Constrained Iterative Speech Enhancement Using Phonetic Classes

Amit Das, *Student Member, IEEE*, and John H. L. Hansen, *Fellow, IEEE*

***Abstract*—The degree of influence of noise over phonemes is not uniform since it is dependent on their distinct acoustic properties. In this study, the problem of selectively enhancing speech based on broad phoneme classes is addressed using Auto-(LSP), a constrained iterative speech enhancement algorithm. Multiple enhanced utterances are generated for every noisy utterance by varying the Auto-LSP parameters. The noisy utterance is then partitioned into segments based on broad level phoneme classes, and constraints are applied on each segment using a hard decision solution. To alleviate the effect of hard decision errors, a Gaussian mixture model (GMM)-based maximum-likelihood (ML) soft decision solution is also presented. The resulting utterances are evaluated over the TIMIT speech corpus using the Itakura–Saito, segmental signal-to-noise ratio (SNR) and perceptual evaluation of speech quality (PESQ) metrics over four noise types at three SNR levels. Comparative assessment over baseline enhancement algorithms like Auto-LSP, log-minimum mean squared error (log-MMSE), and log-MMSE with speech presence uncertainty (log-MMSE-SPU) demonstrate that the proposed solution exhibits greater consistency in improving speech quality over most phoneme classes and noise types considered in this study.**

***Index Terms*—Auditory masked threshold, Auto-LSP, constrained iterative speech enhancement.**

I. Introduction

NOISE is present in almost all environments where speech systems are used, and therefore the need arises for designing effective speech enhancement algorithms. The objective of any speech enhancement algorithm is to suppress background noise, improve perceived quality (subjective) and intelligibility (objective), reduce listener fatigue, and improve performance for automatic speech recognition or speaker identification systems. It is difficult to address all these objectives simultaneously in a single enhancement algorithm since this essentially means that noise should be suppressed in a way which does not introduce processing artifacts, musical noise, or speech distortions. Hence, enhancement algorithms can be broadly classified as perceptual centric [3], [5]–[8], [15], [16], [17] or speech systems centric [1]–[3], [9].

Earlier studies [1] have shown that degradation due to environmental background noise is nonuniform across various phoneme classes of speech. This can be attributed to two reasons: 1) Each phoneme class (and even individual phonemes within the class) has distinct acoustical properties characterized by its time waveform, frequency content, manner of articulation, place of articulation, type of excitation and stationarity or nonstationarity of the vocal tract configuration [12, Ch. 2]. 2) The structure of different noise types can be classified based on their degree of stationarity and their bandwidths. For example, in-vehicle wind noise is a slowly varying narrowband low pass noise while white Gaussian is a stationary broadband noise. For these two reasons, the impact of noise on a phoneme class is determined by the characteristics of both the phoneme class and noise type.

Several research efforts have been devoted on developing phoneme class-based enhancement algorithms. McAulay and Malpass [11] adopted a two-state soft-decision maximum-likelihood algorithm in which speech was classified into equally likely binary—silence and non-silence—states. The resultant clean speech maximum-likelihood spectral envelope estimator was a sum of the products of individual envelope estimators, given the noisy signal and knowledge of the state, and corresponding *a posteriori* probabilities of the states given the noisy signal. The individual spectral envelope estimators, given the noisy signal and knowledge of the state, were optimized in a minimum mean square error (MMSE) sense. In another study, Hansen and Arslan [4] used hidden Markov models (HMMs) to create 13 phoneme class models. Using the forward algorithm scoring procedure, conditional probabilities $p(\vec{X}|\lambda_i)$, $i = 1, 2, \ldots, 13$, were obtained where $\vec{X}$ represents the observation vector from noisy speech, and λ_i is the noisy speech HMM model for phoneme class i. The difference of the top two scores was weighted by the inverse of a cost function to evaluate a confidence measure. Enhancement was done selectively based on this measure.

Later, Wang and Brown developed the computational auditory scene analysis (CASA) model where the objective is to segregate target speech from interfering acoustic mixtures (for example, speech and noise). In one of their studies [28], the input signal was decomposed by passing it through a gammatone filterbank to mimic the response of the auditory filterbank. A time–frequency based analysis was done by using the correl-

Manuscript received April 08, 2011; revised October 03, 2011 and January 19, 2012; accepted January 20, 2012. Date of publication April 11, 2012; date of current version April 30, 2012. This work was supported in part by RADC under Contract FA8750-05-C-0029 and in part by the University of Texas at Dallas under the Distinguished University Chair in Telecommunications Engineering held by J. H. L. Hansen. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Hui Jiang.

A. Das was with the Center for Robust Speech Systems (CRSS), Department of Electrical Engineering, University of Texas at Dallas, Richardson, TX 75083-0688 USA. He is now with the Indian Institute of Technology, Madras 600036, India (e-mail: das_amit_ece@yahoo.com).

J. H. L. Hansen is with the Center for Robust Speech Systems (CRSS), Department of Electrical Engineering, University of Texas at Dallas, Richardson, TX 75083-0688 USA (e-mail: John.Hansen@utdallas.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2012.2191282

ogram which finds the autocorrelation at the output of each auditory filter. The correlogram is effective in separating the fundamental frequency (F0) of each acoustic mixture. Next, based on the assumption that different sources are represented by a block of desynchronized oscillators, a two-layer oscillator network was employed in which acoustic mixtures were tracked based on their F0s. Finally, acoustic waveforms for the mixtures were derived from the composite time–frequency regions.

A constrained iterative speech enhancement model (Auto-LSP[1] [1], [4]) is followed in this study. Although Auto-LSP has been successful in improving context-independent monophone recognition performance [2], there are certain inherent drawbacks present in the formulation.

While noise suppression for high-energy sections (vowels) of speech is significant, it is sometimes overly suppressed for low energy sections (fricatives, stops) at the selected terminating iteration resulting in the introduction of processing artifacts. These artifacts have a pronounced effect on the perceived quality for the entire utterance. Although the number of iterations can be reduced to minimize these artifacts, it will leave noise under suppressed for most high energy sections which does not alleviate the problem. Moreover, degradation due to noise is higher for phoneme sections that lie within the noise bandwidth than for those that lie outside this bandwidth. For example, highway noise in the 0–800 Hz frequency range degrades vowels (first formant, 300–800 Hz) more than unvoiced fricatives (>1500 Hz). Also, there is usually some level of audible residual noise in the enhanced speech due to errors caused during estimation of the model parameters and noise spectrum.

This study addresses these issues by introducing broad phoneme class-based hard and soft decision ROVER[2] solutions. In this approach, multiple enhanced utterances are generated at different enhancement levels for a given noisy utterance. The noisy utterance is partitioned into segments based on the phoneme class, and class specific constraints are applied on each segment. Hard or soft decisions are used to select the best enhanced segment from the set of enhanced utterances. The selected segment is used for reconstruction of the enhanced speech. Also, audible residual noise can be measurably reduced by integrating with the auditory masking threshold framework developed originally by Tsoukalas *et al.* [6].

The remainder of the paper is organized as follows. In Section II, a brief overview of the baseline Auto-LSP system is explained. The algorithm formulation of the ROVER approach with its hard-decision and soft-decision solutions are presented in Section III. Detailed experimental evaluations based on Itakura–Saito, segmental SNR, and perceptual evaluation of speech quality (PESQ) metrics and improvements over other baseline algorithms are reported in Section IV. In Section V, directions for future work along with the conclusions are summarized.

II. Iterative Speech Enhancement

Assuming that the noise is additive and statistically independent of the speech signal, the additive noise model can be given by

$$y(n) = s(n) + d(n) \tag{1}$$

where $y(n)$, $s(n)$, and $d(n)$ represent the realizations of zero mean random processes of noisy speech, clean speech, and noise, respectively at discrete time instant n.

If N samples are observed from (1) to constitute a frame of noisy speech, then the noisy observations and the corresponding clean speech representing the hidden observations can be denoted by $\vec{y}$ and $\vec{s}$, respectively. The power spectrum of the noisy speech P_y is assumed to follow an all-pole model parameterized by autoregressive coefficients (ARC) $\vec{a}$, where $\vec{a} = [a_1\ a_2\ \ldots\ a_P]$, and gain G. This model is given as

$$P_y(\omega) = \frac{G^2}{\left|1 - \sum_{p=1}^{P} a_p e^{-j2\pi\omega p/K}\right|^2} \tag{2}$$

where P is the order of the all-pole model, K is the size of the discrete Fourier transform (DFT), and ω is the frequency index such that $0 \le \omega \le K-1$. Values of P and K considered for this study are 10 and 256, respectively, and the sampling frequency was set to 8 kHz.

The objective is to maximize the joint probability density function $p(\vec{a}, \vec{s}, G|\vec{y}; \vec{s}_I)$ assuming Gaussian priors for the unknowns $\vec{a}$, $\vec{s}$, G. Here, $\vec{s}_I$ denotes the initial speech condition. This results in a set of nonlinear equations involving partial derivatives with respect to $\vec{a}$. To remove this nonlinearity, a linear suboptimal iterative sequential maximum *a posteriori* (MAP) estimation technique was proposed by Lim and Oppenheim [10] where instead of jointly estimating $\vec{s}$ and $\vec{a}$, they were determined in a two step approach. The problem simplifies to finding an estimate of clean speech $\hat{\vec{s}}_\gamma$ at iteration γ given the noisy speech $\vec{y}$ and a previous estimate of ARC $\hat{\vec{a}}_{\gamma-1}$. This is followed by estimating $\hat{\vec{a}}_\gamma$ given $\hat{\vec{s}}_\gamma$ which was obtained from the previous MAP step. The sequential MAP estimation procedure is summarized by

$$\text{Step 1}: \quad \max \quad p(\vec{s}_\gamma|\hat{\vec{a}}_{\gamma-1}, \vec{y}; G, \vec{s}_I) \quad \text{to give } \hat{\vec{s}}_\gamma \tag{3}$$

$$\text{Step 2}: \quad \max \quad p(\vec{a}_\gamma|\hat{\vec{s}}_\gamma, \vec{y}; G, \vec{s}_I) \quad \text{to give } \hat{\vec{a}}_\gamma \tag{4}$$

where gain G and the initial speech condition $\vec{s}_I$ are assumed to be known [10]. These two steps are carried out iteratively until a convergence criterion is met.

Since $p(\vec{s}_\gamma|\hat{\vec{a}}_{\gamma-1}, \vec{y}; G, \vec{s}_I)$ is Gaussian distributed, then it can be shown that the MAP solution to (3) is equivalent to the MMSE estimate given by

$$\hat{\vec{s}}_\gamma = E[\vec{s}_\gamma|\hat{\vec{a}}_\gamma, \vec{y}] \tag{5}$$

[1]There are several flavors of Auto-LSP as reported by Hansen and Clements [1]. The one which is used here is known as "Auto:I,FF-LSP:T". The "Auto-I" refers to the intra-frame constraints on autocorrelation lags across iterations (I) and "FF-LSP:T" refers to the fixed frame (FF) line spectral pairs (LSP) constraints across time (T).

[2]The term ROVER (Recognizer Output Voting Error Reduction) is a connotation to the NIST automatic speech recognition (ASR) system [23] which produces a composite ASR output when outputs from multiple ASR systems are available. Since the enhancement system addressed in this study uses outputs from multiple Auto-LSP systems, it is appropriate to address this system as a ROVER based enhancement system.

