Phoneme Selective Speech Enhancement Using Parametric Estimators and the Mixture Maximum Model: A Unifying Approach

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

***Abstract*—This study presents a ROVER speech enhancement algorithm that employs a series of prior enhanced utterances, each customized for a specific broad level phoneme class, to generate a single composite utterance which provides overall improved objective quality across all classes. The noisy utterance is first partitioned into speech and non-speech regions using a voice activity detector, followed by a mixture maximum (MIXMAX) model which is used to make probabilistic decisions in the speech regions to determine phoneme class weights. The prior enhanced utterances are weighted by these decisions and combined to form the final composite utterance. The enhancement system that generates the prior enhanced utterances comprises of a family of parametric gain functions whose parameters are flexible and can be varied to achieve high enhancement levels per phoneme class. These parametric gain functions are derived using 1) a weighted Euclidean distortion cost function, and 2) by modeling clean speech spectral magnitudes or discrete Fourier transform coefficients by Chi or two-sided Gamma priors, respectively. The special case estimators of these gain functions are the generalized spectral subtraction (GSS), minimum mean square error (MMSE), two-sided Gamma or joint maximum *a posteriori* (MAP) estimators. Performance evaluations performed over two noise types and signal-to-noise ratios (SNRs) ranging from −5 dB to 10 dB suggest that the proposed ROVER algorithm not only outperforms the special case estimators but also the family of parametric estimators when all phoneme classes are jointly considered.**

***Index Terms*—Generalized spectral subtraction, minimum mean square error (MMSE) estimator, joint maximum *a posteriori* (JMAP) estimator, mixture maximum (MIXMAX) model, 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 that impact either human perception or speech language technology performance. Hence, enhancement algorithms can be broadly classified as perceptual centric or speech systems centric. A myriad of algorithms have been developed over the last three decades in both categories. The perceptual centric algorithms [1]–[8] improve subjective quality of speech whereas the speech systems centric algorithms [9]–[13] improve some mathematical scoring metric which could be an objective quality of speech, or speech recognition or speaker identification accuracy percentage.

This study focuses on a speech systems centric framework. Early approaches in speech systems centric algorithms include spectral subtraction (SS) [14] and its variations [15], [16]. The SS method calculates the estimates of the noise spectrum from preceding frames where speech is absent under the assumption that statistics of the noise spectrum do not vary rapidly in time. The clean speech spectral magnitudes are estimated by subtracting the noise spectral magnitude from the spectral magnitude of the degraded speech. However, this scheme has the primary limitation that it is likely to produce musical noise due to random residual noise spectral peaks that are annoying to the listener.

Later, iterative Wiener filtering [17] and subsequent constrained estimation variations [9], [10] were adopted. This filter minimized the estimation error between the clean and estimated signal in the mean-square sense. However, the main drawback of the traditional Wiener filter is that it is assumed to be linear, and its frequency response is a function of the *a priori* signal-to-noise ratio (SNR) only but does not directly take into account the *a posteriori* SNR which is important for reducing musical noise. The Ephraim–Malah minimum mean square error (MMSE) estimator [11] has gained acceptance in the contemporary literature. The MMSE estimator is a non linear Bayesian estimator which minimizes the mean square error (MSE) between the clean and estimated speech spectral magnitudes. In addition, the gain function incorporates both *a priori* and *a posteriori* SNRs. In all these approaches, noise suppression is performed on the degraded or noisy speech only

Manuscript received April 26, 2011; revised November 28, 2011, March 14, 2012; accepted March 20, 2012. Date of publication June 12, 2012; date of current version August 13, 2012. This work was supported in part by the Air Force Research Laboratory (AFRL) under contract FA8750-12-1-0188 (Approved for public release, distribution unlimited), and in part by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J.H.L. Hansen. A. Das was with the Center for Robust Speech Systems (CRSS), Department of Electrical Engineering, University of Texas at Dallas, Richardson, TX, when this work was performed. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Hui Jiang.

A. Das is with the Indian Institute of Technology Madras, Chennai 600 036, India (e-mail: amit.das@ieee.org).

John 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).

Digital Object Identifier 10.1109/TASL.2012.2201471

in the holistic sense (i.e., a configuration that gives reasonably good objective quality score for the *overall* utterance). Such schemes do not adopt any phoneme class selective enhancement approach which is likely due to the inherent assumption that the *a priori* or *a posteriori* SNRs are functions of the amount of degradation in each phoneme class.

