time versus spectral component domain. This nonlinear processing *fills in* spectral noise valleys and *smooths* irregular spectral noise peaks, thereby constraining the frequency spectra so that vocal tract characteristics do not vary wildly from frame-to-frame when speech is present.

The theory of mathematical morphology is based on a signal being viewed as a set in Euclidean space, where morphological operations are applied using a predetermined structuring element [44]. Employing Minkowski set operations [38] of addition and subtraction as followsabovedisplayskip10pt

$$(f \oplus g)(x) = \sup_{y\in D}\{f(y) + g(x-y)\} \quad (17)$$

$$(f \ominus g)(x) = \inf_{y\in D}\{f(y) - g(x-y)\} \quad (18)$$

the four basic morphological operations of dilation, erosion, closing, and opening can be formed as

$$\mathcal{D}(f,g) = (f \oplus g^s)(x) = f(x) \oplus g(-x) \quad (19)$$

$$\mathcal{E}(f,g) = (f \ominus g^s)(x) = f(x) \ominus g(-x) \quad (20)$$

$$\mathcal{C}(f,g) = [(f \oplus g^s) \ominus g](x) \quad (21)$$

$$\mathcal{O}(f,g) = [(f \ominus g^s) \oplus g](x). \quad (22)$$

By combining morphological operations, an extensive class of morphological filters can be formulated that can replace standard linear filters in many signal processing applications. A more detailed treatment of Minkowski function addition, subtraction, and the four gray-scale morphological operators can be found in [36] and [44].

Let the estimated spectral component prior to magnitude averaging at time t_n be written as

$$\hat{\mathbf{P}}^{\beta}_{\vec{s}}(j\omega_i;\gamma_{b_j},t_n). \quad (23)$$

Let $g(t_n, \omega_i)$, be a 3-by-3 parabolically shaped structuring element centered at the spectral line ω_i and time locations (t_{n-1}, t_n, t_{n+1}). If spectral components from (16) contain residual noise peaks, morphological operations of erosion or opening are able to attenuate them while preserving the overall frequency response structure. Dilation and closing operations can also be used to fill in persistent irregular spectral valleys. Morphological operators offer the advantage over nonlinear median filters in their ability to control the removal of positive and negative impulse noise separately. The resulting estimated spectral component after application of a dilation operation is

$$\mathcal{D}(\hat{\mathbf{P}}^{\beta}_{\vec{s}},g) = \left[\hat{\mathbf{P}}^{\beta}_{\vec{s}}(j\omega_i;\gamma_{b_j},t_n)\right]_{\mathcal{D}_g} = \hat{\mathbf{P}}^{\beta}_{\vec{s}}(j\omega_i;\gamma_{b_j},t_n) \oplus g(-t_n,-\omega_i). \quad (24)$$

With this morphological constrained spectral component, (16) can be expressed as

$$\hat{\vec{\mathbf{S}}}_{\gamma_{b_j},\alpha,\beta,\mathcal{D}_g}(j\omega_i) = \left\{\frac{1}{3}\sum_{i=-1}^{1}\left[\left|\vec{\mathbf{Y}}_{\gamma_{b_j}}(j\omega_i;t_{n+i})\right|^{\beta} - \alpha E\left[\left|\vec{\mathbf{D}}(j\omega_i)\right|^{\beta}\right]\right]_{\mathcal{D}_g}\right\}^{\frac{1}{\beta}} e^{j\Theta_{\vec{y}}(j\omega_i)}. \quad (25)$$

employing a dilation morphological constraint across spectral lines $\omega_i \in [0,\pi]$. The particular morphological operator applied reduces spectral fluctuations caused by errors in noise characterization, as well as reducing the chance for erratic movements of individual spectral harmonics across time. A brief evaluation of the spectral line estimator in (25) is presented in the next section. For enhancement evaluation, detected source boundary information b_j was used to adjust the extent of the structuring element, power domain β, and weighting term α. Although a variety of adaptive constraint methods are possible for the sequence of source generators, the recognition evaluations presented later employ only an opening operation with fixed values of β and α across the estimated spectral line components. This was necessary due to

TABLE III
SPEECH QUALITY COMPARISON OF ENHANCEMENT ALGORITHMS ACROSS SOUND TYPES FOR WHITE GAUSSIAN NOISE. SNR = +10 dB.

Sound Type	*Itakura-Saito Likelihood Measure*			
	Original	*Spec.Sub.*	*ST-Wiener*	*MO-α,β,b_j*
Silence	1.438	1.632	1.453	0.941
Vowel	2.782	2.280	2.474	1.631
Nasal	15.695	13.783	8.223	7.049
Stop	5.272	4.128	3.714	1.744
Fricative	3.066	3.269	2.932	2.221
Glide	1.412	1.533	1.353	0.655
Liquid	6.923	3.647	5.415	1.636
Affricate	2.988	2.023	2.547	0.885
Voiced + Unvoiced	4.335	3.522	3.339	2.006
Total	3.095	2.713	2.532	1.547

additional processing needed for Lombard effect compensation of enhanced noisy Lombard speech.

3) Feature Enhancement Evaluation: Evaluation of the spectral line estimator in (25) is considered in terms of estimated speech quality across phoneme classes. This will determine if characteristics important to speech quality can be improved for source generator parameterization under noisy conditions. A subset of continuous speech from the TIMIT [39] database was degraded by additive white Gaussian noise and was processed. A Bartlett spectral estimate was used to characterize noise using data outside primary detected endpoints. Objective speech quality measures that have been shown to possess good correlation with subjective quality [42] were used to determine enhanced speech quality performance. The measure used in this evaluation is the Itakura–Saito likelihood ratio. Performance over sound classes was accomplished by partitioning speech into segments, processing entire sentences, and computing objective measures for each class. Table III summarizes this comparison between basic spectral subtraction [3] (three-frame magnitude averaging with half-wave rectification), short-time Wiener filtering [30] (terminated at the fourth iteration), and the morphological constrained approach $MO\text{-}\alpha,\beta,b_j$. $MO\text{-}\alpha,\beta,b(i)$ results in improved quality for all types of speech and consistently outscored short-time Wiener filtering and spectral subtraction techniques.