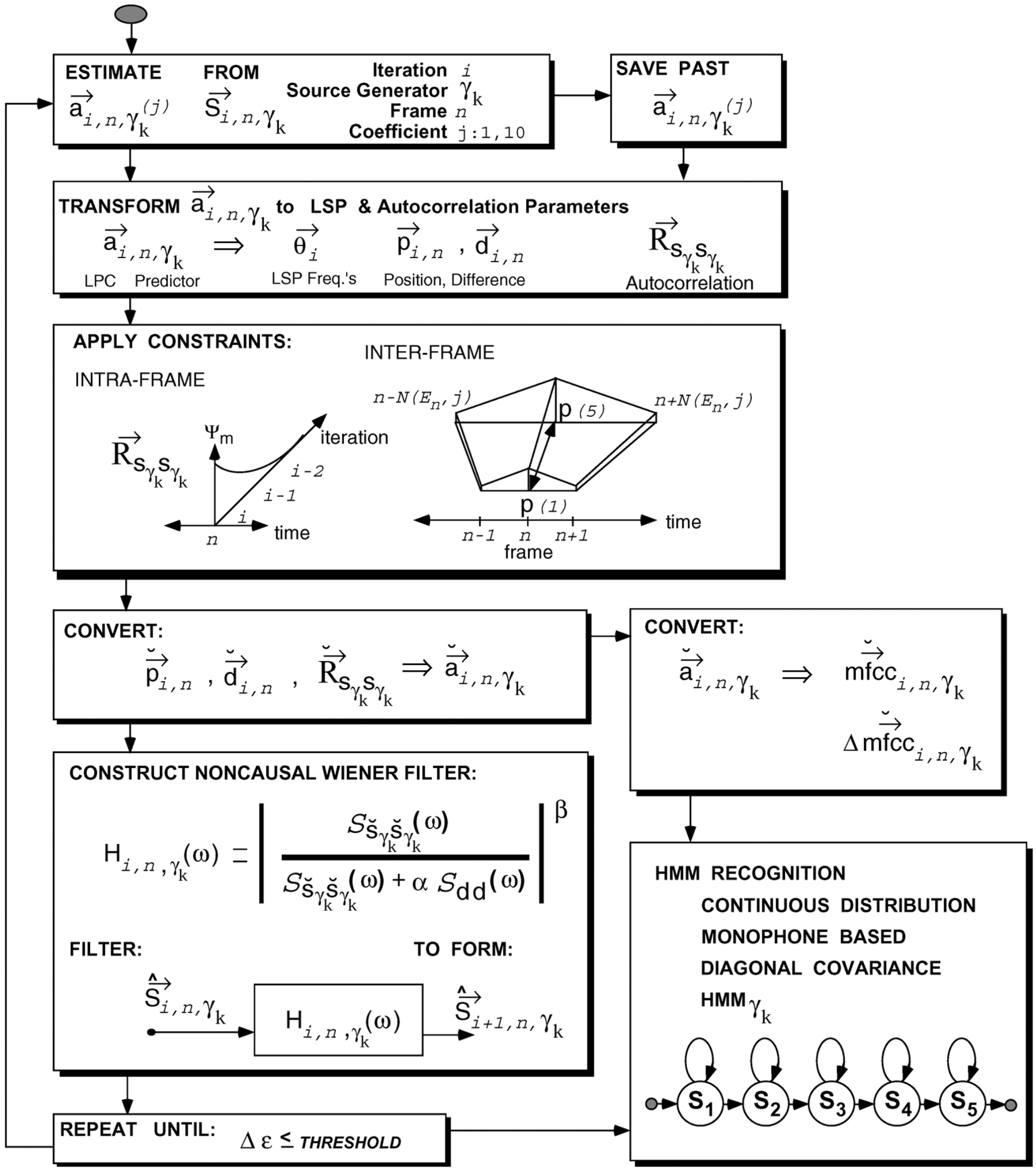

Fig. 1. Auto (I)-LSP(T) framework as a front-end application for feature recognition.

which can be obtained by Wiener filtering the noisy speech.

In the sequential MAP estimation method 1) formant bandwidths decreased and formant locations randomly shifted as the number of iterations increased, and 2) frame-to-frame pole jitter was observed resulting in ragged movement of poles (formants) across frames causing unnatural or metallic sounding speech as reported by Hansen and Clements [1].

To overcome these limitations, Sreenivas and Kirnapure [26] proposed a codebook based scheme to achieve faster convergence. In this scheme, a codebook was constructed from linear predictor coefficients (LPC) vectors derived from clean speech. At each iteration of the MAP steps, a clean speech LPC vector from the codebook was selected that was closest to the LPC vector of the clean speech estimate. The LPC vector of the clean speech estimate was replaced by the codebook entry closest to the clean speech LPC vector and used to construct the *a priori* power spectrum in (3). While such a procedure is expected to yield improvements in segmental SNR, no other perceptual quality metric was reported in [26]. SNR is a good indicator of noise suppression and a high SNR does not necessarily improve perceptual quality.

In the Auto-LSP approach (see implementation blocks in Fig. 1) proposed by Hansen and Clements [1], intra-frame and inter-frame constraints are dynamically applied to the autocorrelation lags and the position parameters of the line spectral pair (LSP) frequencies, respectively. The algorithm first extracts the autocorrelation lags from an input frame of noisy speech. Intra-frame constraints are applied over identical frame indices across iterations by applying weights to the present and past autocorrelation lags and updating the present autocorrelation lag with the weighted lag. This is represented as

$$^{k}R_{s_\gamma s_\gamma}[l] = \sum_{m=0}^{M} \psi_m {}^{k}R_{s_{\gamma-m} s_{\gamma-m}}[l] \tag{6}$$

where ${}^{k}R_{s_\gamma s_\gamma}[l]$ is the lth autocorrelation lag at the γth iteration for a given frame index k and $\sum_{m=0}^{M} \psi_m = 1$ is the weighting constraint over the previous M iterations. Here, $M = 1$ is considered to weight the autocorrelation lags for the current and previous iteration. The weights considered are $\psi_m \in \{0.8, 0.2\}$. This constraint ensures that the rate of convergence is more even across phoneme classes so that the all-pole model remains stable with restricted movement over iterations and possesses speech-like characteristics (i.e., poles do not migrate too close to the unit circle causing narrow bandwidths).

Next, LSPs and LPCs are derived from the autocorrelation lags. Based on the frame energies, each frame is classified as one of voiced, unvoiced, or noise-only frame. Constraints are applied to the position parameters of the LSPs using a weighted triangular window across frames. The weighted triangular window is evaluated based on the frame energy classification. No constraints are applied on difference parameters. With the application of inter-frame constraints, (3) becomes

$$\begin{aligned} &p(\vec{s}_\gamma | \hat{\vec{a}}_{\gamma-1,k}, \vec{y}; G, \vec{s}_I) \\ &= p\Big(\vec{s}_\gamma | \mathcal{F}\{\ldots, \hat{\vec{a}}_{\gamma-1,k-1}, \hat{\vec{a}}_{\gamma-1,k}, \hat{\vec{a}}_{\gamma-1,k+1}, \ldots; \omega\}, \vec{y}; G, \vec{s}_I\Big) \end{aligned} \tag{7}$$

where $\hat{\vec{a}}_{\gamma-1,k}$ is the ARC estimate at iteration $\gamma - 1$ and frame index k, and $\mathcal{F}\{\ldots; \omega\}$ denotes the constraint function that depends on frequency (ω). With this, the new LPCs are obtained from the constrained LSPs which are used to construct an improved all-pole speech spectrum estimate. After the application of Wiener filter (8), an enhanced speech spectrum estimate is obtained which is input for the next iteration. The algorithm is terminated after some stopping criterion which normally entails several iterations. For a detailed discussion on the implementation of Auto-LSP, we encourage the readers to follow the work of Hansen and Clements [1].

III. Algorithm Formulation

The main components of the ROVER framework are the creation of an archive of enhanced utterances, classification of broad phoneme classes (BPC), and hard or soft decision based synthesis. In this study, the 61 individual phonemes according to the NIST phonetic labels are grouped into one of the eight BPCs namely vowels, semivowels, nasals, affricates, fricatives, stops, closures, and silence.

A. Archive

An archive of enhanced frames is created using the Auto-LSP algorithm as presented in Section II. The Wiener filter, used in Auto-LSP, at each frequency component ω is given by

$$H(\omega) = \left(\frac{\hat{P}_s^{(\gamma)}(\omega)}{\hat{P}_s^{(\gamma)}(\omega) + \alpha \hat{P}_n(\omega)} \right)^\beta \tag{8}$$

where $\hat{P}_s^{(\gamma)}(\omega)$ is the LP based *a priori* power spectrum estimate of the clean speech at the γth iteration obtained using (2) after the application of inter-frame and intra-frame constraints. $\hat{P}_n(\omega)$ is the noise power spectrum estimate. With reference to the Wiener filter in Fig. 1, it is the same filter used in (8) where we have dropped the subscripts i, n, γ_k in $H_{i,n,\gamma_k}(\omega)$ for ease of representation. Since this filter is parameterized by the noise over-suppression factor (α), the exponent term (β) and the iteration (γ), it can be represented as $H(\alpha, \beta, \gamma)$, where the frequency term ω is also dropped for ease of representation. In Auto-LSP, up to four sets of filters $H(1,1,1)$, $H(1,1,2)$, $H(1,1,3)$, and $H(1,1,4)$ are used which limits the amount of enhancement that can be achieved. In the ROVER framework, a larger set of filters are used which sufficiently spans the entire enhancement space to obtain a broader range of enhancement levels. This range was chosen heuristically since it achieves minimal to maximal noise suppression for a wide range of noise types and levels. However, inter-frame and intra-frame constraint parameters remain constant although they can also be varied to achieve greater adaptation levels. The filter parameter set is comprised of

$$\begin{aligned} &\alpha \in [0.25, 0.5, \cdots, 2.0, 2.5, \cdots, 4.5] \\ &\beta \in [0.25, 0.5, \cdots, 1.25] \\ &\gamma \in [1, 2, \ldots, 4]. \end{aligned} \tag{9}$$

The total number of values taken by α, β, and γ are 13, 5, and 4, respectively. Hence, the parametric space spanned by (α, β, γ) can be viewed as a three dimensional Auto-LSP system. The minimum step size for α and β was determined from offline experiments. Hard and soft decisions, as explained later in Section III-C2 and Section III-C1, respectively, are made in this space to select the best sequence of enhanced segments based on their BPC knowledge.