Earlier studies [9] 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 [18, Ch. 2]; 2) the structure of different noise types can be classified based on their degree of stationarity and their bandwidths.

Several research efforts have been devoted on developing phoneme class-based enhancement algorithms. In one study, Hansen and Arslan [19] used hidden Markov models (HMMs) to create 13 phoneme class models. Using the forward algorithm scoring procedure, conditional probabilities $p(\vec{X} \mid \lambda_i)$, $i = 1, 2, \ldots, 13$, 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, based on the Auto-LSP [9] algorithm, was performed selectively using this measure.

In this contribution, we present a class selective enhancement approach based on the generalized spectral subtraction (GSS) derived by Sim *et al.* [20], Ephraim–Malah MMSE estimator [11], Martin's discrete Fourier transform (DFT)-based MMSE estimator [21], and Wolfe–Godsill joint maximum *a posteriori* (JMAP) estimator [22] and demonstrate overall improvement in objective quality metrics for each broad level phoneme class. We have presented a preliminary portion of this study in [23] and [24]. For this purpose, we split phonemes into three broad classes—sonorants, obstruents, and silence. In GSS/MMSE/JMAP algorithms, the parameters of the gain functions influencing the enhancement performance are fixed over the entire utterance. In this paper, we develop parametric gain functions for each of these algorithms using flexible cost functions and generalized priors of the clean speech spectral magnitudes and/or phases and show that the parameters in these gain functions can be varied to obtain better enhancement levels per phoneme class than their corresponding baselines. However, no single set of parameters exist that attain equivalent enhancement potential across *all* classes. Although versatile, this drawback makes the parametric gain functions less attractive for real-world implementation.

[1]The term ROVER is a connotation to the National Institute of Standards and Technology (NIST) automatic speech recognition (ASR) system [25] which produces a composite ASR output when outputs from multiple ASR systems are available. In the context of NIST ASR, ROVER stands for Recognizer Output Voting Error Reduction. Since the enhancement system addressed in this study combines outputs from multiple estimators, it is appropriate to address this system using the term ROVER.

To overcome this limitation, we propose a novel ROVER[1] based solution. In the proposed scheme, for a given noisy utterance, three different prior enhanced utterances are generated from the parametric estimators—each customized for a specific phoneme class. The noisy utterance is partitioned into speech and non-speech regions using a voice activity detector (VAD). For a given short time frame, the ROVER solution applies soft decisions using a mixture maximum (MIXMAX) model [26] to weigh and combine the phoneme segments obtained from the prior enhanced utterances. This results in a single composite enhanced utterance that provides better levels of enhancement than the parametric estimators, and also reinforces the applicability of the ROVER solution across different enhancement algorithms.

The remainder of this paper is organized as follows. In Section II, derivations for the parametric estimators of GSS, MMSE, and JMAP algorithms are outlined. In Section III, the application of the MIXMAX model in making enhancement-based soft decisions is presented. A comprehensive objective quality evaluation is performed using the segmental SNR, the Itakura–Saito, and the DFT distortion for every broad level phoneme class at different global SNR levels in Section IV. Finally, conclusions are summarized in Section V.

II. Enhancement Models

A. Generalized Spectral Subtraction

Assuming noise is additive and statistically independent of the speech signal, the representation of noisy speech in the frequency domain can be given as follows:

$$Z_k e^{j\phi_{z,k}} = X_k e^{j\phi_{x,k}} + D_k e^{j\phi_{d,k}}. \tag{1}$$

Here, Z_k, X_k, D_k represent the spectral magnitudes and $\phi_{z,k}$, $\phi_{x,k}$, $\phi_{d,k}$ represent the phases of the discrete Fourier transforms (DFT) of noisy speech $(Z(\omega_k))$, clean speech $(X(\omega_k))$ and noise $(D(\omega_k))$, respectively, at frequency bin k. If $\hat{X}_k$ is the spectral magnitude estimate of clean speech obtained from the GSS [20] algorithm, then this can be represented as

$$\hat{X}_k^{\alpha} = a_k Z_k^{\alpha} - b_k E[D_k^{\alpha}] \tag{2}$$

where the term α is an exponent term with $\alpha > 0$, and a_k, b_k are weighting parameters for frequency bin k. Although these weighting parameters are functions of α, we will drop α from their subscripts for notational simplicity. In [20], the estimator in (2) is optimized by minimizing the MSE between X_k^{α} and $\hat{X}_k^{\alpha}$. In this paper, we define the generalized weighted Euclidean distortion (WED) between the clean speech and estimate of clean speech spectral magnitude in (3) and seek to minimize this error. If the clean speech spectral magnitude vector is $\mathbf{X} = [X_1, X_2, X_3, \ldots, X_K]^T$ in a short-time analysis frame of speech, then the WED error between clean speech and estimate of clean speech spectral magnitude vector is

