Robust Feature-Estimation and Objective Quality Assessment for Noisy Speech Recognition Using the Credit Card Corpus

John H. L. Hansen, *Senior Member, IEEE,* and Levent M. Arslan, *Student Member, IEEE*

Abstract— **It is well known that the introduction of acoustic background distortion into speech causes recognition algorithms to fail. In order to improve the environmental robustness of speech recognition in adverse conditions, a novel constrained-iterative feature-estimation algorithm, which was previously formulated for speech enhancement, is considered and shown to produce improved feature characterization in a variety of actual noise conditions such as computer fan, large crowd, and voice communications channel noise. In addition, an objective measure based MAP estimator is formulated as a means of predicting changes in robust recognition performance at the speech feature extraction stage. The four measures considered include**

i) NIST SNR
ii) Itakura–Saito log-likelihood
iii) log-area-ratio
iv) the weighted-spectral slope measure.

A continuous distribution, monophone based, hidden Markov model recognition algorithm is used for objective measure based MAP estimator analysis and recognition evaluation. Evaluations were based on speech data from the Credit Card corpus (CC-DATA). It is shown that feature enhancement provides a consistent level of recognition improvement for broadband, and low-frequency colored noise sources. Average improvement across nine noise sources and three noise levels was +9.22%, with a corresponding decrease in recognition rate variability as represented by standard deviation in recognition from 12.4 to 6.5. As the stationarity assumption for a given noise source breaks down, the ability of feature enhancement to improve recognition performance decreases. Finally, the log-likelihood based MAP estimator was found to be the best predictor of recognition performance, while the NIST SNR based MAP estimator was found to be poorest recognition predictor across the 27 noise conditions considered.

I. Introduction

A. An Overview to Robust Speech Recognition

The issue of robustness in speech recognition can take on a broad range of problems. A speech recognizer may be robust in one environment and inappropriate for another. The main reason for this is that performance of existing recognition systems which assume a noise-free tranquil environment, degrade rapidly in the presence of noise, distortion, and stress. In order to discuss the signal processing needed to achieve robust speech recognition, we consider the potential sources of distortion that are introduced into a speech signal under adverse conditions. In Fig. 1, a general speech recognition scenario is presented which includes a variety of speech signal distortions.

Clearly, the distortions illustrated may not be present in unison. For this scenario, we assume that a speaker is exposed to some adverse environment, where ambient noise is present and a user task is required. Such scenarios include a noisy automobile environment where cellular communications is used, noisy helicopter or aircraft cockpits, noisy factory environments, and others. Since the user task could be demanding, the speaker is required to divert a measured level of cognitive processing, leaving formulation of speech for recognition as a secondary task. Workload task stress has been shown to significantly impact recognition performance [51], [9], [28]–[30], [32], [53]. Since background noise is present, the speaker will experience the Lombard effect [48]: a condition where speech production is altered in an effort to communicate more effectively across a noisy environment. The level of Lombard effect will depend on the type and level of ambient noise $d_1(n)$. In addition, a speaker may also experience situational stress (i.e., anger, fear, other emotional effects), which will alter the manner in which speech is produced. If we assume $s(n)$ to represent a neutral, noise-free speech signal, then the acoustic signal at the microphone is written as

$$\left\{ s(n) \middle| \begin{array}{l} \text{WORKLOAD TASK} \\ \text{STRESS} \\ \text{LOMBARD EFFECT}\{d_1\} \end{array} \right\} + d_1(n). \qquad (1)$$

In general, the acoustic background noise $d_1(n)$ will also degrade the speech signal. Next, if the speech recognition system is trained with one microphone and another is used for testing, then microphone mismatch will cause a distortion. This can be modeled as a frequency mapping with the impulse response $h_{\text{MIKE}}(n)$. If the speech signal is to be transmitted over a telephone line or cellular phone channel, another level of distortion is introduced (modeled as either additive noise $d_2(n)$, or a frequency distortion with impulse response $h_{\text{CHANNEL}}(n)$). Further noise could also be present (or modeled) at the receiver $d_3(n)$. Therefore the neutral noise-free distortionless speech signal $s(n)$, having been produced and transmitted under adverse conditions, is transformed into the degraded signal $y(n)$, as shown in (2), at the bottom of the next page.

Manuscript received February 2, 1994; revised December 1, 1994. The associate editor coordinating the review of this paper and approving it for publication was Prof. Richard J. Mammone.

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

IEEE Log Number 9410232.

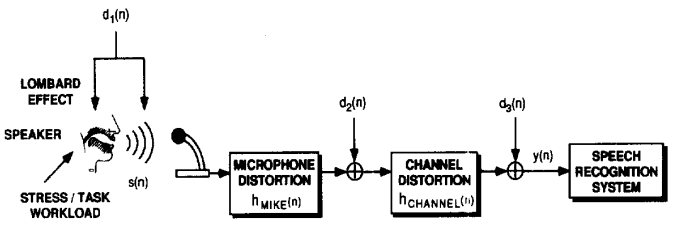


Fig. 1. General framework for the types of distortion which may be addressed for robust speech recognition.

Approaches for robust recognition can be summarized under the following three research areas

i) better training methods
ii) improved front-end processing
iii) improved back-end processing or robust recognition measures.

These recognition approaches have in turn been used to address improved recognition of speech in

a) noisy environments
b) Lombard effect
c) workload task stress or speaker stress
d) microphone or channel mismatch.

To formulate automatic speech recognition algorithms which are more effective in changing environmental conditions, it is important to understand the effects of noise on the acoustic speech waveform, the acoustic-phonetic differences between normal speech and speech produced in noise, and the acoustic-phonetic differences between normal speech and speech produced under stressed conditions. Several studies have shown distinctive differences in phonetic features between normal and Lombard speech [4], [43], [25], [28], [23], [54], [56], and speech spoken in noise [19]. Other studies have focused on variation in speech production brought on by task stress or emotion [25], [28], [29], [35], [5]. The primary purpose of these studies has been to improve the performance of recognition algorithms in noise [42], Lombard effect [43], [37], [55], stressed speaking styles [46], [51], [9], noisy Lombard effect [28], [31], [33], [34], [7], and noisy stressful speaking conditions [53], [28], [30], [34], [35].

Approaches based on improved training methods include multi-style stress training [46], [51], simulated stress token generation [5], training and testing in noise [14], and others [42]. Improved training methods have increased recognition performance, however as suggested by Juang, recognition performance degrades as testing conditions drift from the original training data. A solution which has been suggested is fast update methods for recognition models under varying noise environments.

Another area which has received much attention is front-end processing/speech feature-estimation for robust recognition. Here, many studies have attempted to uncover that speech representation which is less sensitive to various levels and types of additive, linear filtering, or convolutional distortion. For example, some studies focus on identifying better speech features [39], [37], or estimation of speech features in noise [27], or processing to obtain better speech representations [38]. If the primary distortion is additive noise, then a number of speech enhancement algorithms can be used such as

i) short-time spectral amplitude estimation (spectral subtraction) [3]
ii) model based optimal filtering (Wiener filtering) [45], [17], [18], [26], [27]
iii) adaptive noise canceling [60].

Other front-end processing incorporates feature processing for noise reduction and stress equalization [35], [7], [33], additive and convolutional noise [38].

The last approach for robust recognition is in improved back-end processing or robust recognition measures. Such processing methods refer to changes in the recognizer formulation such as the hidden Markov model (HMM) structure, or developing better models of noise within the recognizer [59]. Robust recognition measures are included here because they seek to project either the test data space closer to the trained recognition space, or trained space toward test space [36], [49], [8]. Studies related to robust metrics include those processing for linear filtering or microphone mismatch distortion [47].

B. Outline of Paper: Robust Recognition Under Additive Noise

In this study, we focus on the area of robust features with respect to additive background noise. The method under consideration is a previously formulated scheme based on sequential maximum *a posteriori* (MAP) estimation of the speech waveform and speech modeling parameters followed by the application of inter and/or intra-frame spectral constraints between iterations [24], [27]. The paper is organized as follows. In Section II, three objective speech quality measures are discussed in the context of quality assessment for automatic recognition in noise. In Section III, a simple objective measure based MAP estimation approach for recognition performance is presented, followed by a discussion of constrained-iterative

$$y(n) = \left(\left(\left(\left\{ s(n) \middle| \begin{array}{l} \text{WORKLOAD TASK} \\ \text{STRESS} \\ \text{LOMBARD EFFECT} \{d_1\} \end{array} \right\} + d_1(n) \right) * h_{\text{MIKE}}(n) + d_2(n) \right) * h_{\text{CHANNEL}}(n) \right) + d_3(n) \qquad (2)$$

TABLE I
COMPARISON OF THE AVERAGE CORRELATION COEFFICIENT $|\hat{\rho}|$ BETWEEN OBJECTIVE AND SUBJECTIVE SPEECH QUALITY AS MEASURED BY COMPOSITE ACCEPTABILITY OF DAM [52]. CORRELATION COEFFICIENTS ARE FOR OVERALL DISTORTION *ALL:* 322 TYPES OF DISTORTION, AND SPECIFIC DISTORTION CLASSES: *WFCD:* 66 WAVEFORM CODER DISTORTIONS, *WBD:* 126 WIDEBAND DISTORTIONS FROM WAVEFORM CODERS AND CONTROLLED DISTORTIONS, *NBD:* 36 NARROWBAND FREQUENCY DISTORTIONS, *FDD:* 36 DIFFERENT COLORED FREQUENCY DEPENDENT DISTORTIONS

OBJECTIVE SPEECH QUALITY MEASURE	*D A.M. CORRELATION*		
SNR		$\|\hat{\rho}\|_{\text{WFCD}} = 0.24$	
Segmental SNR		$\|\hat{\rho}\|_{\text{WFCD}} = 0.77$	
LPC Based Measures			
Linear Predictor Coefficient	$\|\hat{\rho}\|_{\text{ALL}} = 0.06$		
Log Predictor Coefficient	$\|\hat{\rho}\|_{\text{ALL}} = 0.11$		
Linear Reflection Coefficient	$\|\hat{\rho}\|_{\text{ALL}} = 0.46$		
Log Reflection Coefficient	$\|\hat{\rho}\|_{\text{ALL}} = 0.11$		
Log Area Ratio (LAR)	$\|\hat{\rho}\|_{\text{ALL}} = 0.62$	$\|\hat{\rho}\|_{\text{WBD}} = 0.65$	$\|\hat{\rho}\|_{\text{NBD}} = 0.91$
Log Likelihood Ratio (IS: Itakura-Saito)	$\|\hat{\rho}\|_{\text{ALL}} = 0.59$	$\|\hat{\rho}\|_{\text{WBD}} = 0.61$	$\|\hat{\rho}\|_{\text{NBD}} = 0.80$
Weighted Spectral Slope (WSSM:Klatt)	$\|\hat{\rho}\|_{\text{ALL}} = 0.74$	$\|\hat{\rho}\|_{\text{WBD}} = 0.61$	$\|\hat{\rho}\|_{\text{FDD}} = 0.90$

robust feature-estimation using the (Auto:I,LSP:T) algorithm in Section IV. This procedure is evaluated using CCDATA and TIMIT data for a number of noise conditions. The transformed (Auto:I,LSP:T) estimated features are then employed within a monophone model based speech recognition algorithm across actual noise conditions in Section VI. Finally, the MAP recognition rate estimator is evaluated in Section VII, and conclusions drawn in Section VIII.