TABLE I
Codebook Sizes of VQ Broad Phoneme Classifier

Class	Size	Class	Size
Vowel	64	**Fricative**	32
Semivowel	16	**Stop**	16
Nasal	16	**Closure**	16
Affricate	4	**Silence**	8

B. Phoneme Classifier

A set of LBG-based vector quantized (VQ) codebooks [13] are used to classify each short-time frame belonging to one of eight BPCs. Phoneme classification is critical in the ROVER framework because of its influence on the class dependent search constraints used in the decision making step discussed in the next section. Instead of using degraded speech as the test utterance for classification, enhanced speech obtained from $H(1,1,1)$ was used since it ensures the improvement in recognition rate for all phoneme classes. In the training phase, a total of 600 TIMIT tokens from the training set (approximately 30 minutes of data) degraded by flat communications channel noise at an SNR of 5 dB were enhanced using $H(1,1,1)$ to generate class based codebooks. The same classifier was used across different noise types for the test tokens considered in this study. Frames belonging to the same BPC are used for constructing class based codebooks and the codebook sizes are determined from the number of individual phonemes belonging to a BPC group. The BPCs with their codebook sizes are summarized in Table I. Using a 30-ms frame size with a 75% overlap rate, each short-time frame was parameterized using 12-dimensional linear predictor cepstral coefficients (LPCC) derived from the AR model parameters [12, pp. 376, Eq.(6.44)]. The utterances were pre-emphasized using the first-order FIR

TABLE II
VQ BROAD PHONEME CLASS RECOGNITION PERFORMANCE FOR TOKENS DEGRADED AT SNR OF 5 dB AND ENHANCED BY $H(1,1,1)$. PERCENTAGE CORRECTLY RECOGNIZED ALONG THE MAIN DIAGONAL (PHONEME CLASS KEY: VOW=VOWELS, SEMI=SEMI-VOWELS, NAS=NASALS, AFF=AFFRICATES, FRIC=FRICATIVES, STOP= STOPS, CLOS=CLOSURES, SIL=SILENCE SEGMENTS)

	VQ Recognized Class $\rightarrow$							
True Class $\downarrow$	**Vow**	**Semi**	**Nas**	**Aff**	**Fric**	**Stop**	**Clos**	**Sil**
Vow	**70.33**	11.02	5.51	0.27	3.57	5.12	3.31	0.87
Semi	17.84	**46.69**	10.87	0.52	6.78	8.14	6.06	3.10
Nas	13.22	11.21	**42.96**	1.70	8.08	8.52	6.77	7.54
Aff	3.79	1.55	2.59	**56.04**	10.51	9.14	10.52	5.86
Fric	3.59	1.61	5.56	4.83	**52.08**	11.04	13.89	7.40
Stop	3.63	4.31	10.43	2.51	15.30	**41.45**	17.31	5.06
Clos	4.29	3.14	3.38	2.41	20.25	10.91	**39.72**	15.90
Sil	1.06	1.73	2.87	2.79	13.33	7.41	17.17	**53.64**

filter $1 - 0.97z^{-1}$. The codebooks were optimized in a minimum mean square error (MMSE) sense. The distance between the test vectors $(\vec{C}_t)$ and codebook entries $(\vec{C}_r)$ was defined using a cepstral projection measure [14]. This distance metric uses the property that noise corrupted cepstral vectors are less sensitive to angle perturbation and is given by

$$l(\vec{C}_r, \vec{C}_t) = |\vec{C}_t| - \frac{\vec{C}_t^T \vec{C}_r}{|\vec{C}_r|}. \tag{10}$$

The confusion matrix in Table II summarizes the recognition performance of the VQ classifier for 128 test tokens of 16 speakers. The tokens were degraded with flat frequency response communications channel noise at an SNR of 5 dB and enhanced by $H(1,1,1)$ before performing phoneme classification.

For any given row in Table II, the percentage of correct classification is given by the main diagonal element and the percentages of misclassifications are given by the remaining elements of that row. Phoneme classification errors occurred mostly due to inter-class confusions arising between low-energy classes like fricatives, stops, closures, and silence. These errors were corrected using a simple forced classification technique in which any intermediate frame that had a different phoneme class from its neighboring (leading and trailing) frames was forced to match the phoneme class of its neighbors. This is a reasonably valid assumption since it is unlikely that two phoneme class transitions occur within three overlapping frames spanning a duration of only 45 ms.

C. *Hard and Soft Decision Synthesis*

An effective decision strategy is required for the reconstruction of the enhanced speech. Fig. 2 illustrates an overview of the ROVER enhancement framework. The objective is to choose the best set of enhanced frames from the archive of enhanced frames, discussed in Section III-A, using a search space constructed from Itakura–Saito (IS) distortion [18]. The IS distortion between clean and enhanced speech spectra is given by

$$d_j\left(P_s(\omega), \hat{P}_s(\omega,\theta)\right) = \frac{1}{2\pi}\left(\int_{-\pi}^{\pi} r(\omega) - \ln\left(r(\omega)\right) - 1\right) d\omega$$
$$\text{where} \quad r(\omega) = \frac{P_s(\omega)}{\hat{P}_s(\omega,\theta)} \tag{11}$$

and $P_s(\omega)$ and $\hat{P}_s(\omega,\theta)$ are obtained using the AR model power spectra of clean and enhanced speech, respectively, at frame j, and θ represents a filter configuration in (9) such that $\theta = (\alpha, \beta, \gamma)$. Therefore, only the enhanced speech spectrum is a function of θ whereas the clean speech spectrum is not a function of θ. The IS distortion measure is used for constructing the search space because it bears a high correlation with the subjective quality of speech [19]. Using 600 TIMIT sentences from the training set and degraded by flat communications channel noise at 5-dB SNR, a training archive is generated using (8) and an enhancement space using $\alpha \in [0.25, 0.5, \cdots, 6.0]$, $\beta \in [0.25, 0.5, \cdots, 4.0]$, $\gamma \in [1, 2, \ldots, 6]$. From this archive, the segment wise IS distortions between clean and enhanced speech, and degraded and enhanced speech, (D_δ), across all possible filter configurations of θ are calculated. Since the knowledge of phoneme class segment locations are known apriori from the phoneme level transcriptions provided in the training set, the segment wise IS distortions are grouped together based on the phoneme class. From this, the median (μ_δ) and standard deviation (σ_δ) per phoneme class are easily determined and stored. These parameters are used in generating the upper and lower bounds of a search space.

If the enhancement space in (9) is denoted by Γ and if $T_\alpha, T_\beta, T_\gamma$ represent the total number of values used by the parameters α, β, and γ, respectively, then $T_\Gamma = T_\alpha \times T_\beta \times T_\gamma$. Since θ represents any filter configuration in Γ, then $\Gamma = \{\theta_1, \theta_2, \cdots, \theta_{T_\Gamma}\}$. Furthermore, let a block or segment comprising of any set of contiguous frames belonging to a single BPC δ and enhanced from the filter $H(\theta)$ be given by

$$F_\delta(\theta) = \left\{f_{\delta,k}(\theta), f_{\delta,k+1}(\theta), \ldots, f_{\delta,k+n-1}(\theta)\right\}. \tag{12}$$

Here, the individual frames are denoted by $f_{\delta,i}$ where i denotes the index of the frame and given by $i = k, \ldots, k+n-1$. Therefore, the segment $F_\delta(\theta)$ is comprised of n contiguous frames. A contiguous sequence of frames (or a segment) belonging to a single BPC are selected for processing instead of individual frames in order to impose a level of naturalness to allow a reasonable rate for the speech spectrum to be allowed to change. However, selection of noise-only regions can be broken into individual frames because they do not contain any useful speech information. Hence, each noise-only segment is limited to no more than three contiguous frames.

With these considerations, the goal is to find the segment $F_\delta(\theta^\star)$ generated from the filter $H(\theta^\star)$ such that the average

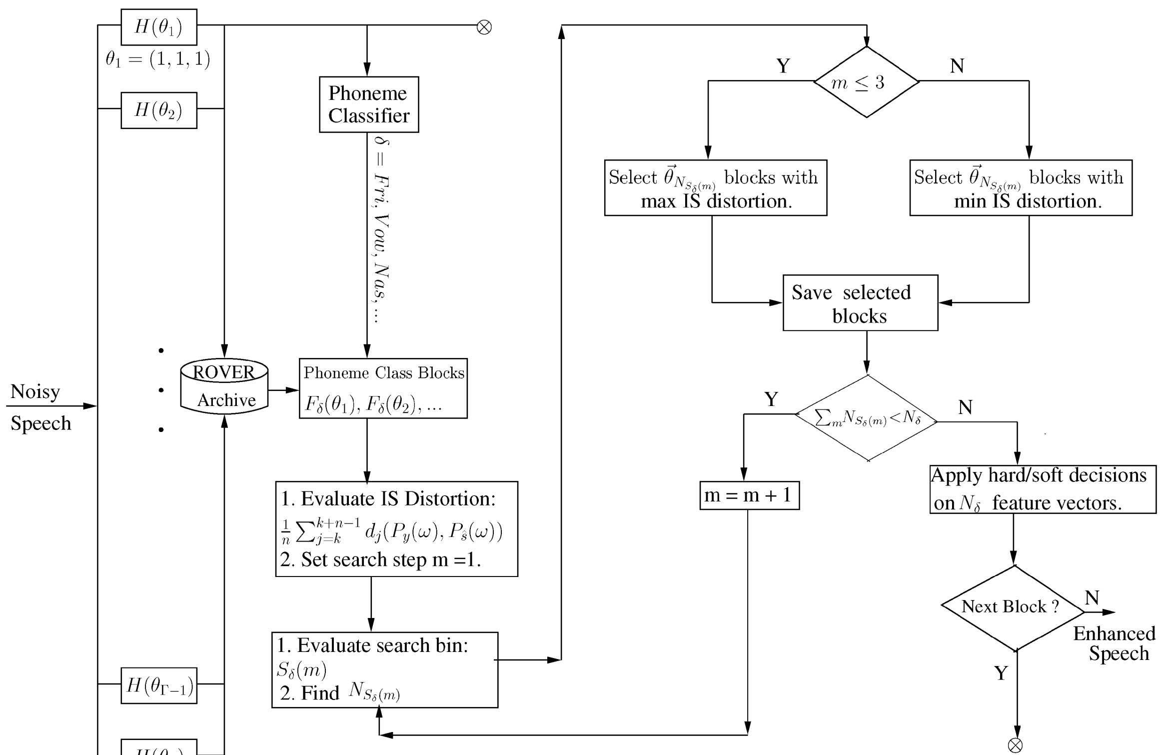

Fig. 2. ROVER enhancement framework.

IS distortion between clean and enhanced speech for a specific block of segment is minimized over enhanced utterances generated from all possible filter configurations. In other words, the goal is to find $F_\delta(\theta^\star)$ from $F_\delta(\theta_1), F_\delta(\theta_2), \ldots, F_\delta(\theta_{T_\Gamma})$.

1) Soft Decision: Based on the foregoing foundation, the following steps outline the soft decision solution:

1) Using the VQ phoneme classifier approach, a contiguous sequence of frames belonging to the same phoneme class δ is determined to select the segment $F_\delta(\theta)$.
2) For each $F_\delta(\theta)$ where $\theta \in \{\theta_1, \theta_2, \cdots, \theta_{T_\Gamma}\}$, the IS distortions are evaluated from degraded speech using

$$\frac{1}{n}\sum_{j=k}^{k+n-1} d_j\left(P_y(\omega), P_{\hat{s}}(\omega,\theta)\right). \tag{13}$$

We introduce the term *search step*, denoted by m, and initialize this to 1. The search step is as an index to the array of *search bins*. The search bins are explained in the next step.

3) For a given BPC δ, a bounded search bin $S_\delta(m)$ at the mth search step is constructed using

$$S_\delta(m) \\ =\{D_\delta : \max(0, \mu_\delta - m\epsilon_1\sigma_\delta) \le D_\delta \le \mu_\delta + m\epsilon_2\sigma_\delta\}. \tag{14}$$

The search bin is a bounded region of IS distortions. As was explained earlier, the IS distortions between clean and enhanced speech, and degraded and enhanced speech (D_δ) from the *training* corpus were calculated for every BPC δ and the median (μ_δ) and standard deviation (σ_δ) parameters were determined. Those IS distortions that fall within these bounds are used to fill the search bin $S_\delta(m)$. Here, ϵ_1 and ϵ_2 are set to 0.1 and represent the backward and forward weights, respectively, on σ_δ. As an example, the first search bins for vowels $(S_{vowel}(1))$ and stops $(S_{stop}(1))$ degraded by flat communications channel noise at 5 dB is shown in Fig. 3. Since the correlation between the IS distortions is higher in vowels than for stops, the initial search bin is narrower for vowels. This accounts for the reason why search bin parameters $(\mu_\delta, \sigma_\delta)$ in (14) are class dependent. Also, since the distribution of D_δ is skewed, it is more meaningful to use the median instead of the mean to determine the bounds. It is to be noted that the size of the search bin, denoted by vertical lines in Fig. 3, increases with increase in m.