$$C_{\epsilon}(\mathbf{X}^{\alpha}, \hat{\mathbf{X}}^{\alpha}) = (\mathbf{X}^{\alpha} - \hat{\mathbf{X}}^{\alpha})^T W (\mathbf{X}^{\alpha} - \hat{\mathbf{X}}^{\alpha}) \tag{3}$$

where $W = \mathrm{diag}(X_1^{\beta}, X_2^{\beta}, \ldots, X_K^{\beta})$, K is the length of the DFT of the analysis frame, and α, β are constant exponent

terms. It should be noted that when $\beta > 0$, the error function penalizes the errors in the spectral peaks more heavily than spectral valleys. When $\beta < 0$, the errors in the spectral valleys are penalized more than those in spectral peaks. Since MSE weighs the errors equally in all regions of the spectrum, the value of β in WED error offers flexibility in modifying the error function. The musical noise present in spectral subtraction approaches can be attributed to the random occurrence of sinusoidal peaks along the spectrum. In an additive white noise scenario, musical noise can be considered predominant in the region of spectral valleys since these are regions of low SNR. Therefore, we focus on values where $\beta < 0$.

Taking the expectation of the square of the magnitudes on both sides of (1) results in $E[Z_k^2] = E[X_k^2] + E[D_k^2]$ since the expectation of the cross-term is 0. Using the assumption that this can be replaced by the sample estimate of the ensemble average, we get $Z_k^2 \approx X_k^2 + D_k^2$. Further, in [20], the power term of 2 was replaced by α to form the ideal generalized model. With this the clean speech spectral magnitude can be written as

$$X_k^\alpha = Z_k^\alpha - D_k^\alpha. \tag{4}$$

Substituting (4) and (2) into (3), and finding the expectation of the WED estimation error at frequency bin k results in the following:

$$\begin{aligned} E[C_\epsilon] &= E\left[X_k^\beta \left((1-a_k)X_k^\alpha - a_k D_k^\alpha + b_k E\left[D_k^\alpha\right]\right)^2\right] \\ &= (1-a_k)^2 E\left[X_k^{\beta+2\alpha}\right] + a_k^2 E\left[X_k^\beta\right] E\left[D_k^{2\alpha}\right] \\ &\quad + \left(b_k^2 - 2a_k b_k\right) E\left[X_k^\beta\right] E^2\left[D_k^\alpha\right] \\ &\quad + 2(1-a_k)(b_k - a_k) E\left[X_k^{\beta+\alpha}\right] E\left[D_k^\alpha\right]. \end{aligned} \tag{5}$$

Assuming mutual independence between frequency components, we attempt to minimize the WED error at each frequency bin independently to minimize the WED error of a frame in (3). The optimum values of a_k, b_k that minimize the error in (5) can be obtained by differentiating (5) with respect to a_k, b_k separately and setting them to zero as

$$\frac{\partial E[C_\epsilon]}{\partial a_k} = 0, \quad \frac{\partial E[C_\epsilon]}{\partial b_k} = 0. \tag{6}$$

Solving for a_k, b_k, and letting $\delta_k = E^2[X_k^{\beta+\alpha}]/E[X_k^\beta]$, the optimum values are given by

$$a_k = \frac{E\left[X_k^{\beta+2\alpha}\right] - \delta_k}{E\left[X_k^{\beta+2\alpha}\right] - \delta_k + E\left[X_k^\beta\right]\left(E\left[D_k^{2\alpha}\right] - E^2\left[D_k^\alpha\right]\right)} \tag{7}$$

$$b_k = a_k - (1-a_k)\frac{E\left[X_k^{\beta+\alpha}\right]}{E\left[X_k^\beta\right] E\left[D_k^\alpha\right]}. \tag{8}$$

Assuming the case where real and imaginary parts of the clean speech and noise DFTs in the analysis frame are independent and Gaussian distributed with zero means, their spectral magnitudes will be Rayleigh distributed. The rth moment of a Rayleigh distributed probability density function (pdf) $f(X_k)$ can be simplified as

$$E\left[X_k^r\right] = \int_0^\infty X_k^r f(X_k) dX_k = \Gamma(r/2+1) E\left[X_k^2\right]^{r/2}. \tag{9}$$