C. Lombard Effect Compensation

Given a detected sequence of speech source generators $\gamma_{b_j}{:}(b_j = b_1, \ldots, b_L)$ and their feature enhanced spectral representations, we now turn to the issue of modeling the change in source generator class from neutral to Lombard effect speaking conditions. From (4), stressed speaking conditions are addressed by the choice of an alternate or modified source generator for each phoneme-like section. Let the estimated speech vector under the noisy neutral and Lombard stress conditions be written as $\hat{\vec{s}}_{\gamma_{b_j}}(t_n)$ and $\hat{\vec{s}}_{\Psi[\gamma_j]_i}(t_n)$ respectively, where $\Psi[\cdot]_i$ represents a stress-based change in the source generator.

Next, let $\vec{\mathbf{C}}_{\gamma_{b_j}}(t_n){:}t_n \in [1, N_{b_j}]$ be a sequence of mel-cepstral vectors over time index t_n from source generator γ_{b_j} under neutral speaking conditions. We further assume that this sequence is obtained over an input utterance token set during the training phase of the HMM recognizer. Let the same sequence for generator γ_{b_j} under Lombard effect stress be modeled as

$$\vec{\mathbf{C}}_{\Psi[\gamma_{b_j}]_i}(t_n) = \vec{\mathbf{C}}_{\gamma_{b_j}}(t_n) + \vec{\mathbf{C}}_{\Psi_i(b_j)} : t_n \in [1, N_{b_j}], \quad (26)$$

where $\vec{\mathbf{C}}_{\Psi_i(b_j)}$ represents an additive stress effect component that depends on the particular stress class Ψ_i and source generator b_j. It is assumed that the output of each source generator can be modeled as a sequence of independent identically distributed random vectors with an estimated mean and variance. In a manner similar to [7], we assume that this component takes on an exponential form, but as suggested in [17], the actual exponential form is unique for each source generator across stressed speaking conditions

$$\vec{\mathbf{C}}_{\Psi_i(b_j)} = C_{\Psi_i(b_j)}(k{:}k = 1, \ldots, 9) \quad (27)$$

$$C_{\Psi_i(b_j)}(k) = \mu(b_j, \Psi_i)e^{-\nu(b_j,\Psi_i)(k-1)} \quad k = 1, \ldots, 9. \quad (28)$$

Here, $\mu(b_j, \Psi_i)$ and $\nu(b_j\Psi_i)$ are fixed for each source generator under stress condition Ψ_i. The kth mel-cepstral parameter under stressed conditions from (26) will have the following probability density function:

$$f_{C_{\Psi[\gamma_{b_j}]_i}}\left(C_{\Psi[\gamma_{b_j}]_i}(k)\right) = \frac{1}{\sqrt{2\pi}\sigma_{k,b_j}} e^{\left\{ -\frac{\left[C_{\Psi[\gamma_{b_j}]_i}(k) - \left(C_{\gamma_{b_j}}(k) + C_{\Psi_i(b_j)}(k)\right)\right]^2}{2\sigma^2_{k,b_j}} \right\}} \quad (29)$$

assuming a sequence of source generator vectors over time instances $t_1 \le t_n \le t_{N_{b_j}}$, which are statistically independent random variables. The mean of the above random variable $C_{\Psi[\gamma_{b_j}]_i}(k)$ will consist of the kth varying mel-cepstral parameter $C_{\gamma_{b_j}}(k, t_n)$ under neutral conditions, plus a modeled stress term $C_{\Psi_i(b_j)}(k)$. The variance σ^2_{k,b_j} for each coefficient k will depend on source generator γ_{b_j}. Given an estimate of the mel-cepstral coefficients over time t_n, and the stress component $C_{\Psi_i(b_j)}(k)$, the log-likelihood of $\vec{\mathbf{C}}_{\Psi[\gamma_{b_j}]_i}(t_n)$ can be found as follows:

$$\begin{aligned}&\mathcal{L}\Big\{C_{\Psi[\gamma_{b_j}]_i}(k, t_1), \ldots, C_{\Psi[\gamma_{b_j}]_i}(k, t_{N_{b_j}}) \mid C_{\gamma_{b_j}}(k, t_1), \ldots, \\ &\qquad C_{\gamma_{b_j}}(k, t_{N_{b_j}}), C_{\Psi_i(b_j)}(k)\Big\} \\ &= -(t_{N_{b_j}} - t_1)\log\frac{1}{\sqrt{2\pi}} \\ &\quad - \sum_{t_n=t_1}^{t_{N_{b_j}}} \frac{\left[C_{\Psi[\gamma_{b_j}]_i}(k, t_n) - \left(C_{\gamma_{b_j}}(k, t_n) + C_{\Psi_i(b_j)}(k)\right)\right]^2}{2\sigma^2_{k,b_j}}.\end{aligned} \quad (30)$$

The unknown model parameter $C_{\Psi_i(b_j)}(k)$ is estimated by maximizing (30), resulting in the maximum likelihood estimate

$$\hat{C}_{\Psi_i(b_j)}(k) = \frac{1}{N_{b_j}} \sum_{t_n=t_1}^{t_{N_{b_j}}} C_{\Psi[\gamma_{b_j}]_i}(k, t_n) - \frac{1}{N_{b_j}} \sum_{t_n=t_1}^{t_{N_{b_j}}} C_{\gamma_{b_j}}(k, t_n). \quad (31)$$

The sequence of detected source generators is obtained during the HMM training phase and used to obtain the sample estimate for $C_{\gamma_{b_j}}(k,t_n)$. A sample estimate for $C_{\Psi[\gamma_{b_j}]_i}(k,t_n)$ is obtained from tokens of actual Lombard effect stressed speech, which is also obtained during the HMM training phase. Since the number of observations from a given source generator γ_{b_j} will vary under stress, the observation number N_{b_j} will normally differ in each summation in (31). Since the number of Lombard effect training tokens is limited, a smoothed exponential decay of the form in (28) is applied to $\hat{\vec{C}}_{\Psi_i(b_j)}$ with the conditions that for each source generator, $\mu(b_j,\Psi_i) = \hat{C}_{\Psi_i(b_j)}(1)$, and $\nu(b_j,\Psi_i) = -\ln(\hat{C}_{\Psi_i(b_j)}(k)/\hat{C}_{\Psi_i(b_j)}(1))$ for $\hat{C}_{\Psi_i(b_j)}(k)\hat{C}_{\Psi_i(b_j)}(1) > 0$ and $|\hat{C}_{\Psi_i(b_j)}(1)| > |\hat{C}_{\Psi_i(b_j)}(k)|$. A compensation model vector $\hat{\vec{C}}_{\Psi_i(b_j)}$ is estimated for each detected source generator section during HMM training and applied during recognition evaluation. The result of morphological constrained enhancement and adaptive cepstral compensation is the provision of a sequence of speech generator parameters that characterize the changing features of $\vec{s}_{\gamma_j}$ across sources but limit the effects of additive noise and stress-induced Lombard effect on speech production for improved recognition performance. The following section briefly describes the final details of MCE-ACC-HMM recognition framework.