Another point to be noted is that certain BPCs can be grouped together to form bigger groups. For example, vowels and semivowels or fricatives and closures may be grouped together depending on the energy levels. Intra-class acoustical characteristics within these bigger groups are similar (for example, vowels versus semivowels) but differ significantly when compared across inter-class (for example, vowels versus fricatives). From Table II, it is clear that most misclassifications occur due to class confusions among similar groups (e.g., vowels and semivowels) resulting in wrong selection of search bins. However, distributions of D_δ do not vary widely among similar groups like vowels and semivowels. This, to some extent, alleviates the misclassification errors caused by the VQ classifier.

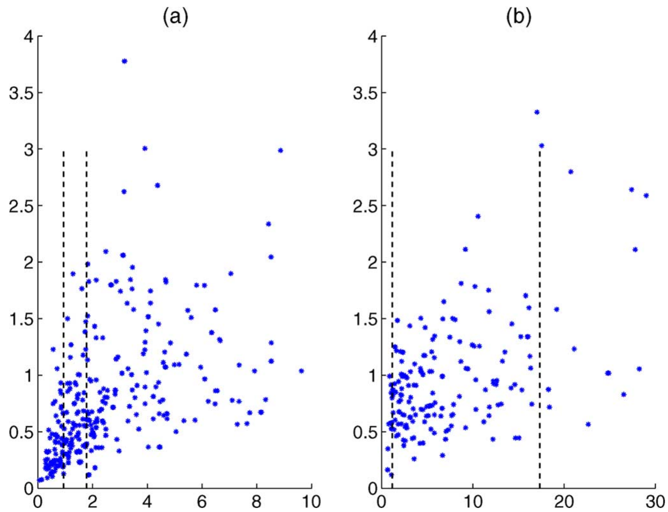

Fig. 3. Initial ($m = 1$) search space (within the vertical lines) for (a) vowels and (b) stops. X-axis: IS (degraded, enhanced), Y-axis: IS (clean, enhanced).

However, for misclassifications occurring across distinctly different groups (e.g., vowels and fricatives) reconstruction becomes more difficult.

In the next two steps 4) and 5), the objective is to find N-best segments $\{F_\delta(\theta_1^\star), F_\delta(\theta_2^\star), \ldots, F_\delta(\theta_N^\star)\}$ for reconstruction. Since N is dependent on BPC δ of the segment, N can be referred to as N_δ. Different values used for N_δ are given in the last column of Table III. Therefore, the higher the number of individual phonemes per BPC δ, the larger is the acoustic space spanned by δ, and hence, more number of segments (N_δ) are required to capture the characteristics of the phoneme. As will be mentioned subsequently in Section III-C1(6), the N-best segments are weighted by normalized maximum-likelihood scores to determine the relevance of each segment for the reconstruction of BPC in consideration. It may be noted that although all segments T_Γ may be selected instead of N_δ, and subsequently weighted by their maximum-likelihood scores, such a step would add unnecessary computational burden to the algorithm considering $N_\delta \ll T_\Gamma$. This is because for a given BPC not all configurations in T_Γ will aid during the reconstruction of the BPC. For example, segments generated using high values of α, β in (8) are ideal for the reconstruction of silence segments. However, these segments lack any useful information that might aid in the reconstruction of higher energy BPCs like vowels, semivowels, or nasals. Using the method of choosing N-best segments, it is expected that segments that lack useful information are excluded from maximum likelihood evaluations.

4) Selection of all N-best segments are not limited to the same search step but spread out over different search steps. A constraint is applied on the number of segments that will be selected out of N-best segments at each search step. This is performed in order to select a diverse range of segments. Segments selected at lower search steps are expected to retain more noise and less artifacts while those selected at higher search steps are expected to be more noise-free and/or possess more artifacts.

TABLE III
NUMBER OF INDIVIDUAL PHONEMES, NUMBER OF MIXTURES, AND NUMBER OF SEGMENTS USED PER BROAD PHONEME CLASS

Broad phoneme class	Individual phonemes	Mixtures	Segments($\mathbf{N}_\delta$)
Vowel	22	64	16
Semivowel	7	16	6
Nasal	7	16	6
Affricate	2	4	2
Fricative	9	32	8
Stop	7	16	6
Closure	7	16	6
Silence	3	8	4

If $N_{S_\delta(m)}$ represents the total number of segments out of N-best to be found at the mth search step for BPC δ, and if it is assumed that the initial condition is $N_{S_\delta(0)} = 0$ (i.e., no segment has been found prior to the first search step), then the required number of segments to be found at the mth search step is given by

$$N_{S_\delta(m)} = \min\left(N_\delta - \sum_{k=0}^{m-1} N_{S_\delta(k)}, 4\right) \tag{15}$$

where the second argument of $\min(.)$ operator indicates that the number of segments is restricted to 4 in the mth search step even if there are more than 4 present. On the other hand, if there are less than the required $N_{S_\delta(m)}$ segments, then $N_{S_\delta(m)}$ is set to the number of segments that are actually present.

5) This is the decision step for the selection of $N_{S_\delta(m)}$ segments at the mth search step. Assuming each segment $F_\delta(\theta)$ in search bin $S_\delta(m)$ is an equally likely candidate for selection, then $N_{S_\delta(m)}$ segments are selected as

$$\vec{\theta}^\star_{N_{S_\delta(m)}} = \begin{cases} \arg\max\limits_{\vec{\theta}_{N_{S_\delta(m)}}} \frac{1}{n} \sum\limits_{j=k}^{k+n-1} d_j\Big(P_y(\omega), \hat{P}_s(\omega,\theta)\Big) & m \le 3 \\ \arg\min\limits_{\vec{\theta}_{N_{S_\delta(m)}}} \frac{1}{n} \sum\limits_{j=k}^{k+n-1} d_j\Big(P_y(\omega), \hat{P}_s(\omega,\theta)\Big) & m > 3 \end{cases} \tag{16}$$

where $\vec{\theta}^\star_{N_{S_\delta(m)}} = \{\theta_1^\star, \theta_2^\star, \ldots, \theta^\star_{N_{S_\delta(m)}}\}$. It is to be noted that a particular selection of θ in search bin $S_\delta(m)$ precludes its reselection at a larger search bin $S_\delta(k)$ (where $k > m$) even though (16) might be satisfied in $S_\delta(k)$.

For any search bin $S_\delta(m)$, segments near the lower bound (lower IS distortion) are noisy but better at retaining the overall spectral structure. However, segments near the upper bound (higher IS distortion) are expected to have more noise suppression but overall spectral structure may be distorted. Therefore, for $m \le 3$, the search bin given by (14) is narrow and retains more noisy segments near the lower bound of $S_\delta(m)$ than noise suppressed segments near the upper bound. Hence, selection during the first three searches are biased towards choosing frames near the upper bound of $S_\delta(m)$. At subsequent iterations when $m > 3$, the upper bound of search bin is increased to accommodate more noise suppressed segments across a wider range having more distortions in spectral structure. Then, the selection procedure is not ideal for choosing segments near the upper bound. Hence, it is *reversed*. Noisy frames near the lower bound are chosen over noise suppressed frames near the upper bound. The core idea behind reversing has been in finding a tradeoff between suppressing noise and introducing processing artifacts so that segments with acceptable speech quality are used for reconstruction while others are rejected.

6) Next, it is determined whether to continue with the search process or proceed for reconstruction.

(6.i.) At the end of search step m if $\sum_m N_{S_\delta(m)} = N_\delta$, then it means all N-best segments have been found. Hence, the segments $F_\delta(\theta_1^\star), F_\delta(\theta_2^\star), \ldots, F_\delta(\theta_{N_\delta}^\star)$ are selected (and others rejected) and used for reconstruction of the enhanced speech using the following formulation. A Gaussian mixture model (GMM) based constrained soft decision solution is proposed here. Using clean speech from the training set, GMMs were constructed for each BPC from 12 dimensional LPCC vectors. The number of mixtures for the GMMs were determined from the number of individual phonemes present in the BPCs as given in Table III. Using GMMs, weights ϕ_i, $i = 1, 2, \ldots, N_\delta$, are assigned to the LPCC vectors obtained from the selected segments $F_\delta(\theta_1^\star), F_\delta(\theta_2^\star), \ldots, F_\delta(\theta_{N_\delta}^\star)$. If $\vec{X}(\theta_i^\star)$ represents the corresponding LPCC vector generated from the segment $F_\delta(\theta_i^\star)$, then the soft decision method finds the resultant feature vector $\vec{X}$ given by

$$\vec{X} = \sum_{i=1}^{N_\delta} \phi_i \vec{X}\left(\theta_i^\star\right) \tag{17}$$

where the term ϕ_i is defined by

$$\phi_i = \frac{p\left(\vec{X}\left(\theta_i^\star\right)|\lambda_\delta\right)}{\sum_{i=1}^{N_\delta} p\left(\vec{X}\left(\theta_i^\star\right)|\lambda_\delta\right)} \tag{18}$$

where λ_δ is the clean speech GMM model for BPC δ and $p(\vec{X}(\theta_i^\star)|\lambda_\delta)$ is the maximum-likelihood score of the model λ_δ for the feature vector $\vec{X}(\theta_i^\star)$. Assuming independence between frames in each segment, the term $p(\vec{X}(\theta_i^\star)|\lambda_\delta)$ can be further written as

$$p\left(\vec{X}\left(\theta_i^\star\right)|\lambda_\delta\right) = \prod_{t=k}^{k+n-1} \sum_{m=1}^{M} p\left(\vec{X}_t\left(\theta_i^\star\right)|\lambda_\delta, m\right) p(m|\lambda_\delta) \tag{19}$$

where $\vec{X}(\theta_i^\star)$ is the set of LPCC vectors $[\vec{X}_k(\theta_i^\star), \vec{X}_{k+1}(\theta_i^\star), \ldots, \vec{X}_{k+n-1}(\theta_i^\star)]$ for the sequence of frames with frame indices $[k, k+1, \ldots, k+n-1]$ of the segment $F_\delta(\theta_i^\star)$ and M is the total number of components in the GMM. The individual component density for the mth mixture is given as

$$p\left(\vec{X}_t\left(\theta_i^\star\right)|\lambda_\delta, m\right) = \omega_m \exp\left\{-\frac{1}{2}\Delta^{\mathrm{T}}\Sigma_m^{-1}\Delta\right\} \tag{20}$$

where $\Delta = \vec{X}_t(\theta_i^\star) - \vec{\mu}_m$, and $\omega_m = (1/(2\pi)^{D/2}|\Sigma_m|^{1/2})$. Here, the individual component densities for the mth mixture is parameterized by mean vector and diagonal covariance matrix $\{\vec{\mu}_m, \Sigma_m\}$ and weighted by the term $p(m|\lambda_\delta)$.

Once reconstruction of enhanced speech for the current segment is complete, the algorithm returns to Step 1) for the next segment. However, if all segments have been enhanced then the algorithm can terminate at this point.

(6.ii.) At the end of search step m if $\sum_m N_{S_\delta(m)} < N_\delta$, then all of N-best segments have not been found yet. Hence, the search process is continued by increasing the search step size by 1, i.e., by setting $m = m + 1$, and returning to Step 3).