II. NOISE AND OBJECTIVE SPEECH QUALITY

When noise is introduced into a speech utterance, its impact on speech quality is nonuniform. As a result, the impact of additive background noise on speech recognition performance will depend on how each phoneme for an input text sequence is effected. In this section, three objective speech quality measures are considered as a means of representing the impact of various additive noise sources on speech quality for recognition. A fourth measure based on SNR is also discussed. For this study, the focus is on distortion which primarily introduces an additive spectral mismatch into the frequency response of the speech signal across time. This, of course, only reflects a small portion of the potential types of distortion which may be introduced for voice recognition applications. It is suggested that the change in quality could be used to predict the robustness of feature enhancement/estimation front-ends for automatic speech recognition in noise.

The choice of an objective measure rests on its ability to predict quality for a particular distortion. Research has been performed in the formulation of objective quality measures for coding [52], [11], [58], and the application of these measures to speech enhancement [24] and recognition [13], [20]–[22], [36], [40]. In one study of over 2000 different objective measures using the multidimensional diagnostic acceptability measure (DAM) [57], several measures were identified which have a noticeable degree of correlation to subjective quality for a broad range of distortions [52].

The following three objective measures are considered in this study, log-likelihood ratio (Itakura–Saito) $d_{\text{IS}(j)}$, log-area-ratio $d_{\text{LAR}(j)}$, and the weighted-spectral slope measure (Klatt) $d_{\text{WSSM}(j)}$. Table I summarizes correlation results from [52] for the objective measures considered, along with several others for comparison. Here, $|\hat{\rho}|_{\text{ALL}}$ refers to correlation with composite acceptability of the DAM across all 322 tested distortions and therefore reflects the objectives measures overall performance. Some of these distortions included

i) coding (e.g., ADM, ADPCM, LPC, MPLPC, etc.)
ii) controlled distortion (e.g., additive noise, clipping, echo, lowpass filtering, etc.)
iii) frequency variant (e.g., narrowband noise, pole distortions, etc.).

Other correlation values reflect more specific distortion classes such as

i) WBD: wideband distortion
ii) NBD: narrowband distortion
iii) WFCD: waveform coder distortion
iv) FDD: frequency dependent distortion.

From this study, the log-likelihood ratio (IS) resulted in one of the higher LPC based quality measures, and of those measures employing an aural model, the weighted spectral-slope measure (Klatt) possessed the highest correlation coefficient with subjective quality. This table shows that the three selected measures possess good degrees of correlation across overall distortion types; and higher degrees of correlation for narrowband and frequency dependent distortion classes.

The method in which each measure is estimated will now be considered. The measures have the property that if the degraded/feature-enhanced and original speech spectra are identical, the resulting measure is zero. Each distance measure represents a measure of distortion between a frame of original and degraded/processed speech. Global quality measures are obtained by averaging individual frame distances d_i over a sentence or database.

One of the more successful quality measures based on the magnitude spectrum is the log likelihood ratio [40], [20], [10]. This measure is based on the dissimilarity between all-pole models of the reference s and processed speech $\acute{s}$ as follows:

$$d_{\text{IS}(i)} = d_i(\vec{a}_s, \vec{a}_{\acute{s}}) = 10 \cdot \log_{10}\left[\frac{\vec{a}_s \mathbf{R}_{\acute{s}} \vec{a}_s^T}{\vec{a}_{\acute{s}} \mathbf{R}_{\acute{s}} \vec{a}_{\acute{s}}^T}\right] \tag{3}$$

where $\vec{a}_s$, $\vec{a}_{\acute{s}}$ are the all-pole model coefficients from the ith frame of the original and feature-enhanced signals respectively, and $\mathbf{R}_{\acute{s}}$ is the corresponding autocorrelation matrix

TABLE II
SUMMARY OF CRITICAL BAND FREQUENCY LOCATIONS AND BANDWIDTHS FOR CALCULATION OF THE WEIGHTED SPECTRAL SLOPE MEASURE

Critical Band Center Frequency Locations and Bandwidths (Hz)							
f_1	50.0000	b_1	70.0000	f_{14}	1148.30	b_{14}	140.423
f_2	120.000	b_2	70.0000	f_{15}	1288.72	b_{15}	153.823
f_3	190.000	b_3	70.0000	f_{16}	1442.54	b_{16}	168.154
f_4	260.000	b_4	70.0000	f_{17}	1610.70	b_{17}	183.457
f_5	330.000	b_5	70.0000	f_{18}	1794.16	b_{18}	199.776
f_6	400.000	b_6	70.0000	f_{19}	1993.93	b_{19}	217.153
f_7	470.000	b_7	70.0000	f_{20}	2211.08	b_{20}	235.631
f_8	540.000	b_8	77.3724	f_{21}	2446.71	b_{21}	255.255
f_9	617.372	b_9	86.0056	f_{22}	2701.97	b_{22}	276.072
f_{10}	703.378	b_{10}	95.3398	f_{23}	2978.04	b_{23}	298.126
f_{11}	798.717	b_{11}	105.411	f_{24}	3276.17	b_{24}	321.465
f_{12}	904.128	b_{12}	116.256	f_{25}	3597.63	b_{25}	346.136
f_{13}	1020.38	b_{13}	127.914	Max. Freq. : 4000.0 Hz, 25 Bands			

of the feature-enhanced signal. The measure has been shown to assign a high weight when an error due to mismatch in formant location occurs, and a lower weight for error in spectral valleys. This is desirable, since the auditory system is more sensitive to errors in formant location, then to the spectral bandwidths or valleys between peaks.

Other objective measures can be formed based on linear prediction coefficients (LPC). A variety of coefficients can be used to represent the LPC model, though it has been shown [52] that of all LPC based measures, the log-area-ratio measure has the highest correlation with subjective quality. The log-area-ratio parameters are obtained from the reflection coefficients r_i as

$$\mathrm{LAR}_i = \log\left[\frac{A_{i+1}}{A_i}\right] = \log\left[\frac{1+r_i}{1-r_i}\right] \quad 1 \le i \le P \quad (4)$$

where P represents the order of the LPC analysis. The objective quality measure is formed as follows:

$$d_{\mathrm{LAR}(i)} = \left|\frac{1}{M}\sum_{k=1}^{M}[\mathrm{LAR}_{s_i} - \mathrm{LAR}_{\hat{s}_i}]^2\right|^{\frac{1}{2}} \quad (5)$$

where LAR_{s_i} is the set of log-area-ratios from the original speech (ith frame), $\mathrm{LAR}_{\hat{s}_i}$ the log-area-ratios for the feature-enhanced frame, and M the number of parameters for each frame. Since spectral distance measures have perhaps been the most widely investigated quality measure, many variations exist [6], [10], [20], [21], [52].

The third measure, entitled weighted spectral-slope measure (WSSM) by Klatt [44], is based on an auditory model in which overlapping filters of progressively larger bandwidth are used to estimate the smoothed short-time speech spectrum. The filter bank bandwidths are chosen to be proportional to the ear's critical bands so as to give equal perceptual weight to each band. Once the filter bank is formed (see Table II), the measure finds a weighted difference between the spectral slopes in each band. The magnitude of each weight reflects whether the band is near a spectral peak or valley, and whether the peak is the largest in the spectrum. Klatt computes the weight for each spectrum separately, then averages the two sets of weights to obtain $w_a(k)$. Next, a per-frame spectral distance measure in decibels is found using

$$d_{\mathrm{WSSM}(i)} = K_{\mathrm{spl}}(K - \hat{K}) + \sum_{k=1}^{25} w_a(k)\{S(k) - \hat{S}(k)\}^2 \quad (6)$$

where K, $\hat{K}$ are related to overall sound pressure level of the reference and processed signals. K_{spl} is a parameter which can be varied to increase overall performance. The resulting measure is therefore sensitive to differences in formant location, yet less sensitive to differences in the height of those peaks or differences in spectral valleys. Next, we consider the various sources of distortion in this study for robust speech recognition, and present the final measure of recognition performance estimation.

The last measure is based on the well-known signal-to-noise ratio (SNR). Two definitions of SNR are employed. The first will be referred to as SNR or $\mathrm{SNR}_{\mathrm{GLOBAL}}$, and is represented as

$$\mathrm{SNR}_{\mathrm{GLOBAL}} = 10\log\left(\frac{\sum_{n=1}^{N} s^2(n)}{\sum_{n=1}^{N} d^2(n)}\right) \quad (7)$$

where the summation is over the entire utterance. The second SNR definition will always be referred to as NIST SNR, and is determined as,

$$\mathrm{SNR}_{\mathrm{NIST}} = 10\log\frac{\text{Peak Signal Power}}{\text{Mean Noise Power}}, \quad (8)$$

where power refers to the signal variance computed over 20-ms windows.[1]

The basic framework for introducing distortion into the CCDATA speech corpus will now be considered. CCDATA consists of 1737 files from two-way spontaneous telephone conversations concerning credit cards. A subset of CCDATA was selected and degraded with preselected levels of degrading background noise as

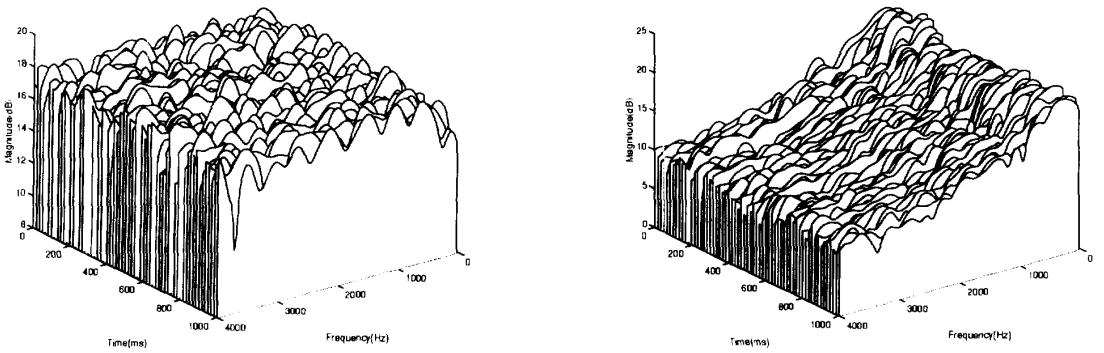
$$y(n) = s(n) + g \cdot d(n) \quad (9)$$

where g is adjusted to achieve an average SNR level of 5, 10, or 15 dB as in (7). The nine noise sources summarized in

[1]Formulation of the NIST SNR measure is discussed in further detail in documentation provided by National Institute of Standards and Technology (NIST) with speech corpus data base (i.e., *snr.doc*); also see [12].

TABLE III
SUMMARY OF NOISE SOURCES CONSIDERED FOR ROBUST FEATURE ENHANCEMENT AND RECOGNITION EVALUATION. *Stationarity* REFERS TO A SUBJECTIVE MEASURE OF NOISE STATIONARITY BASED ON FIRST AND SECOND MOMENT ANALYSIS (1: WIDE-SENSE STATIONARITY TO 10: NONSTATIONARY)

NOISE SOURCES FOR ROBUST RECOGNITION EVALUATION		
Noise	Stationarity	Description
WGN	1	computer generated white Gaussian noise
FLN	1	noise from a flat communications channel
SUN	1	noise recorded from the cooling fan of a Sun 4/330 workstation
PS2	3	noise recorded from the cooling fan of an IBM PS-2/80 workstation
HEL	3	noise recorded from a helicopter fly-by
LCI	3	noise recorded from a large city
LCR	5	noise recorded from a large crowd
HWY	5	noise recorded inside an automobile traveling on an interstate highway at 60 miles/hour
BAB	9	noise recorded under multiple speaker babble conditions



(a) (b)

Fig. 2. Time versus power spectral response for two background noise distortions: (a) Flat noise; (b) highway noise.