D. Hidden Markov Model Recognition

The MCE-ACC-HMM algorithm 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 source generator-compensated mel-cepstral coefficients. A 64-state vector quantizer is used and trained using a binary-split procedure similar to the Lloyd algorithm. Next, a speaker-dependent, isolated word, five-state left-to-right hidden Markov model is formulated for each entry in the recognizer dictionary. Forward $\alpha_t(i)$ and backward probabilities $\beta_t(i)$ are obtained and used to form the familiar Baum–Welch forward-backward estimator

$$P(O \mid \lambda) = \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_t(i)a_{ij}\beta_j(O_{t+1})\beta_{t+1}(j) \qquad 1 \le t \le T-1. \quad (32)$$

Reestimation relations can then be obtained for model parameters $\bar{a}_{ij}$, $\bar{\pi}_i$, $\bar{b}_j(k)$ in $M(\vec{A},\vec{B},\pi)$. Details can be found in [10] and [29]. In the training phase, each model was initiated with essentially random choices for nonzero elements and then iteratively adjusted to increase $P(\vec{\Phi}|\underline{\mathbf{M}})$, where the probability of the observation sequence $\vec{\Phi}$ has been generated by model $\underline{\mathbf{M}}$. A separate Lombard stress compensation model is obtained and associated with each HMM word model for recognition.

E. Phonetic Consistency Rule

Since $v/t/uv$ profiles are obtained during training and recognition, a probabilistic decision criterion was employed to augment the HMM-based recognition procedure. Using mean $v/t/uv$ training profiles, a phonetic consistency rule was implemented for preclassification of the input utterance under test. The consistency rule was based on the following:

i) number of labeled phonemes
ii) sequence particular phonemes
iii) duration of phonemes
iv) the contour of speech energy as represented by the first mel-cepstral parameter $c_{0,L(i)}(i)$.

Implementation of this rule required a statistical characterization of duration variation of labeled phoneme sections. A confidence interval of 3σ is used in this phase.

V. MCE-ACC-HMM Evaluation

Performance of the new MCE-ACC-HMM recognition algorithm is considered in five recognition scenarios, which are shown in Fig. 5. The first four scenarios establish baseline recognition scores for comparison, whereas the fifth represents the framework used for MCE-ACC-HMM evaluation in noisy Lombard effect speech conditions. Input speech recognition conditions considered for algorithm evaluation include the following:

i) noise-free neutral speech
ii) noise-free Lombard speech
iii) neutral speech with additive noise
iv) Lombard speech with additive noise.

Speech data used for evaluation consisted of a 35-word vocabulary spoken by three male speakers (denoted as speakers S1, S2, and S3). For each speaker, 12 tokens for each word under neutral noise-free conditions, and two tokens under Lombard conditions, were used. The vocabulary consists of highly confusable word pairs such as /six-fix/, /go-oh-no/, and /wide-white/. In all evaluations, recognition training employed 10 neutral tokens and tested using two neutral and two Lombard tokens of each word for each speaker. Although HMM training is fully open, limited Lombard effect data required the use of both Lombard tokens for statistical characterization of source generators under Lombard condition for each model (i.e., Lombard recognition not fully open). In the evaluation, three noise sources are considered:

i) white Gaussian noise (WGN)
ii) nonstationary cooling fan noise from an IBM PS-2 workstation (PS2)
iii) nonstationary Lockheed C130 aircraft cockpit noise (AIR).

Fig. 6 illustrates time versus spectral plots of the three sample noise sources. Noise levels were adjusted using the following relations to obtain overall signal-to-noise ratios (SNR) of 10, 20, and 30 dB.

$$y(i) = s(i) + G \cdot d(i) \quad (33)$$

$$\mathrm{SNR}_{\mathrm{GLOBAL}} = 10\log\frac{\sum_{i=1}^{N} s^2(i)}{\sum_{i=1}^{N}(G\cdot d(i))^2}. \quad (34)$$

Here, $s(i)$ represents the noise-free neutral or Lombard effect speech, and $d(i)$ a sample noise sequence. A closer analysis of noisy input data revealed that confusable word pairs such as /six-fix/, with distinguishing lead or trailing phonemes, had frame oriented segmental SNR's that were consistently 14–28 dB below global averages. Therefore, recognition perfor-

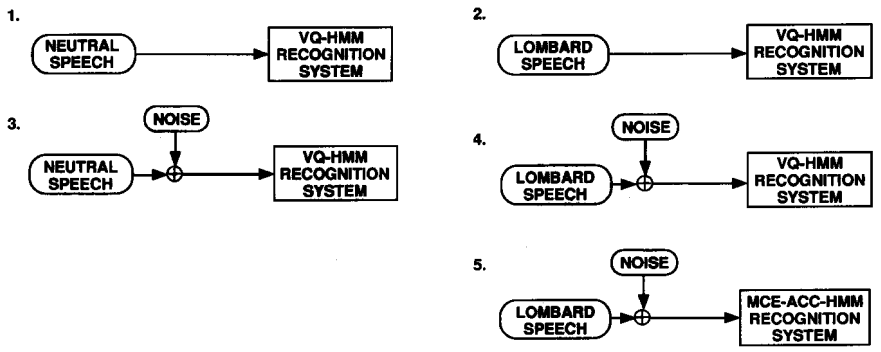


Fig. 5. Automatic speech recognition scenarios. The five environments consist of various levels and/or types of noise for neutral and Lombard effect speech.