For (9) to exist, the exponent of X_k in (9) must be greater than -1. Hence, the constraint is that $r > -2$. Substituting (9) into (7) and (8) further simplifies the optimal value of parameters to

$$a_k = \frac{\xi_k^\alpha \theta_1}{\xi_k^\alpha \theta_1 + \theta_2} \tag{10}$$

$$b_k = \frac{\xi_k^\alpha \theta_1 \Gamma(\alpha/2+1) - \xi_k^{\alpha/2}\theta_3}{\Gamma(\alpha/2+1)\left\{\xi_k^\alpha \theta_1 + \theta_2\right\}} \tag{11}$$

where $\Gamma(\cdot)$ represents the complete Gamma function and ξ_k is the *a priori* SNR at the kth frequency component given by

$$\xi_k = \frac{\lambda_X(k)}{\lambda_D(k)}. \tag{12}$$

Here, $\lambda_X(k) = E[X_k^2]$ and $\lambda_D(k) = E[D_k^2]$ are the variances of the clean speech and noise respectively. It is assumed that the real and imaginary parts of the the DFT coefficients of the clean speech have equal variances (i.e., $\lambda_{X_R}(k) = \lambda_{X_I}(k) = \lambda_X(k)/2$). The same is assumed for the noise as well. The constants $\theta_1, \theta_2, \theta_3$ are given by

$$\begin{aligned} \theta_1 &= \Gamma(\alpha+\beta/2+1)\Gamma(\beta/2+1) - \Gamma^2(\alpha/2+\beta/2+1) \\ \theta_2 &= \Gamma^2(\beta/2+1)\{\Gamma(\alpha+1) - \Gamma^2(\alpha/2+1)\} \\ \theta_3 &= \theta_2 \frac{\Gamma(\alpha/2+\beta/2+1)}{\Gamma(\beta/2+1)} \end{aligned} \tag{13}$$

which are functions of α and β. Substituting (10), (11), (12), (13) into (2), the gain function $\hat{G}_k = \hat{X}_k/Z_k$ can be written as

$$\hat{G}_k = \sqrt[\alpha]{\left(\frac{\xi_k^\alpha}{\xi_k^\alpha \theta_1 + \theta_2}\right)\left(\theta_1 - \left(\theta_1\Gamma(\alpha/2+1) - \theta_3 \xi_k^{-\alpha/2}\right)\gamma_k^{-\alpha/2}\right)} \tag{14}$$

where $\gamma_k = Z_k^2/E[D_k^2]$ is the *a posteriori* SNR at the kth frequency component. The gain equation of (14) may be considered as the GSS β-*unconstrained* parametric estimator. To arrive at the constrained estimator, the constraint $a_k = b_k$ is applied in (2). With this constraint, the expectation of the WED estimation error at frequency bin k is given by

$$\begin{aligned} E[C_\epsilon] = (1-a_k)^2 E\left[X_k^{\beta+2\alpha}\right] + a_k^2 E\left[X_k^\beta\right] E\left[D_k^{2\alpha}\right] \\ - a_k^2 E\left[X_k^\beta\right] E^2\left[D_k^\alpha\right]. \end{aligned} \tag{15}$$

The constrained optimum value of a_k that minimizes the error in (15) is

$$a_k = \frac{E\left[X_k^{\beta+2\alpha}\right]}{E\left[X_k^{\beta+2\alpha}\right] + E\left[X_k^\beta\right]\left(E\left[D_k^{2\alpha}\right] - E^2\left[D_k^\alpha\right]\right)}. \tag{16}$$

This differs from (7) by the term δ_k. This can be further simplified to give

$$a_k = \frac{\xi_k^{\alpha}}{\xi_k^{\alpha} + \theta_4} \tag{17}$$

where

$$\theta_4 = \frac{\Gamma(\beta/2+1)\{\Gamma(\alpha+1) - \Gamma^2(\alpha/2+1)\}}{\Gamma(\alpha+\beta/2+1)}. \tag{18}$$

Using (17), (18), the GSS β-*constrained* gain function can be represented as

$$\hat{G}_k = \sqrt[\alpha]{\left(\frac{\xi_k^{\alpha}}{\xi_k^{\alpha}+\theta_4}\right)\left(1 - \Gamma(\alpha/2+1)\gamma_k^{-\alpha/2}\right)}. \tag{19}$$

It is easy to see that if we set $\beta = 0$ in (14) or (19), then the gain functions of the GSS β-unconstrained and β-constrained estimators simplify to the gain functions of the GSS estimators derived by Sim *et al.* [[20], (32), (33)].