Table III were considered in this study. Each noise source was sampled at 8 kHz. A brief first and second moment analysis across 4 seconds of noise data was also conducted to determine the degree of stationarity for each source. A subjective score of stationarity (i.e., 1: for wide sense stationary, to 10: nonstationary) was assigned to each noise source based on this analysis. Sample time versus power spectral estimates are shown for two of the background noise sources in Fig. 2. Noise sources are grouped as broadband (WGN, FLN), low frequency band (SUN, PS2, HEL, HWY), time varying colored (LCI, LCR, BAB).

III. OBJECTIVE MEASURE BASED MAP RECOGNITION ESTIMATION

In this section, we formulate a simple estimation procedure of output recognition rate based on observed objective speech measures. The motivation for such an estimator is clear; since it provides a basis for obtaining a quantitative measure of recognition performance for a selected speech feature, given a noise source and SNR band, *without* the need of additional recognition simulation. Such an estimator would also indicate the impact level of a noise source for a given speech feature based recognizer, or the level of robustness for a speech feature under consideration for that adverse environment.

Assume that the following is known for a speech recognition application in a given adverse environment:

i) an example noise sequence is available
ii) an operating SNR range is specified
iii) that the noise source possesses a known level of stationarity.

Further assume, that a predetermined speech corpus exists which is representative of the required recognition task. Let the random variable x represent the resulting objective measures from a degraded or feature-enhanced speech corpus. Also, let the random variable y represent the resulting recognition rate for a finite test set of utterances from that corpus. We assume that in the region of interest, both random variables are Gaussian distributed.[2] With this, the joint probability density function (pdf) for objective measure x and recognition rate y is

$$f_{xy}(x_o, y_o) = \frac{1}{(2\pi)|\mathcal{C}_{xy}|^{1/2}} e^{-\frac{1}{2}[x-m_x, y-m_y]\mathcal{C}_{xy}^{-1}\left[\begin{smallmatrix} x-m_x \\ y-m_y \end{smallmatrix}\right]}. \tag{10}$$

The following conditional pdf can then be written

$$\begin{aligned} f_{y|x}(y_o|x_o) &= \frac{f_{xy}(x_o, y_o)}{f_x(x_o)} \\ &= \frac{1}{\sqrt{2\pi}\sigma_y\sqrt{(1-\rho_{xy}^2)}} \\ &\quad \times e^{\frac{-1}{2\sigma_y^2(1-\rho_{xy}^2)}\{m_y+\rho_{xy}\frac{\sigma_y}{\sigma_x}(x-m_x)\}} \end{aligned} \tag{11}$$

where ρ_{xy}, σ_y, m_y, and m_x are estimated from our previous simulations. The resulting maximum *a posteriori* (MAP)

[2]The joint Gaussian assumption is reasonable if the performance of recognition algorithm does not vary too wildly for the given noise source and SNR band.

INTER-FRAME CONSTRAINTS

Step 1. Convert LPC predictor coefficients $\hat{\vec{a}}_{i,n}$ (frame n, iteration i) to LSP position $\vec{p}_{i,n}$ and difference $\vec{d}_{i,n}$ parameters.

Step 2. (i) Perform voiced or unvoiced/silence speech frame classification (based on frame energy E_n, and energy threshold $E_{V/UV}$).
(ii) Maintain count L_j of consecutive frames below $E_{V/UV}$.
(iii) If L_j reaches L_{MAX}, classify all subsequent frames below $E_{V/UV}$ as silence/noise.
(iv) If a single frame rises above $E_{V/UV}$, reset frame energy count L_j to zero and continue.

Step 3. *LSP Position Smoothing Constraints:*
In general, LSP position parameters $\vec{p}_{i,n}$ more closely related to formant location.
- Apply a speech adaptive smoothing constraint across a triangular base of support which is position (i.e., frequency) dependent as follows:

$$\tilde{p}_{i,n}(j) = \sum_{k=-N(p_j)}^{N(p_j)} H(E_n, p_j) \cdot \left[1 - \frac{|k|}{W(E_n, p_j)}\right] \cdot p_{i,n}(j) \quad \forall j = 1, \ldots, 5$$

Here, $H(E_n, p_j)$ represents smoothing window height, $W(E_n, p_j)$ the weighted window base width. (In general, both dependent on frequency and frame energy.)

Step 4. Update even LSPs using constrained position parameters as:

$$LSP_{i,n}(2j+1) = \tilde{p}_{i,n}(j).$$

Step 5. *LSP Difference Constraints:*
LSP difference parameters $\vec{d}_i$ more closely related to formant bandwidth.
- If $|d_i| < d_{min}$, then LPC analysis has underestimating a pole bandwidth. Therefore, set

$$LSP_{i,n}(2j) = \tilde{p}_{i,n}(j) + d_j = \tilde{p}_{i,n}(j) + \text{SIGN}(d_j) d_{min}.$$

- If $|d_i| > d_{max}$, then model pole corresponds to overall vocal tract shaping. Allow bandwidth to varying naturally. Set LSP difference parameter as,

$$LSP_{i,n}(2j) = \tilde{p}_{i,n}(j) + d_j = \tilde{p}_{i,n}(j) + \text{SIGN}(d_j) \frac{\Delta p_{j,j+1}}{2}.$$

- If $d_{min} \le |d_i| \le d_{max}$, then model pole corresponds to a formant. Apply smoothing constraint only to corresponding position coefficient $p_{i,n}(j)$.

Step 6. Form new set of constrained model parameters $\tilde{\vec{a}}_{i,n}$ from constrained $\vec{LSP}_{i,n}(j)$ parameters.

Fig. 3. Procedure for application of inter-frame spectral constraints.

estimator for recognition rate given an objective measure is found as

$$\hat{y}_{\text{MAP}}(x) = \arg\max_y f_{y|x}(y_o \mid x_o). \tag{12}$$

Since we assume that the recognition rate and objective measure are both Gaussian distributed, this resulting MAP estimator $\hat{y}_{\text{MAP}}(x)$ is equivalent to the mean square error estimator $\hat{y}_{\text{mse}}(x) = \int y f_{y|x}(y_o \mid x_o) dy$. The resulting MAP estimation equation which maximizes the conditional pdf in (11) for recognition rate is

$$\hat{y}_{\text{MAP}}(x) = \left[\rho_{xy} \frac{\sigma_y}{\sigma_x}\right]_b x + \left[m_y - \rho_{xy} \frac{\sigma_y}{\sigma_x} m_x\right]_a \tag{13}$$

where $[\cdot]_a$ and $[\cdot]_b$ are used for notation purposes. This estimator will be evaluated in Section VII.

IV. (AUTO:I,LSP:T) CONSTRAINED FEATURE-ESTIMATION

In this section, we consider the feature enhancement based recognition system. It should be noted that this algorithm was previously formulated as one of a number of constrained iterative methods for speech enhancement [27]. Consider a noise corrupted speech vector $\vec{y}_{\gamma_j}$. It is assumed that the input speech signal can be modeled by a set of all-pole model parameters $\vec{a}$ and gain g. A sequential maximum-a-posteriori (MAP) estimation of the clean speech vector $\vec{S}_O$ is obtained given the noisy input speech $\vec{Y}_O$, followed by MAP estimation of the model parameters $\vec{a}$ given $\hat{\vec{S}}_O$, where $\hat{\vec{S}}_O$ is the result of the first MAP estimation. The process iterates between the following two MAP estimation steps:

i) $\text{MAX}\, p(\hat{\vec{a}}_i \mid \hat{\vec{S}}_{O,i}, \vec{Y}_O; g, \vec{S}_{\bar{I}})$ which gives $\hat{\vec{a}}_i$

ii) $\text{MAX}\, p(\hat{\vec{S}}_{O,i} \mid \hat{\vec{a}}_i, \vec{Y}_O; g, \vec{S}_{\bar{I}})$ which gives $\hat{\vec{S}}_{O,i}$

until a convergence threshold is reached. This general sequential MAP estimation approach was first considered for white Gaussian noise conditions by Lim and Oppenheim [45]. In the current feature enhancement approach, constraints are applied to $\hat{\vec{a}}_i$ to ensure the following:

i) The all-pole speech model is stable;
ii) It possesses speech-like characteristics (e.g., poles are in reasonable places with respect to each other and the unit circle);
iii) The vocal tract characteristics do not vary by more than a prescribed amount from frame to frame when speech is present.

Fig. 4 illustrates a flow diagram of the sequential MAP estimation procedure.

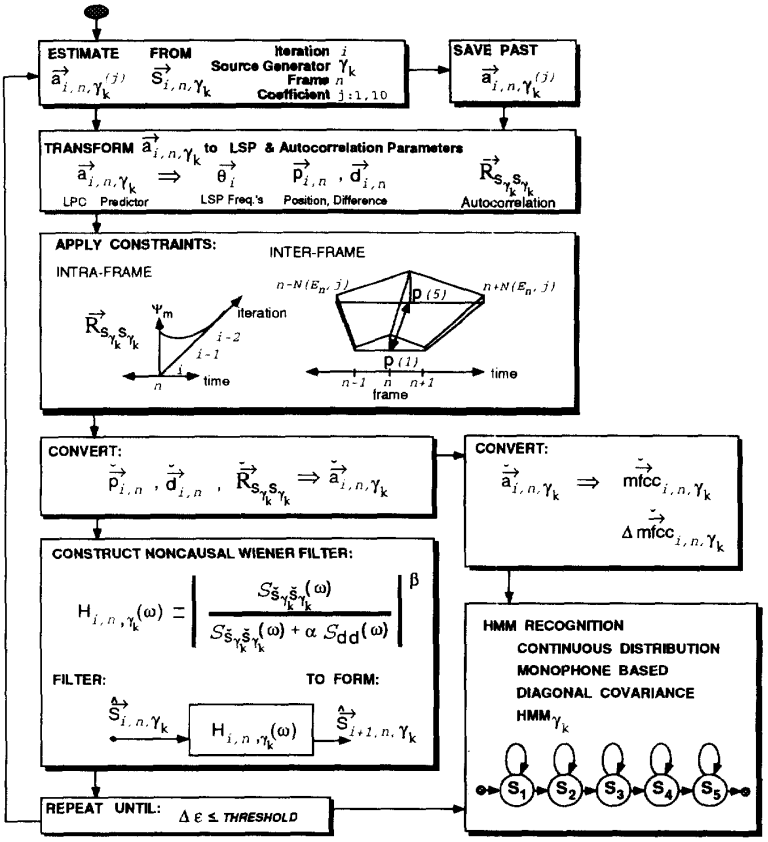


Fig. 4. Framework for the (Auto:I,LSP:T) feature-enhanced recognition algorithm.