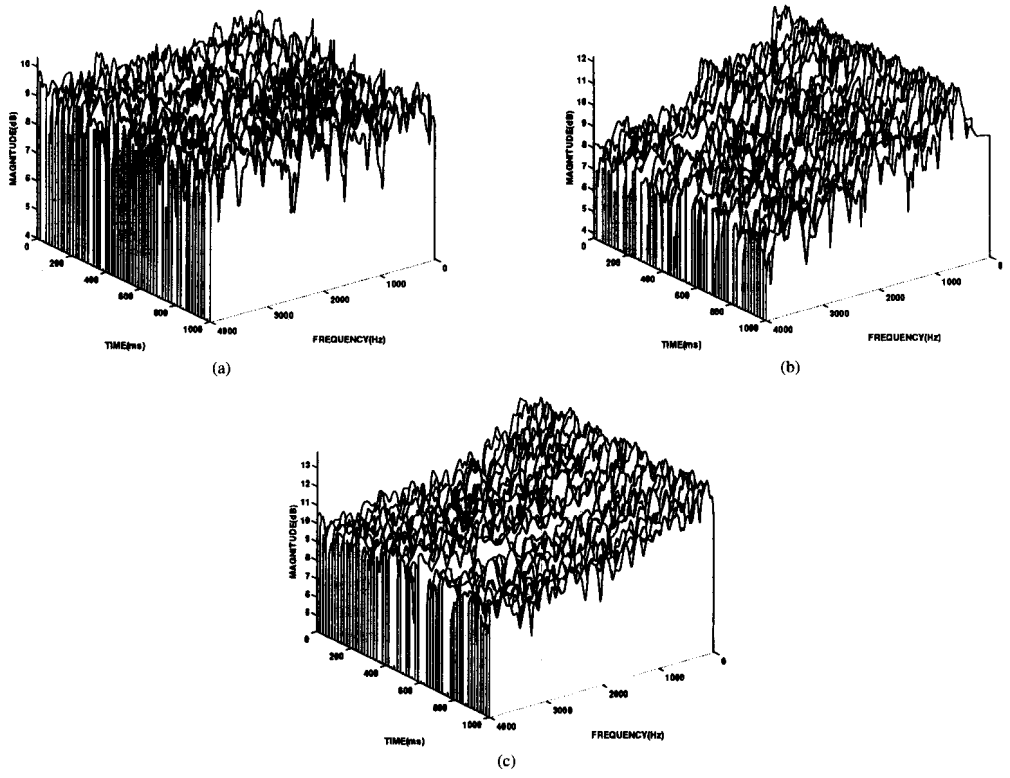
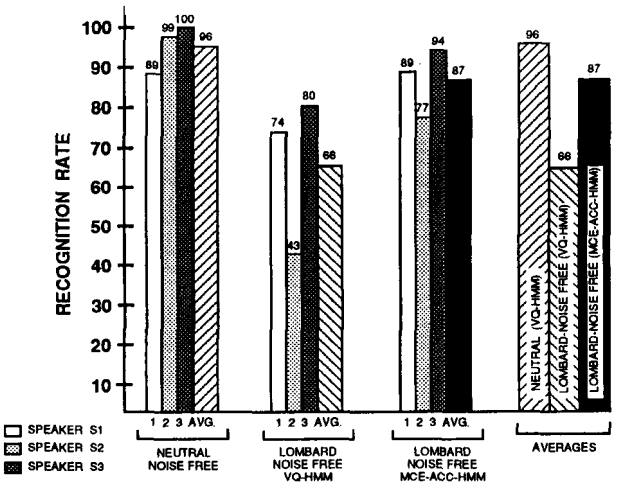


Fig. 6. Time versus power spectral response for three background noise distortions.

mance will be more dependent on local SNR of distinguishing phonemes in confusable word groups than for homogeneous high-energy voiced sections such as the vowel /IY/ in /freeze-three/ or diphthong /AI/ in /wide-white/.

To compare performance, a fairly standard, isolated-word, discrete-observation hidden Markov model (VQ-HMM) recognizer was used. The baseline system was mel-cepstral parameter based with no embellishments. In all experiments, a five-state, left-to-right model was used. A vector quantizer was used to generate a 64-state codebook using 2 min of noise-free training data. The 35 models employed by the VQ-HMM recognizer were trained using the forward-backward algorithm with 10 tokens used to train each model and two for testing. An earlier version of this baseline recognizer based on LPC parameters was used in earlier recognition evaluations for noisy speech under stress [12], [22]. The only modification is that mel-cepstral parameters are used in place of LPC coefficients.[6]

[6]The modification from LPC to mel-cepstral parameters for the baseline recognizer used for Table I was performed for consistency so that differences in input speech parameterization would not effect recognition performance.



NOISE FREE RECOGNITION RESULTS					
Condition	Speaker S1	Speaker S2	Speaker S3	Mean $\bar{x}$	St.Dev. σ_{recog}
Neutral & VQ-HMM	88.6%	98.6%	100.0%	95.7%	6.23
Lombard & VQ-HMM	74.2%	42.9%	80.0%	65.7%	19.98
Lombard & MCE-ACC-HMM	89.4%	77.1%	93.9%	86.8%	8.69

Fig. 7. Individual noise-free speech recognition results across speakers and their recognition means and standard deviations over all speakers for the baseline system (VQ-HMM) and the new robust recognition algorithm (MCE-ACC-HMM).

A. Noise-Free Recognition Results

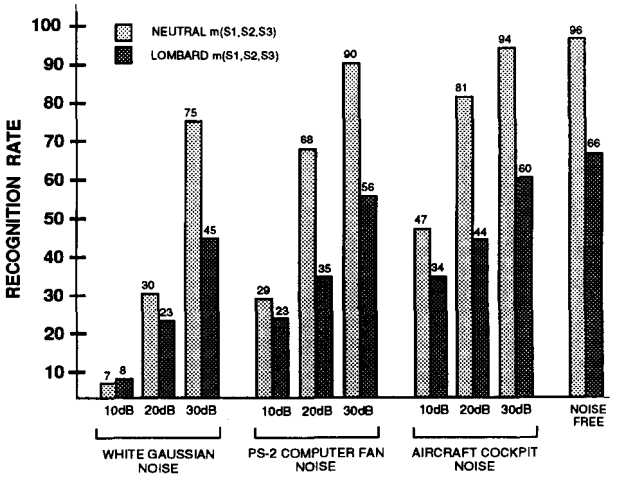
The first evaluation step is to establish performance for noise-free conditions (see Fig. 5, scenarios 1 and 2). Fig. 7 summarizes noise-free recognition results for baseline VQ-HMM and MCE-ACC-HMM recognizers in neutral and Lombard effect speech conditions. The VQ-HMM recognition rate of 96% establishes an upper limit of performance for the chosen confusable vocabulary.[7] Individual results over speakers S1, S2, and S3 are also shown. The standard deviation in recognition rate σ_{RECOG} of 6.23 confirms reliable performance across speakers. The second baseline average recognition rate of 66% establishes a lower limit of performance for noise-free Lombard speech. For this system, speech production variation reflected in speaking under the Lombard effect has reduced recognition performance by an average -30% (individual losses range from -14.4 to -55.7%). The actual loss in performance will depend on vocabulary confusability and the concentration of phonemes most susceptible to Lombard effect speaking style.[8] The corresponding increase in σ_{RECOG} from 6.2 to 19.9 reflects irregular recognition performance due to individual inter-speaker variations caused by the Lombard effect. Next, MCE-ACC-HMM is evaluated in noise-free Lombard speaking conditions (speech enhancement sections are disabled for this evaluation). Individual recognition rates increase for all speakers, with an average improvement over VQ-HMM of +21%. The influence of the Lombard effect on recognition was almost eliminated for speaker $S1$ and significantly improved for speakers $S2$ and $S3$. In addition to raising the mean recognition rate to 87%, a more consistent level of performance is achieved as reflected in by a decrease in σ_{RECOG} from 19.9 to 8.69.