2) Hard Decision: Hard decision-based selection is a special case of soft decision selection where only a single segment $F_\delta(\theta^\star)$ is selected from the search space instead of N-best segments. Hence, $N_\delta = 1$. The search step m is increased until the desired segment is found. Therefore, in (15), the second argument of $\min(.)$ operator is 1 instead of 4. Equation (16) in step 5) can be replaced for scalar parameter $\theta^\star$ instead of vector parameter $\vec{\theta}^\star_{N_{S_\delta(m)}}$. Finally, in Equation (17) in step 6), the weight of the selected segmented is assigned a value of 1.0 (i.e., $\phi_{i=1} = 1.0$) while all the remaining weights are assigned a value of 0 (i.e., $\phi_{i>1} = 0$).

An additional level of audible noise suppression can be achieved using estimates of auditory masking threshold (AMT) from hard decision and soft decision-based enhanced speech. The primary reason for incorporating AMT is to improve perceptual quality since some amount of audible residual noise may persist after grouping individual segments of enhanced speech. Since spectral components of noise are masked by speech, they can be minimized to an audible masking level (or AMT) instead of completely suppressing them. As a result, the spectral components of speech are better preserved and less perceptual distortion is introduced. Originally formulated by Tsoukalas, Mourjopoulos, and Kokkinakis [6], a codebook-based method was later proposed by Sarikaya and Hansen [22]. In the current framework, AMT is calculated from ROVER enhanced speech using the equivalent rectangular bandwidth (ERB) auditory filterbank model. For a detailed discussion on AMT using ERB, readers are advised to follow [15].

IV. Results and Evaluations

In this section, the results of detailed performance evaluations of the ROVER-based hard and soft decision enhancement solutions are summarized. The hard decision and soft decision ROVER solutions in this section will be referred to as HROV and SROV, respectively.

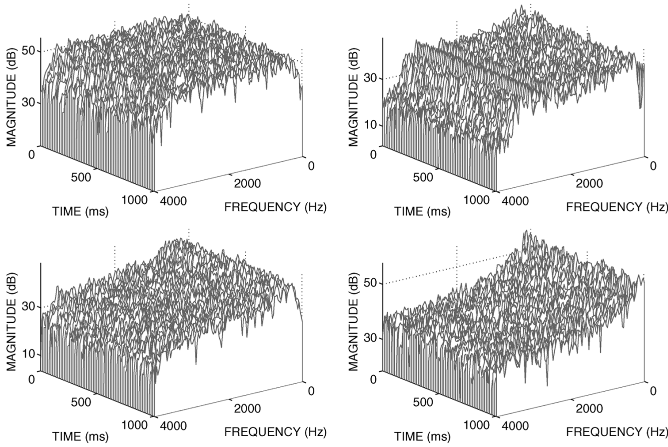

Fig. 4. Time versus frequency spectrograms for different noise types: (top left) Flat communications channel (FLN); (top right) Sun cooling fan (SUN); (bottom left) Large crowd (LCR); (bottom right) in-vehicle wind (BL4).

A. Experimental Setup

The core set of 192 phonetically balanced test utterances from the TIMIT corpus was used for objective quality evaluations. The corpus consisted of speakers from eight dialect regions in the US with two male and one female speakers per region with eight utterances per speaker. The corpus was sampled at 8 kHz and comprised of roughly 69 000 frames (240 samples per frame spanning 30 ms, 75% overlap). Each utterance was corrupted with the following noise types at global SNRs of 0 dB, 5 dB, and 10 dB: flat communications channel noise (FLN), Sun cooling fan noise (SUN), large crowd noise (LCR), and in-vehicle wind noise (BL4). FLN is a wideband stationary noise with a flat response like additive white Gaussian noise and extracted from AT&T voice communication channel. SUN is a stationary noise recorded from the cooling fan of a Sun 4/330 workstation. LCR is primarily a low-frequency slowly varying noise, recorded in a large crowded room with many ongoing conversations. The levels of any one conversation is not sufficient to identify individual speakers or words (i.e., LCR is not babble noise or competing speaker noise). Finally, BL4 is a narrowband (0–800 Hz) slowly varying noise recorded in an automobile (Ford Taurus) traveling at 60 mph on a freeway with windows partially open. The noise estimation was performed after averaging the power spectrum of the first 100-ms noise-only samples present in all TIMIT test utterances. The time versus frequency characteristics of the four noise types are shown in Fig. 4 to illustrate the nature of each noise type.

B. Objective Quality Measures

The quality of enhanced speech is assessed using objective speech quality measures such as the IS [18] (as given in (11)), segmental SNR (SegSNR), PESQ [21], and PESQ-LQ [24].

The IS distortion measure uses the dissimilarity between the all-pole spectra of the clean and enhanced speech and a lower value of IS measure implies better enhanced speech quality. SegSNR is a general measure of the degree of noise suppression and is calculated by taking the average of frame-wise SNRs. Higher values of SegSNR reflect more noise suppression and better signal-to-noise ratio (SNR) although they may not always

TABLE IV
PERCENTAGE IMPROVEMENT OF ITAKURA–SAIRO DISTORTION MEASURED AS $100 \times (IS_{\mathrm{Deg}} - IS_{\mathrm{Enh}})/IS_{\mathrm{Deg}}$ ACROSS PHONEME CLASSES DEGRADED BY FLAT COMMUNICATIONS CHANNEL NOISE AT 0-dB SNR

	Log-MMSE	**Log-MMSE-SPU**	**Auto-LSP**	**HROV**	**SROV**
Vowel	-12.428	-288.01	28.265	39.209	**43.713**
Semivowel	23.429	-112.08	24.870	34.202	**37.578**
Nasal	21.946	-67.91	27.578	30.473	**31.830**
Affricate	-45.702	-144.47	-4.912	-0.655	**-0.281**
Fricative	**56.655**	-26.76	40.616	44.265	47.074
Stop	29.856	-104.26	29.974	31.895	**34.451**
Closure	42.739	39.30	**42.908**	41.184	42.409
Silence	41.943	51.74	43.247	54.764	**56.133**

reflect better speech quality. However, if AMT is engaged, it is normal to expect that SegSNR values will fall. PESQ is an ITU recommendation with a range from 0–4.5. Higher value indicates better speech quality. PESQ assessment is more useful than IS or SegSNR when AMT is engaged since it is a measure of *perceived* speech quality. Finally, PESQ-LQ is a modified score obtained by mapping PESQ score to an average five-point absolute category rating (ACR) listening quality (LQ) scale defined by ITU-T P.800. The five-point ACR LQ scale comprises of excellent, good, fair, poor, bad ratings. PESQ-LQ was proposed to predict MOS scores better than PESQ. MOS scores can be affected by cultural and individual variations [25]. Also, some subjects are likely to get biased to test conditions, (i.e., the subject is likely to rate a poor condition token as excellent if the corpus has a large number of bad condition tokens [25]). PESQ-LQ is likely to give scores that will hold good on an average for a large corpus of subjective tests across different languages and regions. In the following sections, a summary of the results of the proposed enhancement solutions are compared with Auto-LSP, log-MMSE [20], and log-MMSE with speech presence uncertainty (log-MMSE-SPU) [29]. The performance is initially compared without AMT engaged and later with AMT engaged.

C. Performance Across Phoneme Classes

A summary of the IS distortion percentage improvement of enhanced speech over noisy speech calculated as $100 \times (IS_{\mathrm{Deg}} - IS_{\mathrm{Enh}})/IS_{\mathrm{Deg}}$ is shown in Table IV with the highest improvements in each row highlighted in bold. From Table IV, with the exception of closures and fricatives, HROV and SROV outperform Auto-LSP, log-MMSE, and log-MMSE-SPU over all other phoneme classes. It is to be noted that none of the enhancement algorithms could improve the affricates effectively as indicated by the negative values throughout the row. The negative percentage improvement indicates that the IS distortion after enhancement was higher than noisy speech suggesting that the enhancement algorithm caused a further degradation in speech quality. However, the degradation in the quality of affricates enhanced using HROV and SROV algorithms compared to noisy speech is less than just 1% and are difficult to perceive. Further, HROV/SROV experienced least degradation in affricates over the other competing algorithms. Log-MMSE exhibited the best performance for fricatives. However, it suffered higher degradation than noisy speech for affricates and vowels by about 45.70% and 12.43%, respectively. As for closures, the performance improvement of log-MMSE is marginal compared to HROV and SROV

TABLE V
MEAN AND VARIANCE OF ITAKURA–SAITO DISTORTION ACROSS PHONEME CLASSES FOR SPEECH DEGRADED BY FLAT COMMUNICATIONS CHANNEL NOISE AT 0-dB SNR

	Mean				**Variance**			
	Degraded	**Auto-LSP**	**HROV**	**SROV**	**Degraded**	**Auto-LSP**	**HROV**	**SROV**
Vowel	1.699	1.219	1.033	0.956	0.795	0.865	0.562	0.310
Semivowel	3.001	2.257	1.976	1.875	1.303	1.493	0.967	0.714
Nasal	4.201	3.042	2.921	2.864	1.283	1.464	0.991	0.781
Affricate	2.704	2.836	2.721	2.711	0.879	0.966	1.012	0.653
Fricative	3.404	2.021	1.897	1.801	1.181	1.194	1.488	0.926
Stop	2.962	2.074	2.017	1.942	1.284	1.740	1.719	1.012
Closure	7.031	4.014	4.135	4.049	5.376	3.251	5.861	3.503
Silence	8.045	4.566	3.640	3.529	3.134	2.764	1.668	1.235

TABLE VI
BEST ALGORITHMS ACROSS PHONEME CLASSES DEGRADED BY FLAT COMMUNICATIONS CHANNEL NOISE AT 0-dB SNR

Class	**Algorithm**	**Class**	**Algorithm**
Vowel	SROV	**Fricative**	log-MMSE
Semivowel	SROV	**Stop**	SROV
Nasal	SROV	**Closure**	Auto-LSP
Affricate	Degraded	**Silence**	SROV

since IS improvement is higher by approximately only 1.56% (42.739–41.184) and 0.33% (42.739–42.409) than HROV and SROV, respectively. This improvement is not significant to impact the perception of the average human listener. All classes except closures and silence in log-MMSE-SPU enhanced speech suffered further degradation compared to noisy speech.

The mean and variance of each phoneme class is tabulated across noisy, Auto-LSP, HROV and SROV speech in Table V for the same noise condition. Log-MMSE and log-MMSE-SPU has not been included in the table since it exhibited the least overall improvement from Table IV. The mean and variance data is indicative of the degree and consistency of lowering the IS distortion, respectively. Across all phoneme classes in Table V, SROV reported the least variance and hence the most consistent. It may be noted that across low energy phoneme classes (i.e., affricates, fricatives, stops, and closures) the variance of distortion is higher in HROV than noisy speech while it is lower for the remaining high energy classes. The increase in variance for low energy classes suggests the presence of misclassification errors or improper selections of filter parameter $\theta^\star$ during the hard decision step. Since SROV has lower variance, speech quality is expected to be more uniform in SROV than HROV.

The phoneme class results have been summarized in Table VI. Auto-LSP can be considered as the best algorithm for closures since it reported the least mean and variance. For fricatives, the winner is log-MMSE (mean $= 1.475$, variance $= 0.55$ not shown in table). Although no algorithm could enhance the affricates, SROV had the least degradation. For all other classes, SROV outperformed all other algorithms. The important point to consider here is that the ROVER solutions demonstrated a higher degree and consistency in improving the overall speech quality for most phoneme classes compared to other competing algorithms.