Inter-frame spectral constraints are applied to line-spectral-pair parameters across time on a fixed-frame basis. These constraints are applied to ensure that vocal tract characteristics do not vary wildly from frame to frame when speech is present. This method allows constraints to be efficiently applied to speech model pole movements across time so that formants lay along smooth tracks. In order to increase numerical accuracy, reduce computational requirements, and eliminate inconsistencies in pole ordering across frames, the line spectral pair (LSP) transformation [40] is used to implement inter-frame constraint requirements across time. The procedure for application of LSP inter-frame constraints are outlined in Fig. 3.

Intra-frame spectral constraints are applied to autocorrelation parameters across iterations on a single-frame basis. Application of the intra-frame constraints is achieved by weighting the present set of autocorrelation parameters at time frame n with the same frame from previous iterations as

$$\hat{\mathcal{R}}_{s_{\gamma_j} s_{\gamma_j}}[k] = \sum_{m=0}^{M} \psi_m \mathcal{R}_{s_{\gamma_j, i-m} s_{\gamma_j, i-m}}[k] \qquad (14)$$

with the condition that $\sum_{m=0}^{M} \psi_l = 1$ (here, m represents autocorrelation terms from previous iterations). Given the present autocorrelation estimate, this weighting process re-introduces a controlled level of distortion from previous iterations which slows the rate of improved estimation for phoneme sections less sensitive to additive noise.

V. Feature-Enhancement Evaluation

In this section, performance of (Auto:I,LSP:T) feature enhancement is presented. First, evaluations are presented where the algorithm is adjusted to achieve improved subjective performance for a particular noise source. In the second section, the algorithm features are fixed and applied to the nine noise sources for robust feature enhancement/recognition evaluation. These results therefore, represent the default level of improvement for a general recognition algorithm.

A. TIMIT Sentence Evaluation

The general feature estimator possesses several algorithm values which may be adjusted for improved estimation in the presence of different noise sources. For example, the power exponent in the filter and weighting terms in the intra-frame spectral constraints work together to effect the choice for the best terminating iteration for enhanced features. In general, as the value of β is decreased (i.e., in the range $\beta \in [0.1, 0.5]$), the optimal terminating iteration increases. The best improvement in objective speech quality for each of the nine noise sources for a single sentence is shown in Fig. 5. Informal listening

TABLE IV
SUMMARY OF ITAKURA-SAITO (IS) QUALITY MEASURES ACROSS PHONEMES FOR 100 TIMIT SPEECH SENTENCES. DE.: DEGRADED WITH WGN (10DB SNR), FE.: FEATURE-ENHANCED (AUTO:I,LSP:T), Cnt.: FRAME COUNT

OBJECTIVE SPEECH QUALITY ACROSS AMERICAN PHONEMES								
Phoneme		**DE.**	**FE.**	*Cnt.*	*Phoneme*		**DE.**	**FE.**	*Cnt.*
CONSONANTS - nasals					*CONSONANTS - unvoiced stops*				
/m/	me	6.563	1.483	683	/p/	pan	2.228	0.890	508
/n/	no	8.225	1.323	1153	/t/	tan	1.233	0.701	542
/ng/	sing	7.990	1.820	159	/k/	key	2.118	0.982	559
/nx/	many	5.625	0.506	77	*CONSONANTS - voiced stops*				
/em/	problem	5.575	1.324	33	/b/	be	2.206	0.698	135
/en/	subtraction	9.224	1.925	135	/d/	dawn	1.164	0.817	186
/eng/	greasing	3.423	0.908	18	/g/	give	2.217	0.782	142
CONSONANTS - unvoiced fricatives					*CONSONANTS - closure stops*				
/s/	sip	0.664	1.126	1433	/tcl/	it pays	1.869	1.395	999
/th/	thing	1.009	0.836	203	/kcl/	pockets	2.013	1.221	655
/f/	fan	0.881	1.108	796	/bcl/	to buy	3.857	1.432	399
/sh/	show	0.905	0.964	673	/dcl/	sandwich	3.424	1.443	636
CONSONANTS - voiced fricatives					/gcl/	iguanas	3.055	1.327	241
/z/	zip	0.880	1.435	1054	/pcl/	accomplish	1.553	0.935	779
/zh/	garage	0.924	1.232	66	*CONSONANTS - glottal stop, flap*				
/dh/	that	3.338	1.038	270	/q/	allow	3.998	1.481	661
/v/	van	3.574	1.163	273	/dx/	put in	4.299	0.721	142
CONSONANTS - affricates					*CONSONANTS - unvoiced whisper*				
/jh/	joke	1.041	1.398	263	/hh/	had	3.532	1.054	143
/ch/	chop	1.276	1.506	336	*CONSONANTS - voiced whisper*				
					/hv/	you have	6.235	0.832	103
VOWELS - front					*DIPHTHONGS*				
/ih/	hid	1.478	0.426	947	/ay/	hide	1.509	0.300	1033
/eh/	head	1.745	0.375	856	/oy/	coin	3.519	0.654	171
/ae/	had	1.275	0.222	977	/ey/	pain	1.043	0.331	725
/ux/	to buy	1.975	0.584	636	/ow/	code	2.193	0.649	660
VOWELS - mid					/aw/	pout	1.727	0.325	288
/aa/	odd	2.748	0.615	1339	/iy/	new	1.581	0.887	1220
/er/	earth	6.244	1.319	562	*SEMIVOWELS - liquids*				
/ah/	up	2.088	0.480	625	/r/	ran	8.137	1.875	747
/ao/	all	4.528	1.291	750	/l/	lawn	3.783	1.672	1079
VOWELS - back					/el/	chemicals	4.667	2.488	356
/uw/	boot	3.622	1.250	197	*SEMIVOWELS - glides*				
/uh/	foot	2.164	0.534	116	/w/	wet	5.545	1.853	289
VOWELS - front schwa					/y/	you	1.743	1.142	318
/ix/	heed	2.785	0.646	1043	*Silence*				
VOWELS - back schwa					/# /	extended	1.700	0.947	5087
/ax/	a ton	3.418	1.014	628	/pau/	pause	2.552	1.284	175
VOWELS - retroflexed schwa					/epi/	epenthetic	4.358	2.930	98
/axr/	after	8.519	2.141	594					
VOWELS - voiceless schwa					**Overall**		2.750	1.028	36006
/ax-h/	subtraction	2.951	1.634	35	**Overall - /#/**		3.030	1.041	30919

tests and objective measure results show that quality improvement occurs for the following noise conditions: WGN, HEL, HWY, PS2, and SUN computer cooling fans, and (FLN) flat communications channel noise. Little improvement or change was noted for LCI, and LCR. Finally, the feature-estimation procedure introduced further distortion for BAB noise. This is to be expected, since the assumption of a stationary noise source has been violated for the single channel method.

Next, a detailed evaluation was considered across individual phonemes. Since phonetic label data is not available for the CCDATA corpus, TIMIT [1] sentence data was used for analysis. For this evaluation, 100 sentences were selected from a representative sampling of dialect across TIMIT. Each sentence was down-sampled to 8 kHz, and degraded with white Gaussian noise (WGN) at an SNR of 10 dB. Next, feature enhancement was applied to each degraded utterance. Using phonetic label data provided by NIST, individual frames were grouped into NIST labeled phonemes. The IS objective quality measure was obtained for each degraded and enhanced phoneme class (see Table IV). After application of (Auto:I,LSP:T) feature-enhancement, improvement in objective quality is obtained for 55 of the 61 NIST phonemes, with significant improvement for nasals, vowels, diphthongs, glides, and liquids. Fricatives and affricates comprised the six phonemes which did not see significant improvement. This however, does not adversely effect overall performance, since these phonemes were not originally effected by this noise source to the same extent as others (i.e., compare degraded fricatives which did not show improvement with overall test average). These results show that (Auto:I,LSP:T) is able to adapt its feature enhancement across changing vocal tract phoneme structure. It is thus suggested, that such improvement could contribute to improved recognition performance.

B. CCDATA Feature Enhancement Evaluation

Next, we consider the performance of the feature-enhancement algorithm for degraded CCDATA corpus. Results for several objective quality measures are reported to i)

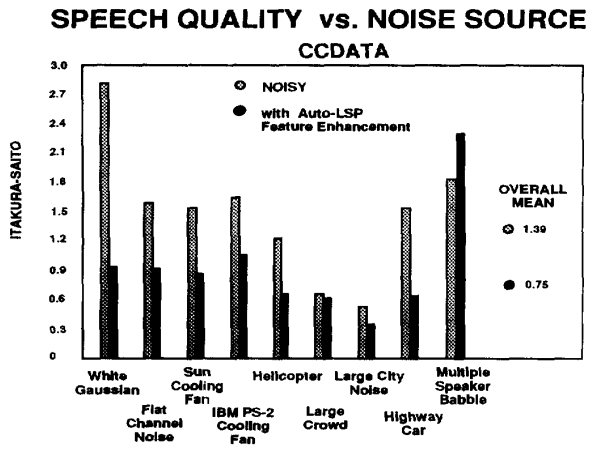


Fig. 5. Enhancement using the (Auto:I,LSP:T) enhancement algorithm for nine noise sources at 10 dB SNR. Improvement is shown with respect to IS distance measures.

illustrate the improvement in feature representation in varying levels and types of additive background noise, and ii) to provide the necessary *a priori* information to formulate a MAP estimator for output recognition performance. Though results have shown that adapting (Auto:I,LSP:T) processing characteristics such as terminating iteration, inter- and intra-frame constraint settings, and power exponent effect the resulting quality of the optimal features, an average algorithm configuration was used across all noise types and levels. Therefore, further improvement in the estimated features employing further knowledge of the noise source should improve the actual recognition rates. This has already been demonstrated using a phoneme class directed constrained iterative enhancement technique [2]. The results obtained by fixing the feature enhancement algorithm therefore represents the level of recognition performance improvement which should be expected from an unknown input noise source (i.e., no *a priori* noise training data).

Time Waveform Analysis: As discussed previously, noise influences speech quality and recognition feature parameters differently across time. To illustrate this, see the first IS objective measure plot of in Fig. 6 for a female CCDATA sentence degraded with WGN. Since WGN has a uniform frequency response, its' impact on speech quality and recognition feature representation will be approximately uniform. However, noise sources such as SUN, PS2, HEL, and HWY effect some phonemes more than others. This results in a nonuniform level of speech quality and speech feature representation for recognition. However, after application of the feature-enhancement algorithm, output speech quality is shown to be more uniform (see Fig. 6). Here the average IS measure is reduced from 2.80 to 0.87, with a significant reduction in IS variance. This result suggests that feature enhancement will increase the quality of extracted speech features for robust speech recognition. Next, we illustrate the effect of additive noise and feature enhancement with objective quality measures.