B. Baseline VQ-HMM Noisy Recognition Results

Baseline VQ-HMM recognition rates are established for noisy neutral and Lombard effect speaking conditions (see Fig. 5, scenarios 3 and 4). Fig. 8 summarizes performance across three speakers for neutral and Lombard effect styles in nine noise conditions. Mean and standard deviation across all speaker evaluations are shown. Performance for VQ-HMM is severely effected in noisy Lombard conditions. These rates establish lower limits of recognition performance in noisy environments for the confusable vocabulary. For noisy neutral speaking conditions, performance was reasonable for speech corrupted by PS-2 cooling fan noise and aircraft cockpit noise at 30 dB. As SNR decreased, however, recognition rates dropped sharply. Additive WGN had a more pronounced impact on recognition performance than either PS2 or AIR noise. This was attributed to the fact that these noise sources contained little high-frequency content; therefore, although average time-domain segmental SNR for distinguishing consonants in confusable word-pairs would be the same for WGN, high-frequency segmental SNR is higher for aircraft and computer fan noise cases. Thus, the effective SNR for

[7] The same VQ-HMM recognizer achieved a recognition rate of 100% for a less-confusable 10-word vocabulary.

[8] An earlier evaluation, using a 20-word vocabulary and an LPC parameterized discrete-observation HMM recognizer, resulted in a -25% loss in noise free recognition rate [12], [22].



BASELINE VQ-HMM NOISY RECOGNITION RESULTS

Noise Type	Signal-to-Noise Ratio					
	10 dB		20 dB		30 dB	
	Neutral	Lombard	Neutral	Lombard	Neutral	Lombard
	Mean $\bar{x}_{recog}(S1, S2, S3)$ *over all speakers*					
WGN	6.7%	8.1%	30.0%	23.3%	75.2%	45.2%
PS-2	29.1%	23.3%	68.1%	35.2%	90.0%	56.2%
Air	47.1%	34.3%	81.0%	44.3%	94.3%	60.5%
	Stand. Dev. $\sigma_{recog}(S1, S2, S3)$ *over all speakers*					
WGN	2.97	2.98	11.16	5.02	5.77	19.29
PS-2	26.97	15.01	18.20	18.09	12.37	19.02
Air	18.24	17.15	9.72	16.48	8.69	22.96

Fig. 8. Summary of mean $\bar{x}_{\text{recog}}(S1, S2, S3)$ and standard deviation $\sigma_{\text{recog}}(S1, S2, S3)$ in recognition rate for the baseline (VQ-HMM) recognizer under nine noisy environmental conditions for neutral and Lombard effect speaking styles.

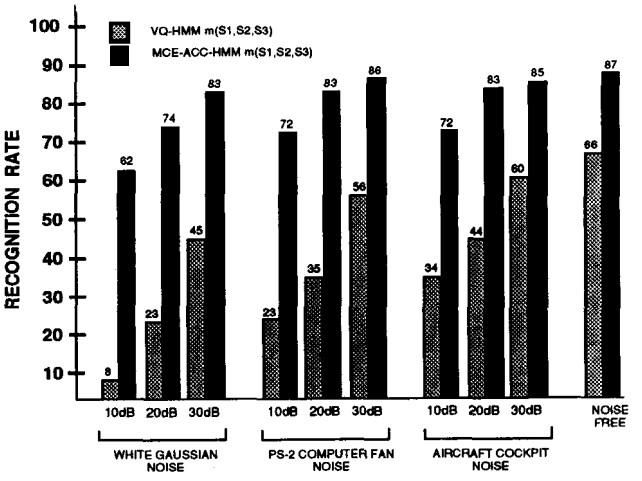
distinguishing consonants in word-pairs such as /six-fix/ is higher for low-frequency degradation such as aircraft cockpit noise but lower for broad-band white Gaussian noise. The final overall mean Lombard effect recognition rate across all speakers, noise types, and SNR's (27 noise conditions) is $\bar{x}_{\text{RECOG}} = 36.7\%$, with a standard deviation in recognition of $\sigma_{\text{RECOG}} = 21.1$.

C. MCE-ACC-HMM Noisy Recognition Results

Noisy Lombard speech recognition performance for the MCE-ACC-HMM recognizer is considered (see Fig. 5, scenario 5). Fig. 9 summarizes recognition results for the nine noisy Lombard speech conditions for both VQ-HMM and MCE-ACC-HMM recognizers. With the exception of source generator statistical characterization for Lombard effect compensation, all results are open tests for noisy Lombard conditions using a neutral trained vector quantized codebook and hidden Markov models. Mean and standard deviation in recognition rate over all speaker evaluations are shown. MCE-ACC-HMM outperforms the baseline VQ-HMM recognizer for all tested noisy conditions. Performance across 10–30 dB of additive white Gaussian noise resulted in a +44.4% improvement in recognition performance over VQ-HMM baseline system (mean increase from 25.7 to 70.1%). For varying levels of C130 aircraft cockpit noise, average recognition rates increased by +30.1% to a mean rate of 76.3%. Finally, for PS-2 cooling fan noise, average recognition rates over 10–30 dB increased +39.4% to 77.8%. The variability of recognition as measured by standard deviation in recognition consistently decreased for each noise type over the SNR range of 10–30 dB. Evaluations showed that σ_i decreases using MCE-ACC-HMM from 19.0 to 11.6 for AWGN, 20.1 to 12.8 for aircraft cockpit noise, and 20.9 to 11.1 for PS-2 cooling fan noise. Employing individual recognition scores for all 27 noisy conditions, the final mean recognition rate increased from 36.7% for VQ-HMM to 74.7% for MCE-ACC-HMM, with a corresponding decrease in the variability of recognition from $\sigma_{\text{RECOG:VQ-HMM}} = 21.1$ to $\sigma_{\text{RECOG:MCE-ACC-HMM}} = 11.9$. These results demonstrate the consistency of MCE-ACC-HMM recognition improvement for noisy Lombard effect speaking conditions.