B. Weighted Euclidean Distortion Bayesian Estimator Based on Chi and Two-Sided Gamma Priors

In this section, we present the formulations of the generalized WED estimators for the cases of Ephraim-Malah MMSE spectral magnitude estimator [11] and Martin's DFT MMSE estimator [21]. Since $Z(\omega_k) = Z_k e^{j\phi_{z,k}}$ is the complex spectrum of noisy speech, the Bayesian estimator of the spectral magnitude that minimizes the WED error in (3) is given by

$$\hat{X}_k = \left\{\frac{E\left[X_k^{\alpha+\beta} \,\middle|\, Z(\omega_k)\right]}{E\left[X_k^{\beta} \,\middle|\, Z(\omega_k)\right]}\right\}^{\frac{1}{\alpha}}$$

$$= \left\{\frac{\int_0^{\infty} X_k^{\alpha+\beta}\, p(X_k \,|\, Z(\omega_k))\, dX_k}{\int_0^{\infty} X_k^{\beta}\, p(X_k \,|\, Z(\omega_k))\, dX_k}\right\}^{\frac{1}{\alpha}}$$

$$= \left\{\frac{\int_0^{\infty} X_k^{\alpha+\beta}\, p(Z(\omega_k) \,|\, X_k) p(X_k)\, dX_k}{\int_0^{\infty} X_k^{\beta}\, p(Z(\omega_k) \,|\, X_k) p(X_k)\, dX_k}\right\}^{\frac{1}{\alpha}} \tag{20a}$$

$$= \left\{\frac{\int_0^{\infty}\int_0^{2\pi} X_k^{\alpha+\beta} p\left(Z(\omega_k)|X_k, \phi_{X_k}\right) p\left(X_k, \phi_{X_k}\right) d\phi_{X_k} dX_k}{\int_0^{\infty}\int_0^{2\pi} X_k^{\beta}\, p\left(Z(\omega_k)|\, X_k, \phi_{X_k}\right) p\left(X_k, \phi_{X_k}\right)\, d\phi_{X_k} dX_k}\right\}^{\frac{1}{\alpha}} \tag{20b}$$

after canceling out the common term $p(Z(\omega_k))$. The resulting estimator in (20) depends on the choice of the distribution of the prior $p(X_k)$. Here, we first investigate the performance using the Chi prior to model the spectral magnitudes of the clean speech. The Chi probability density function is represented by [27]

$$p(X_k) = \frac{2X_k^{2a-1}}{\lambda_X^a(k)\Gamma(a)} \exp\left(-\frac{X_k^2}{\lambda_X(k)}\right) \tag{21}$$

where the term $2a$ indicates the number of degrees of freedom. Furthermore, the phase can be assumed to be independent of the spectral magnitude and uniformly distributed in $[-\pi, +\pi]$, which results in

$$p\left(X_k, \phi_{X_k}\right) = \frac{1}{2\pi} p(X_k). \tag{22}$$

Since it was assumed that the real and imaginary parts of the DFT of noise, D_R and D_I, would be Gaussian random variables with distribution $D_R, D_I \sim \mathcal{N}(0, \lambda_D(k)/2)$, the complex distribution of $p(Z(\omega_k) \,|\, X_k, \phi_{X_k})$ will be centered around the means X_R and X_I with the same variance $\lambda_D(k)/2$. Letting $\Delta_k = \mid Z(\omega_k) - X(\omega_k) \mid$, this can be written as

$$p\left(Z(\omega_k) \,|\, X_k, \phi_{X_k}\right) = \frac{1}{\pi\lambda_D(k)} \exp\left(-\frac{{\Delta_k}^2}{\lambda_D(k)}\right). \tag{23}$$