NIST SNR Analysis: Next, we consider the effect of noise and feature-enhancement on NIST SNR, since improvement in NIST SNR may be a more meaningful numerical measure of improvement for some researchers in robust speech recognition. For this evaluation, 60 CCDATA sentences were degraded at SNR_{GLOBAL} of 5, 10, and 15 dB, and processed with the feature-enhancement algorithm. Fig. 7 summarizes NIST SNR[3] improvement for nine noise conditions and 5–15 dB SNR range (i.e., each scatter plot entry represents a NIST SNR measure for a single degraded input and feature-enhanced output sentence). NIST SNR improvement however, is only meaningful if the background resting noise level is at least stationary with respect to frame-to-frame signal strength. Therefore, for short-time stationary noise sources such as flat communications channel, computer cooling fan, or highway, improvement in NIST SNR is meaningful. However, for such noise sources as large crowd noise (LCR) and background babble (BAB), NIST SNR measurements may be prone to error. This occurs since the non stationarity of the background noise will push individual signal frames into the speech region of the frame density function. Scatter plots between input and feature-enhanced NIST SNR are also shown for WGN, FLN, HWY, and HEL noise. Since an increase in the output feature-enhanced NIST SNR signifies improvement, all entries above the equal input–output line represent improvement. The results show that when additive noise is introduced at 5, 10, and 15 dB SNR values, NIST SNR is concentrated at 12.3, 16.9, and 21.6 dB. After employing (Auto:I,LSP:T) feature enhancement, average NIST SNR increased to 27.0, 31.0, and 35.8 dB. This represents an improvement of 14.1 to 14.7 dB in NIST SNR for the non speech-like noise distortions (i.e., excludes LCR and BAB). More importantly, some noise sources such as WGN, FLN, HEL, SUN, LCI, resulted in concentrated regions of output NIST SNR, while others (HWY, LCR, BAB) showed a wider range of output NIST SNR. This notion is clearly illustrated if NIST SNR scatter plots are compared for FLN and HWY noise sources in Fig. 7. The concentration for each of the 60 CCDATA sentences at each input SNR for WGN, FLN, and HEL suggest a high degree of confidence in (Auto:I,LSP:T) performance. Though the level of feature enhancement was not as consistent for HWY noise, this may be expected, since the impact of highway noise on speech quality varies more across individual phoneme classes than for FLN or WGN. The table of NIST SNR measures in Fig. 7 for the seven non speech-like noise sources show improvement at each input SNR level.

Objective Quality Analysis: It has been shown that feature enhancement improves resulting objective quality across non speech-like noise sources. The three objective quality measures (IS, LAR, WSSM) were obtained for the 60 CCDATA sentences used in the previous evaluation for both degraded and feature-enhanced conditions. These results will be used to obtain the necessary *a priori* information for the objective quality based MAP recognition rate estimator. In this section, we briefly discuss the three objective measure results.

Fig. 8 presents a partial summary of IS objective quality measures for four of the nine noise sources. Scatter plots show input degraded IS measures versus feature enhanced

[3]The signal-to-noise ratio as presented by National Institute of Standards and Technology (NIST) used here is based on their 'Second Method' which uses a 97% spread threshold in the energy histogram.

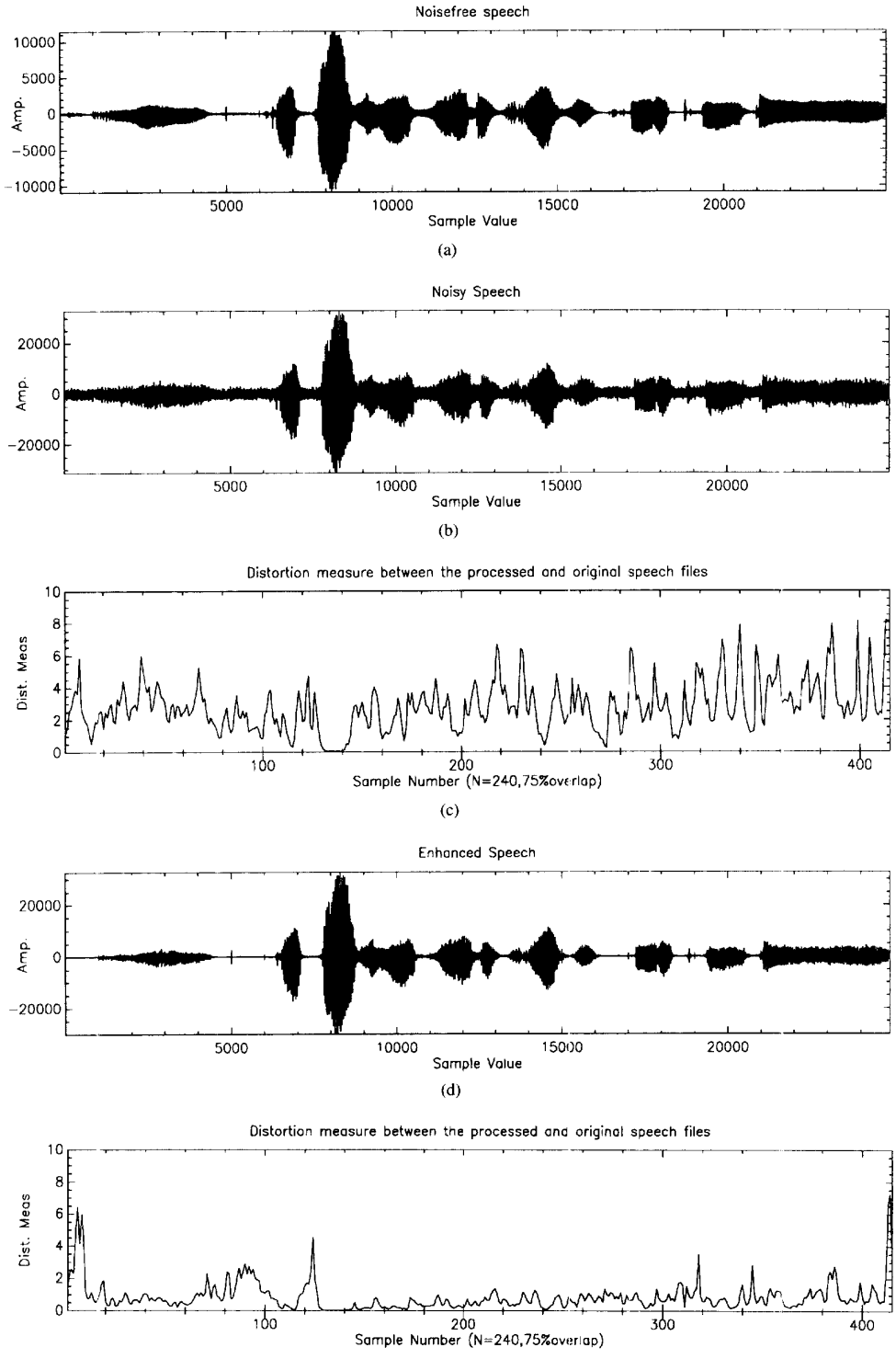


Fig. 6. Time waveforms of (a) an original CCDATA female speech sentence, (b) degraded with additive white Gaussian noise, and (d) (Auto:I,LSP:T) enhanced. Distortion as measured by frame-to-frame IS objective quality measures are shown for (c) noisy and (e) enhanced waveforms.

IS measures for the 60 sentence set over three SNR levels. Since objective measures quantify the level of distortion with respect to an original noise-free data set, entries below the diagonal (equal input versus output quality) line represent improvement using feature-enhancement. Also, the effect of additive noise on a degraded utterance reduces to zero as the measure approaches the origin. The results showed that for broadband noise sources such as WGN and FLN, feature enhancement provided a significant level of improvement. The level of improvement was greater at the 5 dB SNR input case. For narrowband noise sources such as HWY, SUN, and PS2, a measurable improvement was observed, however the variance in output quality was much higher than for WGN and FLN. This can be explained by the nonuniform impact these noise sources have on speech quality, and the fact that the (Auto:I,LSP:T) algorithm was not adjusted as the noise type was varied. Finally, for the speech-like distortions (LCR,BAB), some utterances were enhanced slightly at the lower SNR level for LCR noise; however, feature processing was not successful for background speaker babble (BAB). Since these noise sources violate the stationarity assumption made in using a single noise spectral estimate from the beginning of each sentence, it is not surprising that limited improvement occurs. The few isolated utterances for WGN and FLN which did not show improvement do contribute to a lower general mean IS improvement.

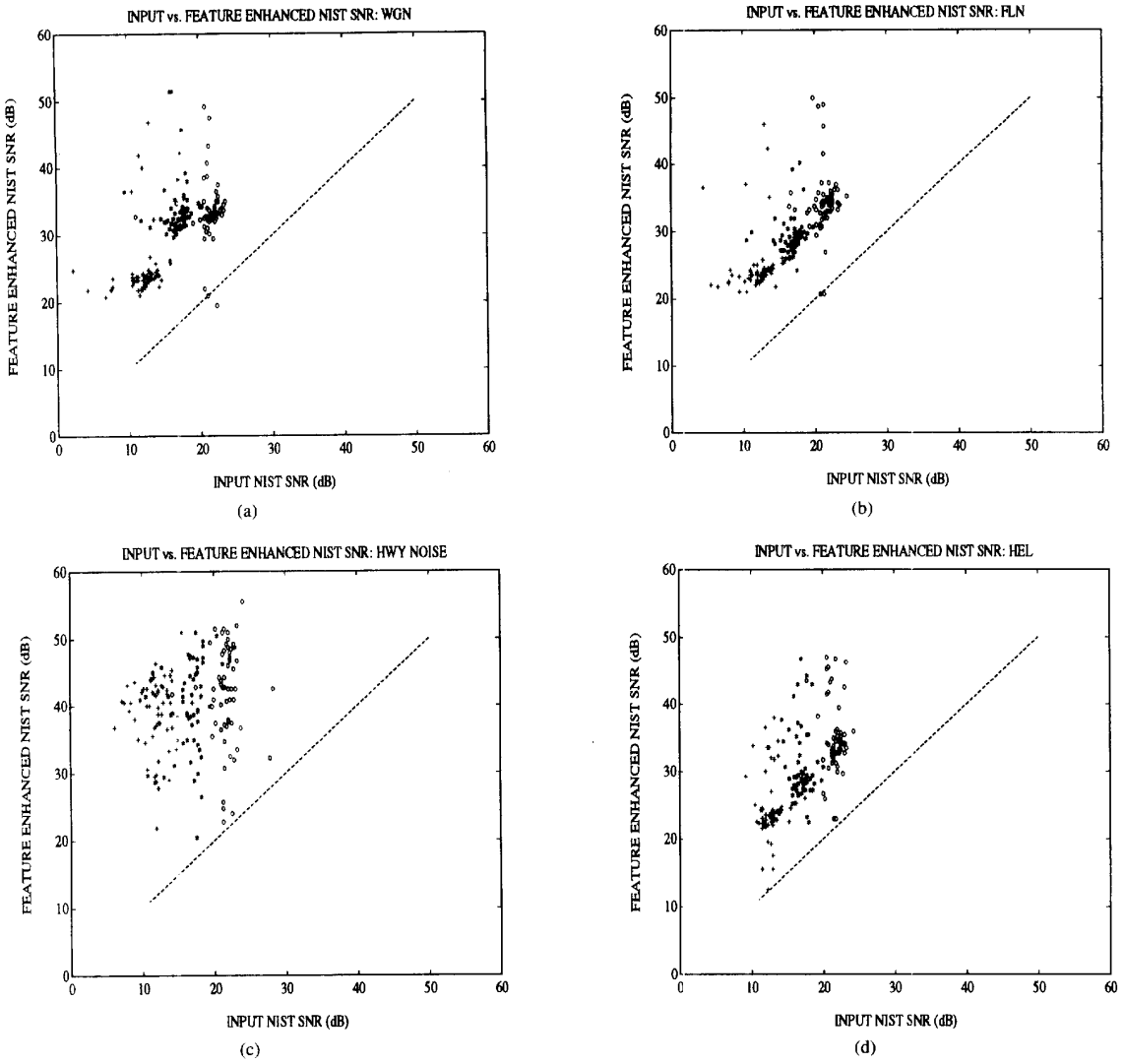
For the LAR measure, consistent improvement was observed for WGN and FLN noise sources. For some narrowband noise (HWY, PS2), a wide range of objective quality resulted; while others (SUN, HEL) gave more consistent results across SNR. In general, a more consistent level of improvement was observed for narrowband noise sources at lower SNR's, with a steady increase in measure variance as SNR increased. Similar results were obtained for WSSM, though measure variance levels were notably lower for most low frequency noise sources.