D. Confusion Matrices

Although mean and standard deviation in recognition rates demonstrate improvement for MCE-ACC-HMM, confusion



NOISY LOMBARD SPEECH RECOGNITION RESULTS						
Recognition Conditions	Signal-to-Noise Ratio					
	10 dB		20 dB		30 dB	
Noise & Recognizer	$\bar{x}_{10}$	σ_{10}	$\bar{x}_{20}$	σ_{20}	$\bar{x}_{30}$	σ_{30}
WGN & HMM-VQ	8.1%	3.0	23.3%	5.0	45.2%	19.3
WGN & MCE-ACC-HMM	62.1%	3.6	74.3%	7.8	83.2%	8.9
AIR & HMM-VQ	34.3%	15.0	44.3%	18.1	60.5%	19.0
AIR & MCE-ACC-HMM	72.1%	14.3	83.4%	15.6	84.6%	7.1
PS2 & HMM-VQ	23.3%	17.2	35.2%	16.5	56.2%	23.0
PS2 & MCE-ACC-HMM	72.3%	5.0	83.3%	15.0	86.7%	9.6

Fig. 9. Recognition results for HMM-VQ and new robust MCE-ACC-HMM recognizers for three types of noise and three SNR's. Overall mean $\bar{x}_{recog}$ and standard deviation σ_{recog} in recognition rate across all speakers are shown.

TABLE IV
OVERALL RECOGNITION RESULTS FOR THE VQ-HMM RECOGNIZER AND THE NEW ROBUST RECOGNIZER MCE-ACC-HMM FOR THREE TYPES OF NOISE. NOISE FREE, AVERAGES OVER ALL NOISY CONDITIONS, AND THE STANDARD DEVIATION OF NOISY RECOGNITION RATES ARE ALSO SHOWN.

OVERALL NOISEFREE & NOISY LOMBARD EFFECT RECOGNITION PERFORMANCE										
	Noise Free		Noisy Lombard Conditions						OVERALL	
Speech & Recognizer	$\bar{x}$	σ	WGN		Aircraft		PS-2 Fan		NOISY LOMBARD	
Neutral & *VQ-HMM*	96.0%	6.1	$\bar{x}$	σ	$\bar{x}$	σ	$\bar{x}$	σ	$\bar{x}_{RECOG}$	σ_{RECOG}
Lombard & *VQ-HMM*	65.7%	19.9	25.7%	19.0	46.2%	20.1	38.4%	20.9	36.7%	21.1
Lombard & *MCE-ACC-HMM*	86.7%	8.7	70.1%	11.6	76.3%	12.8	77.8%	11.1	74.7%	11.9

matrices can more clearly illustrate how mel-cepstral-based Lombard compensation is able to reduce errors caused by low-energy consonants. Fig. 10 shows example confusion matrices for VQ-HMM and MCE-ACC-HMM algorithms using speech under the Lombard effect with additive aircraft cockpit noise (30 dB SNR) for one speaker. A black square refers to two tokens, whereas a gray square indicates a single token in place (i.e., a normal error). Here, 11 of the 27 errors for VQ-HMM, due to confusable word pairs under noisy Lombard effect speech conditions, are corrected when MCE-ACC-HMM is used. As a result, the error rate is reduced from 37 to 17%. This was due to improved feature representation resulting from morphological constrained enhancement and cepstral compensation along with application of the phonetic consistency rule. It is noted that in all recognition evaluations, the phonetic consistency rule worked flawlessly (i.e., the correct model was always included as part of the final search set). This rule also reduces Lombard stress compensation processing by roughly a factor of three by limiting the dictionary search from 35 to an average of 13 word models. Further studies show that half of all errors due to confusable word pairs are eliminated with the new MCE-ACC-HMM algorithm.

VI. Discussion and Conclusion

This paper has described a new low-vocabulary speech recognition algorithm (MCE-ACC-HMM) that provides robust performance in noisy environments with particular emphasis on characteristics due to the Lombard effect. A stressed-based source generator framework is established to achieve improved speech parameter characterization using morphological

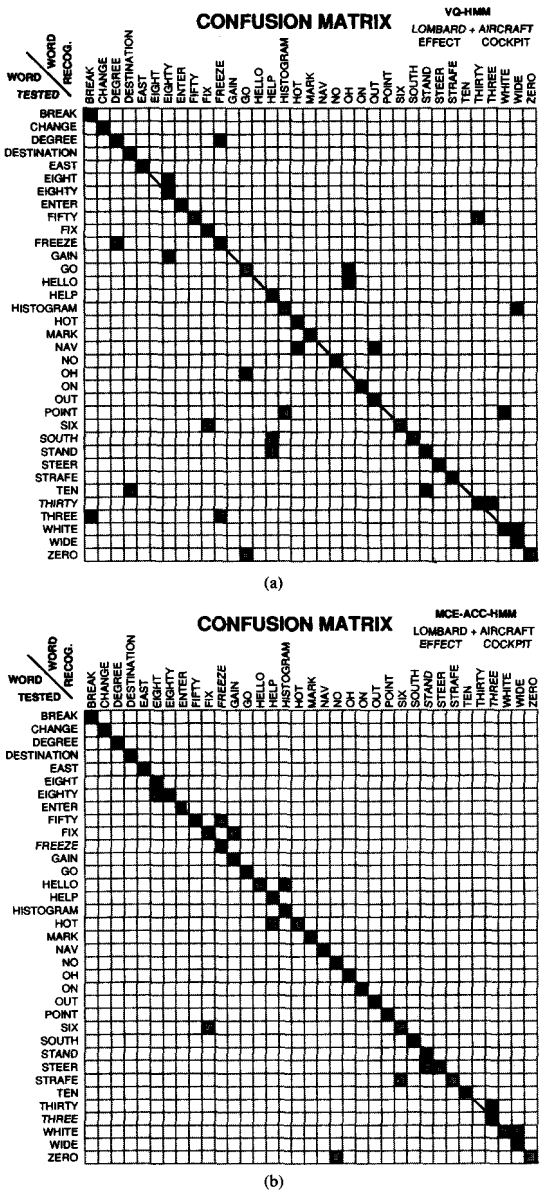


Fig. 10. Sample: (a) VQ-HMM; and (b) MCE-ACC-HMM confusion matrices for one speaker under Lombard effect and aircraft cockpit noise conditions.

constrained feature enhancement and stressed source compensation, which is unique for each source generator across a stressed speaking class. It has been shown that the impact of stress effects the sequence of speech source generators differently in noise-free and noisy conditions. Therefore, the proposed algorithm uses a noise adaptive boundary detector to obtained a sequence of source generator classes, which in turn directs noisy feature enhancement and stress compensation. This allows the parameter enhancement and stress compensation schemes to adapt to changing speech generator types. A phonetic consistency rule is also employed based on input source generator partitioning.