D. Performance Across Objective Measures

In Table VII, the results from the three objective measures—Itakura–Saito, SegSNR, and PESQ—are summarized for the case of flat communications channel noise at 0-dB, 5-dB, and 10-dB SNR levels and compared with those of log-MMSE, log-MMSE-SPU, and Auto-LSP algorithms. Unlike Section IV-C where results were analyzed over individual phoneme classes, the results presented here are averaged over the entire utterance. The numbers highlighted in bold indicate the best performance along the column. SROV reported the least IS distortion compared to all the other algorithms across all SNRs under consideration. The percentage improvement of IS over noisy speech is roughly 37%–38% at 0-dB and 5-dB SNR and about 51% at 10-dB SNR. Also, SROV outperformed HROV by about 2%–3% at lower SNRs and about 6% at 10-dB SNR.

Increase in SegSNR for the ROVER solutions over noisy speech achieved are about 9.1 dB at 0-dB SNR, 7.6 dB at 5-dB SNR, and 6.3 dB at 10-dB SNR. HROV reported the highest improvement over all the other competing algorithms. SegSNR values for SROV are about 0.2 dB lower than HROV. The contrasting performances in IS and SegSNR results for HROV and SROV is mostly due to the ability of SROV to retain better spectral structure due to the selection of diverse segments during the decision phase. This is more likely to happen at low energy phoneme classes like stops, closures, or fricatives. While it is difficult to recover the spectrum in these regions, improvement is observed in the form of noise suppression in HROV. Therefore, there is a system tradeoff between HROV and SROV methods. HROV performs better in suppressing noise and SROV is superior in retaining the spectral structure of the speech.

Finally, the improvement in the performance of the ROVER schemes is further emphasized by the PESQ results which are higher than Auto-LSP, log-MMSE, and log-MMSE-SPU. It is to be noted that all scores reported in Table VII do not take AMT enhancement into account.

In Fig. 5, a plot of frame-to-frame IS distortion is illustrated for the sentence "*Should she wake him?*". It is to be noted that in the segment encompassing the closure /dcl/ frames, HROV has higher distortion than noisy speech due to phoneme misclassification. While it is difficult for HROV to recover from this error, it has been mitigated to some extent in SROV due to soft decision. In the transition region from /iy/ to /w/ (frames 153–158), there is a rise in the distortion level in both HROV/SROV. This is again due to a semivowel (/w/) being misclassified as a vowel which accounts for the highest number of misclassification errors (i.e., 17.84% from Table II). Since both these classes have high energy and similar spectral structures, their filter estimates are not expected to have large variations. As a result, most of

TABLE VII
ITAKURA–SAITO, SegSNR, AND PESQ RESULTS FOR 192 TIMIT SENTENCES DEGRADED BY FLAT COMMUNICATIONS CHANNEL NOISE AT 0-dB, 5-dB, 10-dB SNRs

	Itakura-Saito			SegSNR			PESQ		
Enhancement	0dB	5dB	10dB	0dB	5dB	10dB	0dB	5dB	10dB
Degraded	3.436	2.672	1.966	-7.803	-3.714	0.807	1.515	1.807	2.132
Log-MMSE	2.472	2.723	2.051	-0.781	2.319	4.828	1.948	2.273	2.556
Log-MMSE-SPU	4.687	3.538	2.241	-0.339	2.290	4.678	1.684	2.015	2.323
Auto-LSP	2.354	1.878	1.237	-0.208	3.024	6.350	2.135	2.521	2.518
HROV	2.216	1.713	1.078	**1.384**	**3.958**	**7.214**	2.268	2.631	2.807
SROV	**2.153**	**1.657**	**0.953**	1.187	3.729	7.032	**2.312**	**2.684**	**2.943**

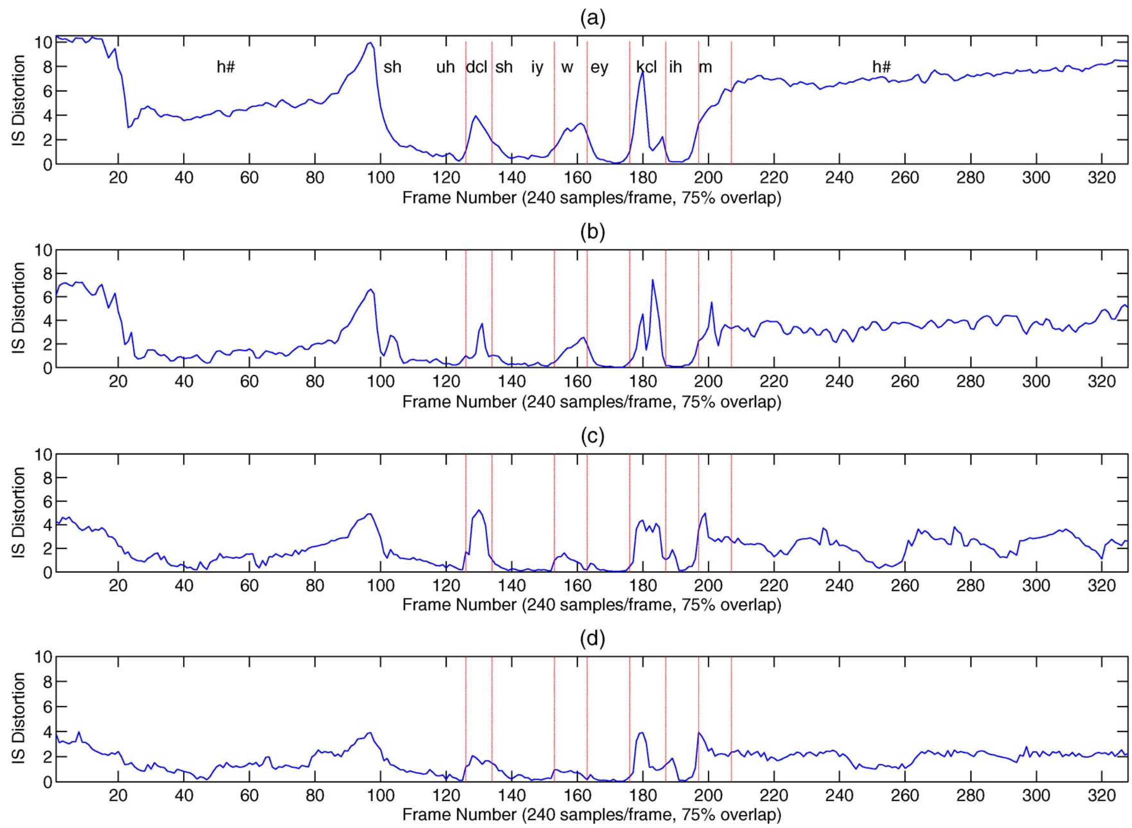

Fig. 5. Frame-to-frame Itakura–Saito distortion at 0-dB SNR. (a) Clean speech degraded by flat communications channel noise (Mean $= 5.098$, Var $= 8.172$). (b) Auto-LSP enhanced speech (Mean $= 2.588$, Var $= 3.330$). (c) HROV enhanced speech (Mean $= 1.925$, Var $= 1.601$). (d) SROV enhanced speech (Mean $= 1.628$, Var $= 0.813$).

the spectral shape is retained unlike /dcl/. The distortion levels in this region for HROV and SROV are still lower than noisy speech. However, in frames 159–163, /w/ was correctly classified resulting in a drop of IS distortion. During the transition from /ey/ to /kcl/, frames 178–183 were misclassified as /k/(unvoiced stop) which is similar to /kcl/ (closure). This did not severely affect the spectral shape and the reduction in distortion level is due to noise attenuation. In the noise only regions, HROV achieves lesser IS distortion levels intermittently while levels in SROV are more consistent.

E. Performance Across Noise Types

To study the performance of the HROV and SROV over all noise types at 0-dB, 5-dB, and 10-dB SNRs, the average IS distortion is compared against the other competing algorithms and is shown in Fig. 6. Log-MMSE-SPU has been excluded since it exhibited higher IS distortions than log-MMSE for FLN/LCR noises and marginally lower IS distortions than log-MMSE for SUN/BL4 noises. In general, log-MMSE-SPU performed better than log-MMSE with respect to SegSNR scores but performed worse with respect to IS or PESQ scores. Across noise types, the highest percentage improvement in IS measure was observed for FLN noise: 42.28% averaged over all SNRs. The least percentage improvement was observed for BL4 noise: 7.39% averaged over all SNRs. This is not an anomaly since at a given SNR the levels of degradation is already low for BL4 noise (degraded IS $=$ 1.07 at 5-dB SNR) and high for FLN noise (degraded IS $=$ 2.67 at 5-dB SNR). The low levels of degradation in BL4 noise limits the efficacy of enhancement algorithms. The results also indicate that the HROV solution caused an additional but marginal level of degradation to the noisy speech in the case of BL4 noise at 0-dB and 5-dB SNRs, whereas this degradation is mitigated in the SROV solution. At 10-dB SNR for BL4, the IS distortion of noisy speech is very low at 0.7 and this does not require any further enhancement. As a result, all enhancement algorithms fail for this particular case. For BL4 noise, Auto-LSP exhibited the best overall performance whereas

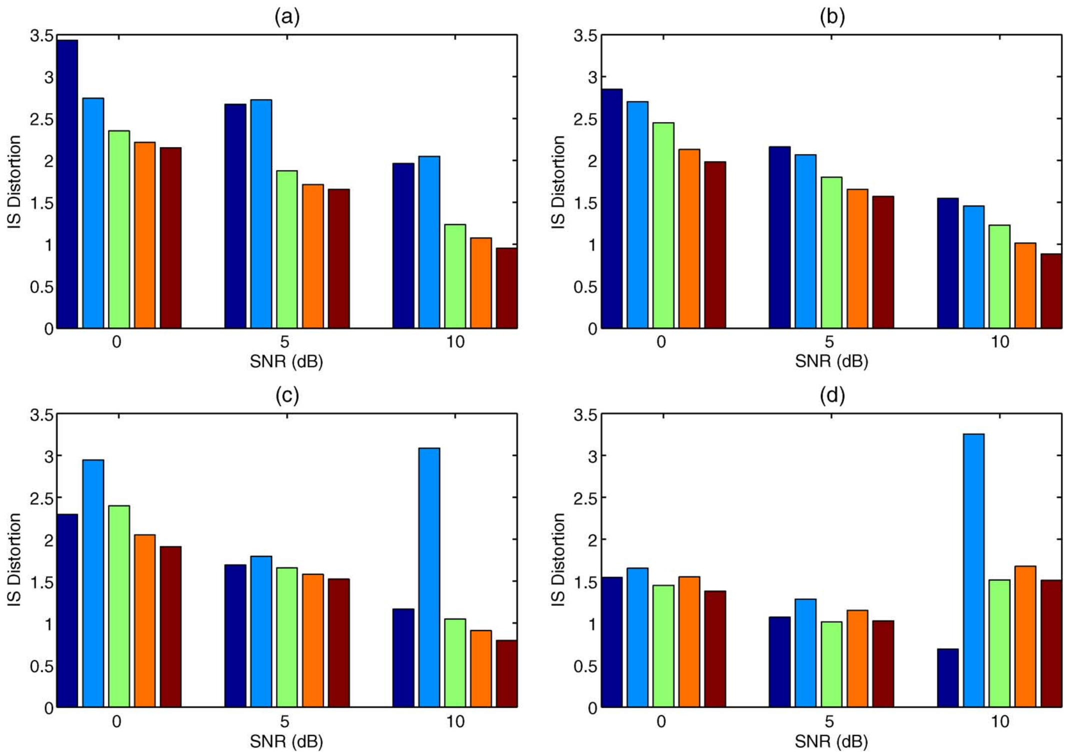

Fig. 6. Average Itakura–Saito distortion for different enhancement algorithms over 192 TIMIT sentences degraded with different noise types at SNRs of 0 dB, 5 dB, and 10 dB. Enhancement algorithms arranged left to right according to the order: Degraded, Log-MMSE, Auto-LSP, HROV, SROV. (a) Flat communications channel noise. (b) Sun cooling fan noise. (c) Large crowd noise. (d) In-vehicle wind noise.