VI. Speech Recognition Using (Auto:I, LSP:T) Derived Features

The (Auto:I,LSP:T) algorithm has been shown to be useful in improving speech features for a limited isolated-word speech recognition task [26]–[28]. Extensions to this approach based on auditory constraints have also been reported [32], and shown to improve the quality of CELP coded speech [50]. Further feature enhancement methods have also been developed and applied to speech recognition in adverse (i.e., noisy stressful) conditions [33]–[35]. In this section, we consider the CCDATA corpus with respect to robust feature enhancement for recognition. As a test of the difficulty of the noise-free CCDATA corpus, an evaluation by BB&N[4] using their BYBLOS system produced a gender dependent phone recognition score of 36.2%, and a gender independent score of 30.1%. The BYBLOS recognition system uses discrete output densities with features based on normalized cepstra and delta cepstra. As shown in Fig. 4, feature-enhanced parameters $\hat{a}_{i,n,\gamma_k}$ were used to derive a combination of ten Mel-cepstral parameters $\hat{\text{mfcc}}_{i,n,\gamma_k}$, ten delta Mel-cepstral parameters $\Delta\hat{\text{mfcc}}_{i,n,\gamma_k}$, and energy. This conversion was performed after terminating feature processing. Cepstral mean removal was performed prior to testing. The recognizer was based on monophone hidden Markov models, with a single Gaussian per state, each possessing a diagonal covariance matrix. A complete noise-free training and testing evaluation using the entire CCDATA corpus with monophone HMM_{γ_k} models resulted in a correct monophone recognition score of 43.7%, with an accuracy of 39.7%. Noise-free trained monophone HMM_{γ_k}'s were used for all recognition evaluations for the remainder of this study. Since we wish to determine performance over a large number of noise conditions, a subset of those sentences used for noise free testing was extracted. A rank ordering of those sentences tested was performed, and the 60 best sentences were extracted for noisy recognition evaluation. This rank ordered set, produced a monophone recognition score of 54.2%, with an accuracy of 51.8%.

Next, the 60 sentence rank order test set was sequentially degraded with each noise source at global SNR's of 5, 10, and 15 dB, and submitted to the context-independent HMM monophone recognizer. Recognition scores summarized in Fig. 9 show that average recognition rates decreased across all noise sources. HWY resulted in the smallest decrease from the 54.2% noise-free rate, with average recognition rates of 30.8%–46.9%. Other noise sources such as WGN, FLN, and SUN introduced more pronounced losses in recognition, with rates ranging from 7.1 to 39.7%. Next, results for (Auto:I,LSP:T) feature enhancement for robust recognition are also summarized. Consistent recognition improvement (+8 to +13%) was observed for the following noise sources: WGN, FLN, SUN, PS2, LCR. A scatter plot for individual SNR levels revealed consistent performance for these noise sources as SNR varied from 5 dB to 15 dB. Recognition performance improvement ranged from 13 to 20% at 5 dB, 4–8% at 15 dB. Feature enhancement for helicopter (HEL) and large city (LCI) noise sources performed well at 5 dB SNR, with recognition improvements ranging from 13 to 21%. However, this improvement decreases as SNR increases to 15 dB. Little change (increase or decrease) in recognition rate was observed for highway (HWY) noise. A measurable decrease in recognition rate was observed for feature enhancement under background babble (BAB) noise conditions. The associated bar graph illustrates a representative sample of the improvement in recognition using feature enhancement. Finally, means and standard deviations in recognition rate for degraded and feature-enhanced conditions across the noise sources are also summarized for 5, 10, and 15 dB SNR (BAB noise not included in these calculations). With feature enhancement, context-independent recognition scores increased +15.2% for 5 dB SNR (i.e., from 14.2 to 29.4%), +9.59% for 10 dB SNR (i.e., 29.1 to 38.7%), and +2.9% for 15 dB SNR (i.e., from 40.3 to 43.2%). Feature enhancement processing significantly reduced the variability in recognition across noise sources. This can seen by the fact that recognition standard deviation $\sigma_{\text{RECOG}}(\text{NOISES}, \text{SNR}_i)$ was cut in half for each SNR level.

[4]These results were obtained by P. Jeanrenaud of BB&N as part of the DoD Workshop on Robust Speech Processing, Rutgers University CAIP Center, July 1993.



CREDIT CARD DATABASE ENHANCEMENT RESULTS (NIST SNR, dB)										
Avg. Input	WGN		FLN		HWY		HEL		SUN	
SNR_{GLOBAL}	DE	EN	DE	EN	DE	EN	DE	EN	DE	EN
5dB	11.7	24.7	11.9	24.8	12.0	39.0	12.8	24.6	12.7	25.2
10dB	16.8	33.9	16.8	29.2	16.5	40.3	17.0	30.7	17.3	29.4
15dB	21.3	33.6	21.3	34.4	21.8	41.9	21.7	35.0	21.9	34.4
Avg. Input	PS2		LCI		LCR		BAB		AVG-7	
SNR_{GLOBAL}	DE	EN	DE	EN	DE	EN	DE	EN	DE	EN
5dB	12.3	25.7	12.6	25.1	12.8	26.7	27.9	33.6	12.3	27.0
10dB	17.0	23.3	17.2	30.2	17.4	31.9	21.4	39.1	16.9	31.0
15dB	21.6	35.9	21.7	35.2	22.3	35.9	17.3	39.6	21.6	35.8

Fig. 7. Improvement in NIST SNR for (a) WGN, (b) FLN, (c) HWY Noise, (d) HEL using 60 Credit Card sentences across 11 noise sources and three average input SNR levels. AVG-7 indicates average NIST SNR value for the seven nonspeech noise sources (i.e., LCR large crowd noise, and BAB: multiple speaker babble noise scores were excluded).

If statistics are collected for degraded and feature-enhanced recognition scores across the eight nonspeech noise sources (excluding BAB), and three SNR's (i.e., 24 degrading noise conditions), average recognition increased by +9.22% (27.9 to 37.1%), with a corresponding decrease in variability as represented by standard deviation in recognition from 12.4 to 6.5. These findings suggest that (Auto:I,LSP:T) can be used to increase speaker independent continuous speech recognition performance in a wide range of noise conditions. It should be noted that the particular HMM recognizer was quite basic. If triphone models, and/or multiple Gaussian mixtures per state were included in the recognizer formulation, higher overall recognition rates could be achieved for an actual working system in adverse conditions.

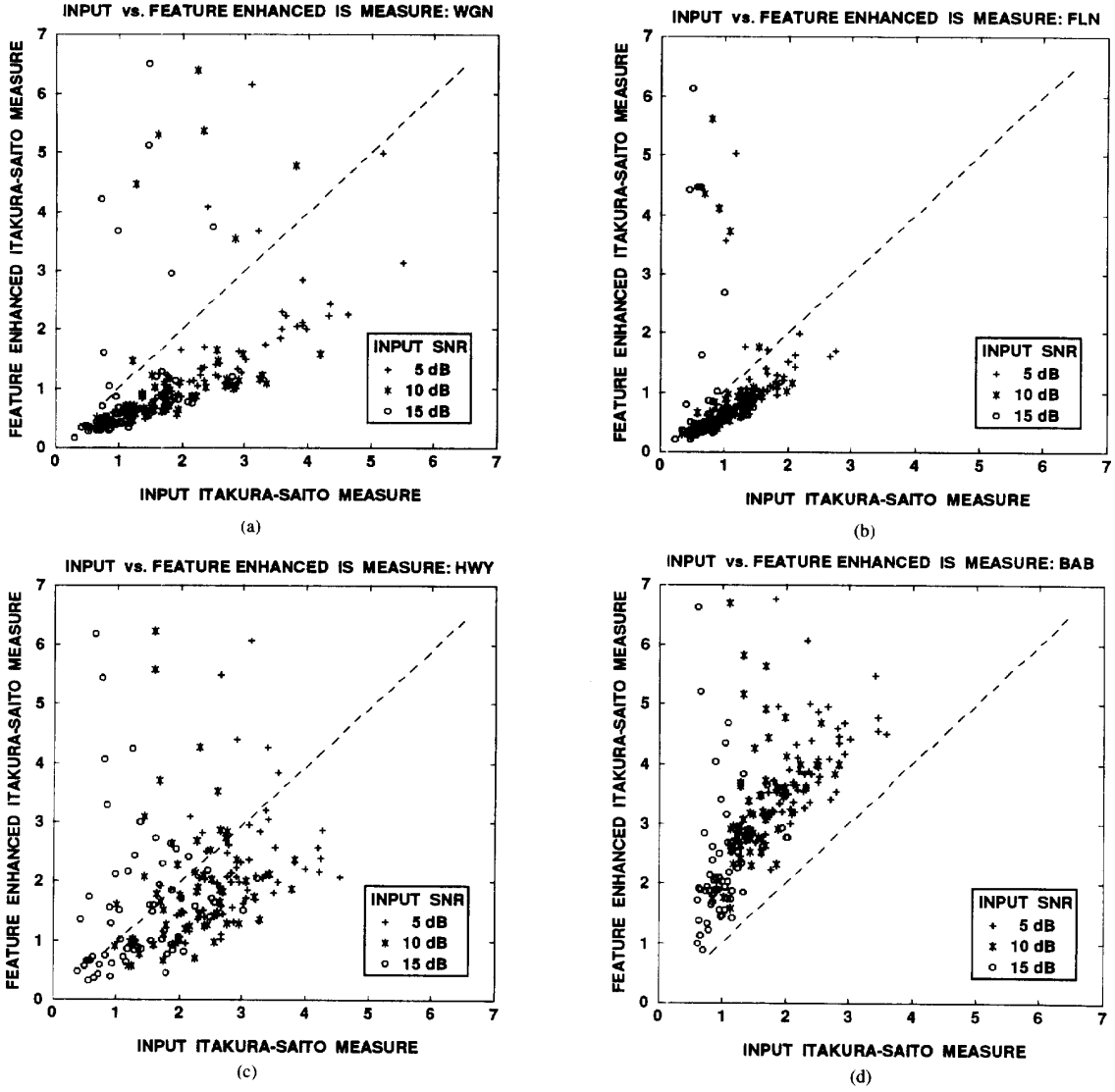
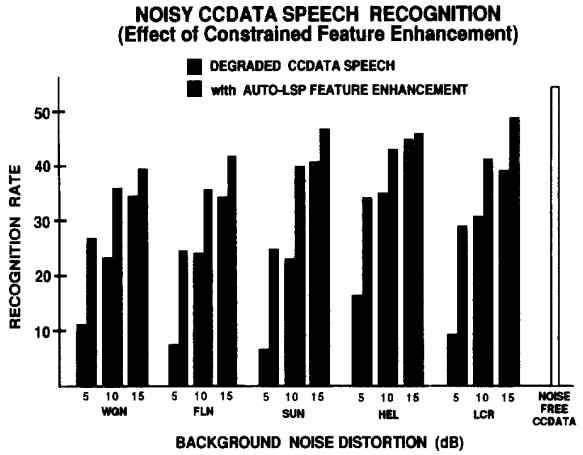


Fig. 8. Scatter plots of sample input degraded (DE) and Auto-LSP feature-enhanced (FE) Itakura–Saito (d_{IS}) objective quality measures using 60 Credit Card sentences across three average input SNR levels: (a) WGN, (b) FLN, (c) HWY, (d) BAB.