A stress source compensation model is obtained for each word in the input vocabulary during training, whereas for the testing phase, a corresponding stress-compensation model for each tested hidden Markov word model is applied to the unknown input source generator sequence. If the correct word model (HMM) is under consideration, then stress compensation will limit spectral variations. If an incorrect word model (HMM) is considered, then stress compensation using the incorrect stress model introduces spectral variations that increases word rejection by the recognizer.

MCE-ACC-HMM was compared with a more traditional discrete-observation VQ-HMM recognizer with no embellishments. The evaluation considered noise-free Lombard effect conditions and nine noisy Lombard conditions using a highly confusable vocabulary. Noise conditions included additive white Gaussian noise, aircraft cockpit noise, and noise from the cooling fan of a computer workstation, all at three levels of SNR. Performance increased for these three noise sources by +44.4, +30.1, and +39.4% respectively, demonstrating the method's ability to perform in stationary and slowly varying colored noise conditions. Overall recognition rates for noisy Lombard speech were shown to increase from an average of 36.7% for the baseline recognizer to 74.7% for the new algorithm (a +38% improvement). The new algorithm was also shown to be more consistent under varying noisy conditions as demonstrated in a decrease in the standard deviation of recognition rates from 21.1 to 11.9 and a reduction in confusable word-pairs shown in confusion matrices. Finally, it is noted that while MCE-ACC-HMM improves recognition performance in noisy Lombard effect conditions, additional computational resources over baseline VQ-HMM are needed. If computational resources are limited, future studies might consider limiting feature enhancement or stress compensation to only those source generators that have the largest influence on overall recognition performance. In addition, further studies might also consider varying the type of morphological processing based on background noise type and level.

REFERENCES

[1] A. Acero and R. M. Stern, "Environmental robustness in automatic speech recognition," in *Proc. 1990 IEEE ICASSP* (Albuquerque, NM), Apr. 1990, pp. 849–952.

[2] M. Berouti, J. Makhoul, and R. Schwartz, "Enhancement of speech corrupted by acoustic noise," in *Proc. 1979 IEEE ICASSP* (Washington DC), Apr. 1979, pp. 208–211.

[3] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust, Speech Signal Processing*, vol. ASSP-27, pp. 113–120, Apr. 1979.

[4] O. N. Bria, "Improved automatic speech recognition under Lombard effect," M.S. Thesis, Dept. of Elect. Eng., Duke Univ., May 1991.

[5] D. A. Cairns, "Real-time speech recognition under Lombard effect and in noise," M.S. Thesis, Dept. of Elect. Eng., Duke Univ., May 1991.

[6] D. A. Cairns and J. H. L. Hansen, "ICARUS: An Mwave based real-time speech recognition system in noise and Lombard effect," in *Proc. ICSLP-92, Int. Conf. Spoken Language Processing* (Alberta, Canada), Oct. 1992, pp. 703–706, vol. II.

[7] Y. Chen, "Cepstral domain talker stress compensation for robust speech recognition," *IEEE Trans. Acoust, Speech Signal Processing*, pp. 433–439, Apr. 1988.