TABLE VIII
PESQ AND PESQ-LQ SCORES WITH AND WITHOUT AMT ACROSS 192 TIMIT SENTENCES DEGRADED WITH FLN, SUN, LCR, AND BL4 NOISES AT AN SNR OF 5 dB. "BL" INDICATES A BASELINE ALGORITHM AND "N/A" INDICATES THAT THE RESULTS CANNOT BE OBTAINED AND HENCE NOT APPLICABLE

		FLN		**SUN**		**LCR**		**BL4**	
Enhancement	**Score**	BL	BL +AMT	BL	BL +AMT	BL	BL +AMT	BL	BL + AMT
Degraded	PESQ	1.807	N/A	1.910	N/A	2.103	N/A	2.445	N/A
Log-MMSE	PESQ	2.273	2.351	2.435	2.482	2.347	2.418	2.630	2.683
Auto-LSP	PESQ	2.521	2.584	2.594	2.622	2.628	2.661	2.667	2.704
HROV	PESQ	2.631	2.672	2.723	2.763	2.747	2.784	2.793	2.826
SROV	PESQ	**2.684**	**2.725**	**2.761**	**2.799**	**2.785**	**2.831**	**2.826**	**2.863**
Degraded	PESQ-LQ	1.097	N/A	1.202	N/A	1.426	N/A	1.890	N/A
Log-MMSE	PESQ-LQ	1.647	1.755	1.875	1.944	1.750	1.851	2.166	2.247
Auto-LSP	PESQ-LQ	2.002	2.096	2.111	2.154	2.163	2.214	2.223	2.280
HROV	PESQ-LQ	2.168	2.230	2.309	2.371	2.346	2.404	2.418	2.470
SROV	PESQ-LQ	**2.249**	**2.312**	**2.368**	**2.427**	**2.405**	**2.477**	**2.470**	**2.527**

for all other noise types SROV outperformed all other enhancement algorithms. The results for BL4 noise confirm that the ROVER algorithms do not significantly improve overall quality, and therefore Auto-LSP is a better candidate for cellular telephony applications in vehicles. However, for flat communications, sun cooling fan, and large crowd noise, improvement was observed consistently.

F. Performance With AMT Integrated

In the next experiment, AMT is engaged as a postprocessor of the enhancement algorithms discussed in this study, and PESQ and PESQ-LQ results over all noise types at an SNR of 5 dB are tabulated in Table VIII where each row represents one enhancement scheme. In Table VIII, any entry in the "BL" or "Baseline" column is a PESQ score of the enhancement scheme present in the corresponding row when there is no AMT engaged. The "BL + AMT" column has AMT engaged as a second level of enhancement. Since the AMT approach followed in this study requires prior knowledge of the clean speech estimate that can be obtained from any of the enhancement schemes, the results for AMT enhancement using noisy speech is not possible. Although noisy speech can be used to estimate the AMT, it is usually not considered a preferred procedure. The PESQ and PESQ-LQ results indicate that SROV performs the best resulting in improved levels of speech quality for all noise types. Improvement in BL4 was the least because of the reduced degradation caused by BL4 noise in comparison to FLN or SUN noises.

G. Performance Across NIST Phonemes

The IS performance summary for the NIST 61 individual phonemes listed in TIMIT 192 sentences is reported in Table IX for HROV and SROV solutions and compared with Auto-LSP (indicated by AUT). The sentences were degraded with FLN

TABLE IX
ITAKURA–SAITO DISTORTION FOR AUTO-LSP, HROV, AND SROV ENHANCED SPEECH ACROSS 61 NIST PHONEMES FROM THE 192 TIMIT UTTERANCES DEGRADED BY FLAT COMMUNICATIONS CHANNEL NOISE AT AN SNR OF 5 dB

OBJECTIVE SPEECH QUALITY ACROSS AMERICAN PHONEMES						
Ph.		DEG	AUT	HROV	SROV	*# Fr*
CONSONANTS – nasals						
/m/	me	3.310	2.312	2.109	2.103	1492
/n/	no	3.398	2.488	2.364	2.344	2049
/ng/	sing	3.318	2.760	2.714	2.651	360
/nx/	many	1.834	1.216	1.105	1.101	125
/em/	problem	3.184	2.413	2.121	2.103	33
/en/	traction	3.474	3.019	2.898	2.863	256
/eng/	greasing	2.338	1.074	1.011	0.951	6
CONSONANTS – unvoiced fricatives						
/s/	sip	2.331	2.269	2.231	2.193	4419
/th/	thing	3.749	1.729	1.652	1.610	354
/f/	fan	2.938	1.510	1.476	1.474	1648
/sh/	show	1.510	1.528	1.479	1.477	999
CONSONANTS – voiced fricatives						
/z/	zip	2.803	2.159	2.053	2.014	1833
/zh/	garage	1.784	3.533	2.876	2.105	104
/dh/	that	3.228	1.578	1.453	1.402	558
/v/	van	3.048	1.724	1.651	1.597	663
CONSONANTS – affricates						
/jh/	joke	2.031	2.469	1.954	1.908	323
/ch/	chop	1.856	1.680	1.645	1.581	428
VOWELS – front						
/ih/	hid	0.843	0.400	0.229	0.193	1856
/eh/	head	0.707	0.354	0.168	0.142	2032
/ae/	had	0.534	0.254	0.158	0.136	1739
/ux/	to buy	1.424	0.760	0.453	0.421	540
VOWELS – mid						
/aa/	odd	1.072	0.618	0.401	0.354	2000
/er/	earth	1.683	1.095	0.913	0.872	1419
/ah/	up	0.963	0.519	0.392	0.355	1365
/ao/	all	1.798	1.135	0.989	0.913	1458
VOWELS – back						
/uw/	boot	1.934	1.202	0.913	0.858	282
/uh/	foot	1.061	0.550	0.401	0.396	261
VOWELS – front schwa						
/ix/	heed	1.559	0.996	0.712	0.548	2268
VOWELS – back schwa						
/ax/	a ton	1.747	1.063	0.953	0.887	998
VOWELS – retroflexed schwa						
/axr/	after	2.326	1.608	1.397	1.263	1339
VOWELS – voiceless schwa						
/ax-h/	sub	2.859	3.744	3.427	3.318	48
CONSONANTS – unvoiced stops						
/p/	pan	2.244	2.293	2.341	2.286	712
/t/	tan	1.676	2.650	2.714	2.434	1005
/k/	key	2.050	2.606	2.293	2.104	1008
CONSONANTS – voiced stops						
/b/	be	2.401	1.006	0.892	0.851	253
/d/	dawn	1.933	1.989	1.865	1.783	322
/g/	give	2.392	2.332	2.193	2.157	223
CONSONANTS – closure stops						
/tcl/	it pays	6.094	3.206	3.627	3.073	1572
/kcl/	pockets	6.175	3.242	3.491	3.177	1445
/bcl/	to buy	6.556	3.721	3.628	3.559	887
/dcl/	sandwich	5.641	3.177	3.103	3.058	1116
/gcl/	iguanas	5.893	3.364	3.302	3.284	486
/pcl/	accomplish	7.029	3.805	3.824	3.237	1125
CONSONANTS – glottal stop, flap						
/q/	_allow	2.767	1.900	1.859	1.801	804
/dx/	put_in	1.903	0.916	0.863	0.793	297
CONSONANTS – unvoiced whisper						
/hh/	had	2.829	2.066	2.257	1.951	372
CONSONANTS – voiced whisper						
/hv/	you have	1.938	2.093	1.982	1.953	249
DIPHTHONGS						
/ay/	hide	0.637	0.311	0.192	0.154	1637
/oy/	coin	1.471	0.869	0.614	0.572	356
/ey/	pain	0.645	0.293	0.195	0.152	1844
/ow/	code	1.525	0.908	0.643	0.573	1385
/aw/	pout	0.743	0.401	0.314	0.216	630
/iy/	new	1.152	0.615	0.413	0.357	2553
SEMIVOWELS – liquids						
/r/	ran	1.995	1.339	1.102	0.953	1867
/l/	lawn	2.140	1.371	1.216	1.163	1701
/el/	chemicals	2.654	1.747	1.542	1.431	635
SEMIVOWELS – glides						
/w/	wet	3.141	2.106	1.936	1.891	1062
/y/	you	1.587	1.033	0.934	0.855	355
Silence						
/# /	extended	7.216	3.755	3.156	2.981	8905
/pau/	pause	5.688	2.599	2.193	1.995	1041
/epi/	epenthetic	4.869	2.148	1.702	1.615	227
Overall		2.672	1.878	1.713	1.657	69329
Overall - /#/		2.425	1.601	1.436	1.395	60424

noise at an SNR of 5 dB. Although the two ROVER solutions outperformed Auto-LSP in most of the phonemes, results conclude that the unvoiced stops (/p/,/t/,/k/), voiced whisper (/hv/), and voiceless schwa (/ax-h/) are slightly distorted after enhancement of any kind. The performance of Auto-LSP was better than HROV for some of the unvoiced stops (/p/, /t/) and closures (/tcl/,/kcl/,/pcl/). However, SROV was able to outperform Auto-LSP in all of these cases.

H. Complexity

In this section, we discuss the time complexity of implementing the algorithm. Since Auto-LSP lies at the core of the ROVER framework, we assume the complexity of running Auto-LSP is A and that we generate a single frame of ROVER enhanced speech. The complexity analysis can be split into different steps: power spectra generation, IS distortion evaluation, feature extraction, VQ classification, and finally finding GMM likelihoods.

To determine the power spectrum of single frame of enhanced speech, there are $Q = K/2+1$ distinct spectrum samples generated from applying DFT. K is the size of the DFT as was defined in (2). The DFT transform uses $2Q\log_2 Q$ multiplications and $3Q\log_2 Q$ additions. To generate the power spectrum, there are $3Q$ more multiplications and Q additions. Hence, considering the enhancement space originating from (9), there are a total of $T_\Gamma(2Q\log_2 Q + 3Q)$ multiplications and $T_\Gamma(3Q\log_2 Q + Q)$ additions. Further, the power spectrum of noisy speech requires $2Q\log_2 Q+3Q$ multiplications and $(3Q\log_2 Q+Q)$ additions.

At the feature extraction stage, an inverse DFT is performed on the power spectrum to get the autocorrelation coefficients from which LPCCs are extracted. The inverse DFT requires $2Q\log_2 Q + 3Q$ multiplications and $3Q\log_2 Q + Q$ additions. Out of Q, only P autocorrelation values (2) are saved to create a Toeplitz matrix. Inversion of the Toeplitz matrix and determining the P LPCs require P^3+P^2 multiplications and P^2-P additions. Next, to determine the kth LPCC, we require $2(k-1)$

multiplications and $k-1$ additions. To determine the D dimensional LPCC vector, we require $2\sum_{k=1}^{D}(k-1) = D(D-1)$ multiplications and $\sum_{k=1}^{D}(k-1) = D(D-1)/2$ additions. Evaluating (10) requires $V(3D+1)$ multiplications and $V(3D-2)$ additions where V is the number of codebook entries obtained from Table I.

Since the power spectrum is already known, calculating IS distortion requires $2T_{\Gamma}Q$ multiplications and $T_{\Gamma}Q$ additions. From this, finding the N-best segments N_{δ} satisfying maximum and minimum criteria in (16), requires scanning through not more than T_{Γ} IS distortion values.