VII. Evaluation of Objective Measure Recognition Estimator

Objective speech quality measures and NIST SNR have been used to illustrate the change in speech quality as additive background noise conditions are varied. These measures have also been used to quantify the change in quality after (Auto:I,LSP:T) feature processing. Next, the recognition rate MAP estimator is evaluated for each of the nine noise sources, for each of the three objective speech quality measures and NIST SNR. Fig. 10 shows an example of recognition rate versus IS objective quality measure for degraded and feature-enhanced white Gaussian noise CCDATA conditions. Note that only mean values for degraded and feature-enhanced IS and recognition rate are shown. Clearly, there exists a relationship between the IS measure and recognition performance. The average mean-square error for each objective measure and NIST SNR were computed and summarized in Table V. The results show that across the degraded speech conditions (nine noise sources, 5–15 dB SNR band), the WSSM based MAP estimator provides the best estimate of recognition performance. After (Auto;I,LSP:T) processing, the IS based MAP estimator proved to be the best, while the WSSM based estimator lost performance. The NIST SNR measure consistently came in last across the nine noise sources in both degraded and feature-enhanced conditions. If we wish to determine the best MAP estimator using either degraded or feature-enhanced parameters for recognition, then the IS based estimator is the best overall predictor of recognition rate, while the NISTSNR estimator is the poorest predictor. Since the IS measure proved to be the better estimator, we choose



CREDIT CARD DATABASE MONOPHONE RECOGNITION RESULTS							Overall Mean Performance (5,10,15dB)		Change
NOISE TYPE	*5dB*		*10dB*		*15dB*				
	noisy	*enhanced*	*noisy*	*enhanced*	*noisy*	*enhanced*	m_{DE}	m_{FE}	in %
WGN	11.02	26.69	23.86	36.10	34.68	39.84	23.2	34.2	+11.0%
FLN	8.29	24.37	23.86	35.89	33.87	41.15	22.0	33.8	+11.8%
SUN	7.08	24.57	23.26	39.13	39.74	45.60	23.3	36.4	+13.0%
PS2	17.80	31.45	27.60	36.30	35.49	42.67	26.9	36.8	+9.9%
HEL	17.09	32.86	33.87	42.26	44.49	45.10	31.8	40.1	+8.3%
LCI	12.03	32.86	31.24	41.66	43.88	43.88	29.1	39.5	+10.4%
LCR	9.91	29.32	30.94	41.05	43.38	46.81	28.1	39.1	+11.0%
HWY	30.84	33.27	38.12	37.01	46.92	40.85	38.6	37.0	-1.5%
BAB	13.55	11.93	25.28	18.71	38.42	29.12	25.7	19.9	-5.8%
$m_{RECOG}(\bar{D}, SNR_i)$	14.2	29.4	29.1	38.7	40.3	43.2			
$\sigma_{RECOG}(\bar{D}, SNR_i)$	7.7	3.8	5.4	2.7	5.1	2.5			
Overall Degraded Results: $m_{RECOG}(\bar{D}, \mathrm{SNR}) = 27.89$, $\sigma_{RECOG}(\bar{D}, \mathrm{SNR}) = 12.39$									
Overall Feature-Enhanced Results: $m_{RECOG}(\bar{D}, \mathrm{SNR}) = 37.11$, $\sigma_{RECOG}(\bar{D}, \mathrm{SNR}) = 6.55$									

Fig. 9. Mean monophone recognition rate across three input SNR levels (5, 10, 15 dB), for degraded input and with (Auto:I,LSP:T) feature enhancement. The noise-free monophone rate was 54.2%. The nine noise sources are as follows: WGN-white Gaussian; FLN-flat communications channel; SUN-Suncomputer fan; PS2-IBM PS-2 computer fan; LCR-large crowd noise; LCI-large city noise; HEL-helicopter noise; HWY-automobile highway noise; BAB-multiple speaker babble noise. Calculation of $m_{\mathrm{RECOG}}(\mathrm{NOISES}, \mathrm{SNR}_i)$, $\sigma_{\mathrm{RECOG}}(\mathrm{NOISES}, \mathrm{SNR}_i)$, $m_{\mathrm{RECOG}}(\mathrm{NOISES}, \mathrm{SNR})$, and $\sigma_{\mathrm{RECOG}}(\mathrm{NOISES}, \mathrm{SNR})$ excluded BAB noise results.

TABLE V
SUMMARY OF THE AVERAGE MEAN SQUARE ERROR FOR EACH RECOGNITION RATE MAP ESTIMATOR ACROSS NINE NOISE SOURCES. DEGRADED, FEATURE-ENHANCED, AND TOTAL ERROR IS SHOWN FOR EACH MEASURE BASED ESTIMATOR

A COMPARISON OF OBJECTIVE MEASURE BASED RECOGNITION ESTIMATORS			
MEASURE	*DEGRADED AVG. ERROR*	*ENHANCED AVG. ERROR*	*TOTAL AVG. ERROR*
IS d_{IS}	0.7620	0.4089	0.5855
LAR d_{LAR}	1.0656	0.8828	0.9742
KLATT d_{WSSM}	0.6193	1.7238	1.1715
NIST SNR	1.4730	4.7050	3.0890

to summarize its performance across the nine noise sources in Table VI. In this table, entries a and b correspond to terms in the MAP estimator (13). Consistent performance is observed for the IS based MAP estimator for most noise sources under degraded speech conditions. A notable loss in performance is seen for BAB and LCR noise conditions, a result attributed to the nonstationarity of the noise source. For feature-enhanced conditions, uniform performance also results, with a decrease in error for all but WGN and HEL noise sources. Though we conclude that the IS based estimator provides good performance for recognition rate estimation, it is hypothesized that a composite estimator based on several measures may result in a better overall prediction of recognition rate.

VIII. CONCLUSION

In order to improve the environmental robustness of speech recognition in adverse conditions, a constrained iterative feature-estimation algorithm (Auto:I,LSP:T), previously employed for speech enhancement, is considered which is shown to produce improved speech characterization in a wide range of actual noise conditions such as various computer fan noises, large crowd noise, and voice communications channel noise. A MAP estimation process was also formulated using one of four measures as a means of predicting changes in recognition performance at the signal extraction phase. The four measures considered included,

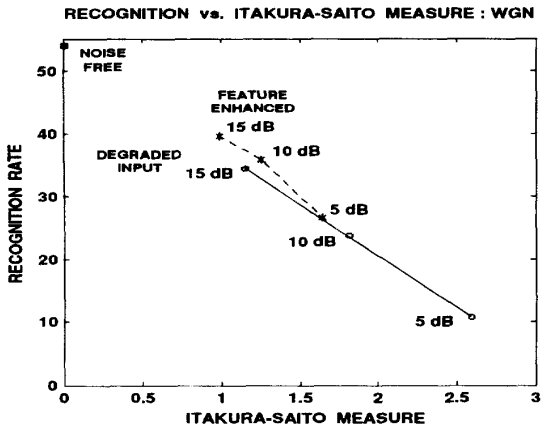


Fig. 10. Context-independent monophone recognition rate versus Itakura–Saito objective quality measures for degraded and feature-enhanced (Auto:I,LSP:T) processing. Results are for WGN at 5–15 dB SNR.

TABLE VI
SUMMARY OF MAP ESTIMATION PARAMETERS FROM (13) FOR THE ITAKURA-SAITO d_{IS} BASED RECOGNITION ESTIMATOR ACROSS NINE NOISE SOURCES. CORRESPONDING TERMS FOR THE MAP ESTIMATOR ACROSS EACH NOISE SOURCE FOR DEGRADED AND FEATURE-ENHANCED CONDITIONS ARE SHOWN. MEAN SQUARE ERROR IS SUMMARIZED FOR EACH ESTIMATOR

d_{IS} MEASURE BASED RECOGNITION RATE ESTIMATION						
NOISE TYPE	*DEGRADED*			*ENHANCED*		
	a	*b*	*error*	*a*	*b*	*error*
WGN	53.4970	-16.4400	0.0080	60.5080	-20.4380	0.6350
FLN	54.6390	-31.2130	0.6900	86.6220	-54.9940	0.2290
HWY	60.9210	-10.4020	0.1370	59.5780	-11.6310	0.0750
BAB	57.6520	-19.5670	0.9160	55.6300	-11.1200	0.1400
PS2	53.1980	-11.1000	0.1870	65.0150	-12.6190	0.1320
HEL	65.9180	-17.2840	0.8200	69.6050	-12.9230	1.4050
LCI	72.8430	-53.4640	0.8820	87.6950	-36.7060	0.3860
LCR	76.7860	-52.6170	2.3830	110.4630	-50.3140	0.0010
SUN	64.4680	-36.9620	0.8350	94.1970	-41.9160	0.6770

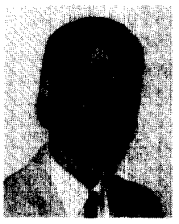
i) NIST SNR
ii) Itakura–Saito log-likelihood
iii) log-area-ratio
iv) weighted-spectral slope measure.

A context-free, continuous distribution, monophone based hidden Markov model algorithm was used for recognition valuation and objective measure analysis. Evaluations based on the Credit Card corpus showed that feature enhancement provides a consistent level of recognition improvement for broadband, and low-frequency colored noise sources. As the assumption of stationarity breaks down for a given noise source, the ability of feature enhancement to improve recognition performance decreases. Finally, the log-likelihood based MAP estimator for output recognition rate was found to be the best predictor of recognition performance, while the NIST SNR based MAP estimator was found to consistently be the poorest recognition predictor across the nine noise sources under consideration. The results show that robust front-end feature enhancement can contribute to improved recognition performance in a variety of adverse recognition conditions, and that a measure of recognition performance can be derived at the feature extraction stage based on an objective measure.