[8] G. J. Clary and J. H. L. Hansen, "A novel speech recognizer for keyword spotting," in *Proc. ICSLP-92, Int. Conf. Spoken Language Processing* (Alberta, Canada), Oct. 1992, pp. 13–16, vol. I.
[9] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, no. 4, pp. 357–366, Aug. 1980.
[10] J. Deller, J. Proakis, and J. H. L. Hansen, *Discrete Time Processing of Speech Signals*. New York: Macmillan, 1993.
[11] M. B. Gardner, "Effect of noise system gain, and assigned task on talking levels in loudspeaker communication," *J. Acoust. Soc. Amer.*, vol. 40, pp. 955–965, 1966.
[12] J. H. L. Hansen, "Analysis and compensation of stressed and noisy speech with application to robust automatic recognition," Ph.D. thesis, Georgia Inst. of Technol., Atlanta, July 1988.
[13] ——, "Evaluation of acoustic correlates of speech under stress for robust speech recognition," in *IEEE Proc. Northeast Bioeng. Conf.* (Boston, MA), Mar. 1989.
[14] ——, "A new speech enhancement algorithm employing acoustic endpoint detection and morphological based spectral constraints," in *Proc. 1991 IEEE ICASSP* (Toronto, Canada), May 1991, pp. 901–904.
[15] ——, "Detection and recognition of key words under noisy, stressful conditions," submitted to Nat. Sci. Foundation, Grant no. NSF-IRI-90-10536, Final Tech. Rep. DSPL-93-3, Robust Speech Processing Lab., 261 pgs, Dept. of Elect. Eng., Duke Univ., Mar. 1993.
[16] ——, "Adaptive source generator compensation and enhancement for speech recognition in noisy stressful environments," in *Proc. 1993 IEEE ICASSP* (Minneapolis, MN), Apr. 1993, pp. 95–98.
[17] J. H. L. Hansen and O. N. Bria, "Lombard effect compensation for robust automatic speech recognition in noise," in *Proc. 1990 Int. Conf. Spoken Language Processing* (Kobe, Japan), Nov. 1990, pp. 1125–1128.
[18] ——, "Improved automatic speech recognition in noise and Lombard effect," in *Proc. EURASIP-92, Sixth Euro. Signal Processing Conf.* (Brussels, Belgium), Aug. 1992, pp. 403–406.
[19] J. H. L. Hansen and M. A. Clements, "Iterative speech enhancement with spectral constraints," in *Proc. 1987 IEEE ICASSP*, Apr. 1987, pp. 189–192.
[20] ——, "Evaluation of speech under stress and emotional conditions," in *Proc. Acoust. Soc. Amer.* (Miami, FL), Nov. 1987.
[21] ——, "Constrained iterative speech enhancement with application to automatic speech recognition," in *Proc. 1988 IEEE ICASSP* (New York, NY), Apr. 1988, pp. 561–564.
[22] ——, "Stress compensation and noise reduction algorithms for robust speech recognition," in *Proc. 1989 IEEE ICASSP* (Glasgow, Scotland), May 1989, pp. 266–269.
[23] ——, "Constrained iterative speech enhancement with application to speech recognition," *IEEE Trans. Signal Processing*, vol. 39, no. 4, pp. 795–805, Apr. 1991.
[24] B. A. Hanson and H. Wakita, "Spectral slope based distortion measures with linear prediction analysis for word recognition in noise," in *Proc. 1986 IEEE ICASSP* (Tokyo, Japan), Apr. 1986, pp. 757–760.
[25] B. H. Juang, "Speech recognition in adverse environments," *Computer Speech Language*, vol. 5, pp. 275–294, 1991.
[26] J. C. Junqua, "The Lombard reflex and its role on human listeners and automatic speech recognizers," *J. Acoust. Soc. Amer.*, vol. 93, pp. 510–524, Jan. 1993.
[27] K. D. Kryter, *The Effects of Noise on Man*. New York: Academic, 1970.
[28] L. F. Lamel, L. R. Rabiner, A. E. Rosenberg, and J. G. Wilpon, "An improved endpoint detector for isolated word recognition," *IEEE Trans. Acoust, Speech Signal Processing*, vol. ASSP-29, pp. 777–785, Aug. 1981.
[29] 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.*, pp. 1035–1074, Apr. 1983.
[30] J. S. Lim and A. V. Oppenheim, "All-pole modeling of degraded speech," *IEEE Trans. Acoust. Speech Signal Processing*, vol. ASSP-26, pp. 197–210, June 1978.
[31] R. P. Lippmann, E. A. Martin, and D. B. Paul, "Multi-style training for robust isolated-word speech recognition," in *Proc. 1987 IEEE ICASSP*, Apr. 1987, pp. 705–708.
[32] P. Lockwood, J. Boudy, and M. Blanchet, "Non-linear spectral subtraction (NSS) and hidden Markov models for robust speech recognition in car noise environments," in *Proc. 1992 IEEE ICASSP*, Mar. 1992, pp. 265–268, vol. I.
[33] E. Lombard, "Le Signe de l'Elevation de la Voix," *Ann. Maladies Oreille, Larynx, Nez, Pharynx*, vol. 37, pp. 101–119, 1911.
[34] F. J. Malkin and K. A. Christ, "Human factors engineering assessment of voice technology for the light helicopter family," *U.S. Army Human Eng. Lab. Tech. Rep.*, pp. 1–20, June 1985.
[35] D. Mansour and B. H. Juang, "A family of distortion measures based upon projection operation for robust speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 1659–1671, 1988.
[36] P. Maragos and R. W. Schafer, "Morphological filters, Parts I and II," *IEEE Trans. Acoust., Speech, Signal Processing*, pp. 1153–1184, Aug. 1987.
[37] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Processing*, pp. 137–145, Apr. 1980.
[38] H. Minkowski, "Volumen und Oberflache," *Math. Annalen*, vol. 57, pp. 447–495, 1903.
[39] "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database," *Nat. Inst. of Standards Technol.* (NIST), Gaithersburg, MD, (prototype as of Dec. 1988).
[40] D. B. Paul, "A speaker-stress resistant HMM isolated word recognizer," in *Proc. 1987 IEEE ICASSP* (Dallas, TX), Apr. 1987, pp. 713–716.
[41] K. S. Pearsons, R. L. Bennett, and S. Fidell, "Speech levels in various noise environments," Office Health Ecological Effects, Rep. No. EPA-600/1-77-025, 1977.
[42] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective Measures of Speech Quality*. Englewood Cliffs, NJ: Prentice-Hall, 1988.
[43] P. K. Rajasekaran, G. R. Doddington, and J. W. Picone, "Recognition of speech under stress and in noise," in *Proc. 1986 IEEE ICASSP* (Tokyo, Japan), Apr. 1986, pp. 733–736.
[44] J. Serra, *Image Analysis and Mathematical Morphology*. New York: Academic, 1982.
[45] C. A. Simpson, "Speech variability effects on recognition accuracy associated with concurrent task performance by pilots," Psycho-Linguistic Res. Associates, Tech. Rep., pp. 1–15, Apr. 1985.
[46] C. Sorin and C. Thouin-Daniel, "Effects of auditory fatigue on speech intelligibility and lexical decision in noise," *J. Acoust. Soc. America*, vol. 74, no. 2, pp. 456–466, Aug. 1983.
[47] B. J. Stanton, L. H. Jamieson, and G. D. Allen, "Acoustic-phonetic analysis of loud and Lombard speech in simulated cockpit conditions," in *Proc. 1988 IEEE ICASSP* (New York, NY), Apr. 1988, pp. 331–334.
[48] D. Van Compernolle, "Noise adaptation in a hidden Markov model speech recognition system," *Comput. Speech Language*, vol. 3, pp. 151–167, 1989.

John H. L. Hansen (SM'93) was born in Plainfield, NJ. He received the B.S.E.E. degree with highest honors from Rutgers University, New Brunswick, NJ, in 1982 and the M.S. and Ph.D. degrees in electrical engineering from the Georgia Institute of Technology, Atlanta, USA, in 1983 and 1988, respectively.

In 1988, he joined the faculty of Duke University as an Assistant Professor in the Department of Electrical Engineering, amd received a secondary appointment in the Department of Biomedical Engineering in 1993. He is founder and coordinator of the Robust Speech Processing Laboratory in the Electrical Engineering Department. Prior to joining the Duke faculty, he was employed by the RCA Solid State Division and Dranetz Engineering Laboratories. He has served as a technical consultant to industry, including AT&T Bell Laboratories and I.B.M., in the areas of voice communications, wireless telephony, and speech recognition. His research interests span the areas of digital signal processing, analysis and modeling of speech under stress or pathology, speech enhancement and feature estimation in noise, robust speech recognition with a current emphasis on auditory-based constrained speech enhancement, and source generator speech modeling for robust recognition in noise, stress, and Lombard effect.

Dr. Hansen is the author of numerous papers and technical reports in the area of speech processing and is coauthor of the textbook *Discrete-Time Processing of Speech Signals* (Macmillan, 1993). He was the recipient of a National Science Foundation Research Initiation Award in 1990 and was named a Lilly Foundation Teaching Fellow in 1991 and 1992. He has served as Chairman for the IEEE Communications and Signal Processing Societies of North Carolina, Advisor for the Duke University IEEE Student Branch, and he is presently serving as Associate Editor for the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING. He has also served as co-editor of a special issue on Robust Speech Recognition for that publication.