Finally, finding the GMM likelihoods for N-best segments is dominated by $N_{\delta}MD$ multiplications and additions. Overall, the most dominant complexity term is in the calculation of (9) involving $T_{\Gamma}(2Q\log_2 Q + 3Q)$ multiplications and $T_{\Gamma}(3Q\log_2 Q + Q)$ additions in addition to the complexity involved in generating the Auto-LSP utterances, i.e., $T_{\Gamma}A$. Using an Intel processor with 1.8-GHz clock rate and MATLAB environment, the average time to enhance a single utterance was approximately 9 s. We are investigating a smaller size of the enhancement space by considering only the most relevant α, β, γ parameters to reduce the computational burden of the dominant complexity term. Since the enhancement framework utilize outputs from multiple iterations, they can be used for offline applications like spoken document retrieval, and news broadcasting.

V. Conclusion

A ROVER-based enhancement algorithm was introduced to enhance speech selectively based on phoneme classes degraded by various noise types. Hard and soft decision ROVER solutions were proposed. In both solutions, multiple enhanced utterances are generated per noisy utterance. The noisy utterance is partitioned into segments based on broad phoneme classes using a vector quantization classifier. From this knowledge, class specific constraints are applied. In the hard decision approach, only one segment from the multiple utterances set is selected for every segment of the noisy speech. Selection errors in hard decision were alleviated using a soft decision approach. In the soft decision method, instead of one segment, several segments are selected and weighted using GMMs. Finally, a second level of enhancement using estimates of auditory masking threshold was applied to the hard and soft decision solutions to remove audible residual noise.

The proposed algorithms were shown to be effective in various objective quality evaluations. Experiments were carried over the TIMIT 192 core test utterances degraded by four noise types and at three SNR levels: 0 dB, 5 dB, and 10 dB. The performance was assessed and analyzed using three objective quality metrics (Itakura–Saito, SegSNR, and PESQ). Across eight broad phoneme classes, it was demonstrated that the levels of perceived quality of speech improved across most of the phoneme classes when compared with the performance of Auto-LSP, log-MMSE, and log-MMSE-SPU. After engaging AMT as an additional level of enhancement, perceptual evaluation using PESQ results confirmed the superiority of the ROVER solutions.

Future studies could consider analysis in the evaluation of the effect of smoothing during transitions between broad phoneme class segments. The effect of using probabilistic decisions, instead of binary decisions, during phoneme class classification could be investigated. Instead of using a single set of search space parameters [$m, \epsilon_1, \epsilon_2$ in (14)] and number of search steps in forward and backward direction, they could be optimized for each phoneme class. Further, integration of the proposed ROVER solutions with other enhancement algorithms such as log-MMSE could also be investigated. In real world environments, these methods could be easily integrated into an overall solution for addressing additive noise suppression, convolutional channel and microphone distortion, and/or room noise acoustics. From a systems normalization perspective, adaptation to different noise sources using limited noise tokens could also be studied.

Acknowledgment

The authors would like to thank P. Angkititrakul and V. Prakash of CRSS at the University of Texas at Dallas for sharing their feature extraction and model building routines. They would also like to thank N. Krishnamurthy, also at CRSS, for his help with computing systems, and the anonymous reviewers for their helpful and constructive comments which significantly improved the quality of the manuscript.

References

[1] J. H. L. Hansen and M. A. Clements, "Constrained iterative speech enhancement with application to speech recognition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 39, no. 4, pp. 795–805, Apr. 1991.

[2] J. H. L. Hansen and L. M. Arslan, "Robust feature-estimation and objective quality assessment for noisy speech recognition using the credit card corpus," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 3, pp. 169–184, May 1995.

[3] D. Sen and W. H. Holmes, "Perceptual enhancement for CELP speech coders," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 1994, vol. 2, pp. 105–110.

[4] J. H. L. Hansen and L. Arslan, "Markov model based phoneme class partitioning for improved constrained iterative speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 1, pp. 98–104, Jan. 1995.

[5] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 2, pp. 126–137, Mar. 1999.

[6] D. E. Tsoukalas, J. N. Mourjopoulos, and G. Kokkinakis, "Speech enhancement based on audible noise suppression," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 6, pp. 497–514, Nov. 1997.

[7] P. J. Wolfe and S. J. Godsill, "Towards a perceptually optimal spectral amplitude estimator for audio signal enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Jun. 2000, vol. 2, pp. 821–824.

[8] P. C. Loizou, "Speech enhancement based on perceptually motivated Bayesian estimators of the magnitude spectrum," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 857–896, Sep. 2005.

[9] J. Wu, J. Droppo, L. Deng, and A. Acero, "A noise-robust ASR front-end using Wiener filter constructed from MMSE estimation of clean speech and noise," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, pp. 321–326, ASRU 2003.

[10] J. S. Lim and A. V. Oppenheim, "All pole modeling of degraded speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-26, pp. 197–210, Jun. 1978.

[11] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, pp. 137–145, Apr. 1980.

[12] J. Deller, J. H. L. Hansen, and J. Proakis, *Discrete Time Processing of Speech Signals*. Upper Saddle River, NJ: Prentice-Hall, 2000.

[13] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. COM 28, no. 1, pp. 84–95, Jan. 1980.

[14] D. Mansour and B. H. Juang, "A family of distortion measures based upon projection operations for robust speech recognition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 11, pp. 1659–1671, Nov. 1989.

[15] A. Natarajan, J. H. L. Hansen, K. H. Arehart, and J. Rossi-Katz, "Perceptual based speech enhancement for normal-hearing and hearing-impaired individuals," *EURASIP J. Appl. DSP: Spec. Iss. Signal Process. Hearing Aids Cochlear Implants*, pp. 1425–1428, Oct. 2005.

[16] S. Nandakumar and J. H. L. Hansen, "Dual-channel iterative speech enhancement with constraints based on an auditory spectrum," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 1, pp. 22–34, Jan. 1995.

[17] J. H. L. Hansen and S. Nandakumar, "Robust estimation of speech in noisy backgrounds based on aspects of the human auditory process," *J. Acoust. Soc. Amer.*, vol. 97, no. 6, pp. 3833–3849, 1995.

[18] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-23, no. 1, pp. 67–72, Jan. 1975.

[19] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective Measures of Speech Quality*. Englewood Cliffs, NJ: Prentice-Hall, 1988.

[20] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-33, no. 2, pp. 443–445, Apr. 1985.

[21] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)—A new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2001, vol. 2, pp. 749–752.

[22] R. Sarikaya and J. H. L. Hansen, "Auditory masking threshold estimation for broadband noise sources with application to speech enhancement," in *Proc. Eurospeech '99*, Budapest, Hungary, Sep. 1999, vol. 6, pp. 2571–2574.

[23] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER)," in *Proc. IEEE Workshop Autom. Speech Recognition Understanding*, pp. 347–354, ASRU 1997.

[24] A. W. Rix, "A new PESQ-LQ scale to assist comparison between P.862 score and subjective MOS," ITU-T SG12 COM12-D86, May 2002.

[25] A. W. Rix, "Comparison between subjective listening quality and P.862 PESQ score," in *Proc. Meas. Speech Audio Quality Netw. (MESAQIN'03)*, May 2003.

[26] T. V. Sreenivas and P. Kirnapure, "Codebook constrained Wiener filtering for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 5, pp. 383–389, Sep. 1996.

[27] J. S. Lim and A. V. Oppenheim, "All pole modeling of degraded speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-26, no. 3, pp. 197–210, Jun. 1978.

[28] D. L. Wang and G. J. Brown, "Separation of speech from interfering sounds based on oscillatory correlation," *IEEE Trans. Neural Netw.*, vol. 10, no. 3, pp. 684–697, May 1999.

[29] I. Cohen, "Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator," *IEEE Signal Process. Lett.*, vol. 9, no. 4, pp. 113–116, Apr. 2002.

Amit Das (S'07) received the B.E. degree in electronics and communications engineering from the University of Madras, Madras, India, and the M.S. degree in electrical engineering from the University of Colorado, Boulder.

He is currently a member of the Speech Processing Research Group, Department of Electrical Engineering, Indian Institute of Technology, Madras. From 2007 to 2011, he worked at Qualcomm, San Diego, CA, enhancing RF front-end capabilities of WCDMA receivers and transmitters used in 3G wireless devices. Prior to that, he was a Research Staff Member at the Center for Robust Speech Systems (CRSS), University of Texas at Dallas, Richardson, and a Research Assistant at the Center for Spoken Language Research (CSLR), University of Colorado at Boulder. His research interests span the areas of speech enhancement, speech recognition, and speaker adaptation.

John. H. L. Hansen (S'81–M'82–SM'93–F'07) received the B.S.E.E. degree from the College of Engineering, 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, in 1983 and 1988.

He joined the Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas (UTD), Richardson, in the fall of 2005, where he is Professor and Department Head of Electrical Engineering and holds the Distinguished University Chair in Telecommunications Engineering. He also holds a joint appointment as Professor in the School of Behavioral and Brain Sciences (Speech and Hearing). At UTD, he established the Center for Robust Speech Systems (CRSS) which is part of the Human Language Technology Research Institute. Previously, he served as Department Chairman and Professor in the Department of Speech, Language, and Hearing Sciences (SLHS) and Professor in the Department of Electrical and Computer Engineering, at the University of Colorado, Boulder (1998–2005), where he cofounded the Center for Spoken Language Research. In 1988, he established the Robust Speech Processing Laboratory (RSPL) and continues to direct research activities in CRSS at UTD. His research interests span the areas of digital speech processing, analysis, and modeling of speech and speaker traits, speech enhancement, feature estimation in noise, robust speech recognition with emphasis on spoken document retrieval, and in-vehicle interactive systems for hands-free human-computer interaction. He has supervised 58 (27 Ph.D., 31 M.S./M.A.) thesis candidates. He is author/coauthor of 417 journal and conference papers and ten textbooks in the field of speech processing and language technology, coauthor of the textbook *Discrete-Time Processing of Speech Signals*, (IEEE Press, 2000), coeditor of *DSP for In-Vehicle and Mobile Systems* (Springer, 2004), *Advances for In-Vehicle and Mobile Systems: Challenges for International Standards* (Springer, 2006), and lead author of the report "The Impact of Speech Under 'Stress' on Military Speech Technology," (NATO RTO-TR-10, 2000).

Prof. Hansen was named IEEE Fellow for contributions in "Robust Speech Recognition in Stress and Noise," in 2007 and is currently serving as Member of the IEEE Signal Processing Society Speech Technical Committee (2005–2008; 2010–2013; elected TC Chair in 2011), and Educational Technical Committee (2005–2008; 2008–2010). He was named International Speech Communications Association (ISCA) Fellow in 2010. Previously, he served as Technical Advisor to a U.S. Delegate for NATO (IST/TG-01), IEEE Signal Processing Society Distinguished Lecturer (2005/2006), Associate Editor for the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING (1992–1999), Associate Editor for the IEEE SIGNAL PROCESSING LETTERS (1998–2000), and Editorial Board Member for the IEEE SIGNAL PROCESSING MAGAZINE (2001–2003). He has also served as Guest Editor of the October 1994 special issue on Robust Speech Recognition for the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING. He has served on the Speech Communications Technical Committee for the Acoustical Society of America (2000–2003), and is serving as a member of the International Speech Communications Association (ISCA) Advisory Council. He was recipient of the 2005 University of Colorado Teacher Recognition Award as voted by the student body. He also organized and served as General Chair for ICSLP/Interspeech-2002: International Conference on Spoken Language Processing, September 16–20, 2002, and has served as Co-Organizer and Technical Program Chair for the IEEE ICASSP-2010, Dallas, TX.