REFERENCES

[1] "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database," National Institute of Standards and Technology (NIST), Gaithersburg, MD, (prototype as of Dec. 1988).
[2] L. M. Arslan and J. H. L. Hansen, "A minimum cost based phoneme class detector for improved iterative speech enhancement," in *IEEE ICASSP-94*, Adelaide, Australia, Apr. 1994, pp. 45–48, vol. 2.
[3] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," in *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27 pp. 113–120, Apr. 1979.
[4] Z. S. Bond and T. J. Moore, "A note on Loud and Lombard Speech," in *Proc. 1990 Int. Conf. Spoken Language Processing*, Kobe, Japan, Nov. 1990, pp. 969–972.
[5] S. E. Bou-Ghazale and J. H. L. Hansen, "Duration and spectral based stress token generation for HMM speech recognition under stress," in *IEEE ICASSP-94*, Adelaide, Australia, Apr. 1994, pp. 413–416, vol. 1.
[6] P. Breitkopf and T. P. Barnwell, "Segmentation preclassification for improved objective speech quality measures," in *Proc. 1981 IEEE ICASSP*, Atlanta, GA, Mar. 1981, pp. 1101–1104.
[7] D. A. Cairns and J. H. L. Hansen, "ICARUS: An Mwave based real-time speech recognition system in noise and lombard effect," in *ICSLP-92, Int. Conf. Spoken Language Processing*, Alberta, Canada, Oct. 1992, pp. 703–706.
[8] B. Carlson and M. Clements, "Speech recognition in noise using a projection-based likelihood measure for mixture density HMM's," in *Proc. 1992 IEEE ICASSP*, San Francisco, CA, Mar. 1992, pp. 237–240, vol. I.
[9] Y. Chen, "Cepstral domain talker stress compensation for robust speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-36, pp. 433–439, Apr. 1988.
[10] P. L. Chu and D. G. Messerschmitt, "A weighted Itakura–Saito spectral distance measure," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-30, no. 4, pp. 545–560, Aug. 1982.
[11] H. J. Coetzee and T. P. Barnwell, "An LSP based speech quality measure," in *Proc. 1989 IEEE ICASSP*, Glasgow, Scotland, May 1989, pp. 596–599.
[12] J. R. Cohen, "Application of an auditory model to speech recognition," *J. Acoust. Soc. Am.*, vol. 85 no. 6, pp. 2623–2629, June 1989.
[13] P. V. de Souza, "Statistical tests and distance measures for LPC coefficients," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, no. 6, pp. 554–559, Dec. 1977.
[14] B. A. Dautrich, L. R. Rabiner, and T. B. Martin, "On the effects of varying filter bank parameters on isolated word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-31, no. 4, pp. 793–806, Aug. 1983.
[15] J. Deller, J. Proakis, and J. H. L. Hansen, *Discrete Time Processing of Speech Signals*. New York: Macmillan, 1993 (Macmillan Series for Prentice-Hall).
[16] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Ann. Royal Stat. Soc.*, pp. 1–38, Dec. 1977.
[17] Y. Ephraim, D. Malah, and B. H. Juang, "Speech enhancement based upon hidden markov modeling," in *Proc. 1989 IEEE ICASSP*, Glasgow, Scotland, May 1989, pp. 353–356.
[18] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proc. IEEE*, vol. 80, no. 10, pp. 1526–1555, Oct. 1992.
[19] 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.
[20] A. H. Gray and J. D. Markel, "Distance measures for speech processing," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 380–391, 1976.
[21] R. M. Gray, A. Buzo, A. H. Gray, and Y. Matsuyama, "Distortion measures for speech processing," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, no. 4, pp. 367–376, Aug. 1980.
[22] F. Gurgen, S. Sagayama, and S. Furui, "Line spectrum pair frequency-based distance measures for speech recognition," in *Int. Conf. Spoken Language Processing*, Kobe, Japan, Nov. 1990, pp. 521–524.
[23] C. N. Hanley and D. G. Harvey, "Quantifying the Lombard effect," *J. Hearing Speech Disorders*, vol. 30, pp. 274–277, Aug. 1965.
[24] J. H. L. Hansen and M. A. Clements, "Iterative speech enhancement with spectral constraints," in *Proc. 1987 IEEE ICASSP*, Dallas, TX, Apr. 1987, pp. 189–192.
[25] ——, "Evaluation of speech under stress and emotional conditions," presented at *Proc. Acoust. Soc. Amer.*, 114th Meeting, H15, Miami, FL, Nov. 1987.

[26] ——, "Constrained iterative speech enhancement with application to automatic speech recognition," in *Proc. 1988 IEEE ICASSP*, New York, NY, Apr. 1988, pp. 561–564.
[27] ——, "Constrained iterative speech enhancement with application to speech recognition," *IEEE Trans. Signal Processing*, vol. 39, no. 4, pp. 795–805, Apr. 1991.
[28] J. H. L. Hansen, "Analysis and compensation of stressed and noisy speech with application to robust automatic recognition," Ph.D. Dissertation, Georgia Institute of Technology, p. 428, July 1988.
[29] ——, "Evaluation of acoustic correlates of speech under stress for robust speech recognition," in *IEEE Proc. of the Fifteenth Annual Northeast Bioengineering Conf.*, Boston, MA, Mar. 27–28, 1989, pp. 31–32.
[30] J. H. L. Hansen and M. A. Clements, "Stress compensation and noise reduction algorithms for robust speech recognition," in *Proc. 1989 IEEE ICASSP*, Glasgow, Scotland, May 1989, pp. 266–269.
[31] 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.
[32] J. H. L. Hansen, "Detection and recognition of key words under noisy, stressful conditions," Duke Univ. Tech. Rep. DSPL-92-2, Grant no. NSF-IRI-90-10536, Nat. Sci. Found., p. 248, Mar. 1992.
[33] J. H. L. Hansen and O. Bria, "Improved automatic speech recognition in noise and Lombard effect," in *EURASIP-92, Sixth Europ. Signal Processing Conf.*, Heverlee, Belgium, Aug. 1992, pp. 403–406.
[34] J. H. L. Hansen, "Morphological constrained enhancement with adaptive cepstral compensation (MCE-ACC) for speech recognition in noise and Lombard effect," *IEEE Trans. Speech Audio Processing*, vol. 2, no. 4, pp. 598–614, Oct. 1994.
[35] ——, "Adaptive source generator compensation and enhancement for speech recognition in noisy stressful environments," in *IEEE ICASSP-93*, Minneapolis, MN, Apr. 1993, pp. 95–98.
[36] 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.
[37] B. A. Hanson and T. Applebaum, "Robust speaker-independent word recognition using instantaneous, dynamic and acceleration features: Experiments with Lombard and noisy speech," in *Proc. 1990 IEEE ICASSP*, Albuquerque, NM, Apr. 1990, pp. 857–860.
[38] H. Hermansky, N. Morgan, and H. G. Hirsch, "Recognition of speech in additive and convolutional noise based on RASTA spectral processing," in *Proc. 1993 IEEE ICASSP*, Minneapolis, MN, Apr. 1993, pp. 83–86.
[39] M. J. Hunt and C. Lefebvre, "A comparison of several acoustic representations for speech recognition with degraded and undegraded speech," in *Proc. 1989 IEEE ICASSP*, Glasgow, Scotland, May 1989, pp. 262–265.
[40] F. Itakura, "Line spectrum representation of linear predictive coefficients of speech signals," *J. Acoust. Soc. Am.*, vol. 57, S35(A), 1975.
[41] F. Jelinek, "Continuous speech recognition by statistical methods," *Proc. IEEE*, vol. 64, pp. 532–556, Apr. 1976.
[42] B. H. Juang, "Speech recognition in adverse environments," *Comput. Speech Language*, vol. 5, pp. 275–294, 1991.
[43] 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.
[44] D. Klatt, "Prediction of perceived phonetic distance from critical-band spectra: A first step," in *Proc. 1982 IEEE ICASSP*, Paris, France, 1982, pp. 1278–1281.
[45] 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.
[46] 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.
[47] F. H. Liu, A. Acero, and R. M. Stern, "Efficient joint compensation of speech for the effects of additive noise and linear filtering," in *Proc. 1992 IEEE ICASSP*, San Francisco, CA, Mar. 1992, pp. 257–260, vol. I.
[48] E. Lombard, "Le Signe de l'Elevation de la Voix," *Ann. Maladies Oreille, Larynx, Nez, Pharynx*, vol. 37, pp. 101–119, 1911.
[49] 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. ASSP-37, pp. 1659–1671, 1988.
[50] S. Nandkumar, J. H. L. Hansen, and R. Stets, "A new dual-channel speech enhancement technique with application to CELP coding in noise," in *ICSLP-92, Int. Conf. Spoken Language Processing*, Alberta, Canada, Oct. 1992, pp. 527–530.
[51] D. B. Paul, "A speaker-stress resistant HMM isolated word recognizer," in *Proc. 1987 IEEE ICASSP*, Dallas, TX, Apr. 1987, pp. 713–716.
[52] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective Measures of Speech Quality*. Englewood Cliffs, NJ: Prentice-Hall, 1988.
[53] 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.
[54] 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.
[55] ——, "Robust recognition of Loud and Lombard speech in the fighter cockpit environment," in *Proc. 1989 IEEE ICASSP*, Glasgow, Scotland, May 1989, pp. 675–678.
[56] W. V. Summers, D. B. Pisoni, R. H. Bernacki, R. I. Pedlow, and M. A. Stokes, "Effects of noise on speech production: Acoustic and perceptual analyses," *J. Acoust. Soc. Amer.*, vol. 84, pp. 917–928, Feb. 1988.
[57] W. D. Voiers, "Diagnostic acceptability measure for speech communication systems," in *Proc. 1977 IEEE ICASSP*, Hartford, CT, May 1977, pp. 204–207.
[58] S. Wang, A. Sekey, and A. Gersho, "Auditory distortion measure for speech coding," in *Proc. 1991 IEEE ICASSP*, Toronto, Canada, May 1991, pp. 493–496.
[59] M. Wang and S. Young, "Speech recognition using hidden Markov model decomposition and a general background speech model," in *Proc. 1992 IEEE ICASSP*, San Francisco, CA, Mar. 1992, pp. 253–256, vol. I.
[60] B. Widrow, *et al.*, "Adaptive noise canceling: Principles and applications," *Proc. IEEE*, vol. 63, no. 12, pp. 1692–1716, Dec. 1975.



John H. L. Hansen (S'81–M'82–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. He received the M.S. and Ph.D. degrees in electrical engineering from Georgia Institute of Technology, Atlanta, in 1983 and 1988, respectively.

In 1988, he joined the faculty of Duke University, as an Assistant Professor in the Department of Electrical Engineering, and received a secondary appointment in the Department of Biomedical Engineering in 1993. He is the coordinator of Robust Speech Processing Laboratory in the Department of Electrical Engineering. Prior to joining the Duke faculty, he was employed by the RCA Solid State Division, Somerville, NJ (1981–82), and Dranetz Engineering Laboratories, Edison, NJ (1978–81). 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 robust speech recognition. His research interests span the areas of digital signal processing, analysis and physical 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 based speech modeling for robust recognition in noise, stress, and Lombard effect.

Dr. Hansen is the author of numerous papers and technical reports in the field of speech processing, and is coauthor of the textbook *Discrete-Time Processing of Speech Signals* (Prentice-Hall, 1993). He was the recipient of a National Science Foundation's Research Initiation Award in 1990, and has been named a Lilly Foundation Teaching Fellow. He has served as Chairman for the IEEE Communications and Signal Processing Society of North Carolina, Advisor for the Duke University IEEE Student Branch, and is presently serving as Associate Editor for IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING. He has also served as co-editor of the Oct. 1994 special issue on Robust Speech Recognition for that publication.

Levent M. Arslan was born in Besni, Turkey on September 2, 1968. He received the B.S. degree in electrical engineering from Bogazici University, Istanbul, Turkey in 1991, and the M.S. degree in electrical engineering from Duke University, Durham, NC, in 1993. During the summer of 1994, he was a visiting speech researcher at Texas Instruments, Dallas, TX.

He is currently working towards the Ph.D. degree in electrical engineering at Duke University. His research interests include digital signal processing, speech enhancement, speech analysis, and speech recognition in noisy environments.