Inserting (21), (22), (23) into (20b) and using the simplification from [[28], (6.631.1), (8.406.3)], and the fact that $\hat{G}_k = \hat{X}_k/Z_k$, we obtain

$$\hat{G}_k = \frac{\sqrt{\nu_k}}{\gamma_k}\left\{\frac{\Gamma\left(\frac{\alpha+\beta}{2}+a\right)\phi\left(1-\frac{\alpha+\beta}{2}-a, 1; -\nu_k\right)}{\Gamma\left(\frac{\beta}{2}+a\right)\phi\left(1-\frac{\beta}{2}-a, 1; -\nu_k\right)}\right\}^{\frac{1}{\alpha}} \tag{24}$$

where

$$\nu_k = \frac{\xi_k}{1+\xi_k}\gamma_k$$
$$\gamma_k = \frac{Z_k^2}{\lambda_D(k)},\ (a\ posteriori\ \mathrm{SNR}) \tag{25}$$

and $\phi(a, b; x)$ denotes the confluent hypergeometric function. The conditions for which (24) is valid is given by

$$a > 0, \quad \alpha \neq 0, \quad \beta > -2a, \quad \alpha + \beta > -2a. \tag{26}$$

This implies that for a given value of a, both the parameters α, β can take on negative values which will be useful for achieving greater degrees of noise suppression. It may be noted that for $a = 1$ and $\alpha = 1$, the WED Chi estimator in (24) becomes the estimator derived by Loizou in [[12], (18)]. In [12], the constraint was $\beta > -2$. This constraint is relaxed in (24) since if $a > 1$, then β can take on values smaller than -2 which helps in penalizing any low SNR region spectral valleys. Also, if $a = 1, \beta = 0$, and α is replaced by 2α in (24), the solution becomes the power spectrum based generalized MMSE (GMMSE) estimator derived by Hansen *et al.* in [[8], (19)]. Furthermore, the Ephraim-Malah MMSE estimator [11] is a special case of the WED Chi estimator when $a = 1, \alpha = 1, \beta = 0$.

Next, we turn our attention to finding the WED optimized solution of the DFT MMSE estimator. In [21], Martin modeled the clean speech DFT coefficients using the two-sided Gamma prior $p(X_R)$ at frequency bin k as

$$p(X_R) = \frac{\sqrt{\mu}}{2\sqrt{\pi\sqrt{\lambda_X(k)}}} |X_R|^{-0.5} \exp\left(-\frac{\mu|X_R|}{\sqrt{\lambda_X(k)}}\right) \tag{27}$$

where $\mu = \sqrt{1.5}$ and R denotes the real part of the DFT coefficient at frequency k. The probability density function of the imaginary part, $p(X_I)$, can be obtained by replacing X_R with X_I in (27). The WED Gamma estimator can be obtained by replacing X_k and $Z(\omega_k)$ with X_R and Z_R, respectively, and dropping the phase term in (20b) since there is no phase component for Z_R. We continue with the same assumption made earlier that the noise DFT coefficients are Gaussian distributed. In general, it can be shown that for some power α

$$\int_{-\infty}^{\infty} X_R^{\alpha} p(Z_R \,|\, X_R) p(X_R)\, dX_R = T_1(\alpha) T_2(\alpha) \tag{28}$$

where

$$T_1(\alpha) = c\left(\frac{\lambda_D(k)}{2}\right)^{\frac{1}{2}(\alpha+\frac{1}{2})} \times \exp\left(-\frac{Z_R^2}{\lambda_D(k)}\right)\Gamma\left(\alpha+\frac{1}{2}\right) \tag{29}$$

$$T_2(\alpha) = \exp\left(\frac{G_{R-}^2}{2}\right) D_{-(\alpha+\frac{1}{2})}(\sqrt{2}G_{R-}) + (-1)^{\alpha}\exp\left(\frac{G_{R+}^2}{2}\right) D_{-(\alpha+\frac{1}{2})}(\sqrt{2}G_{R+}) \tag{30}$$

and

$$c = \frac{\sqrt{\mu}}{2\pi\sqrt{\lambda_D(k)\sqrt{\lambda_X(k)}}} \tag{31}$$

$$G_{R-} = \frac{1}{2}\frac{\mu}{\sqrt{\xi_k}} - \frac{Z_R}{\sqrt{\lambda_D(k)}} \tag{32}$$

$$G_{R+} = \frac{1}{2}\frac{\mu}{\sqrt{\xi_k}} + \frac{Z_R}{\sqrt{\lambda_D(k)}} \tag{33}$$

and $D_p(z)$ is the parabolic cylindrical function [[28], (9.241.2)] where p and z are its order and argument, respectively. Substituting (28)–(33) into (20a) with appropriate power terms, we arrive at the DFT WED Gamma estimator

$$\hat{G}_k = \sqrt{\frac{1}{\gamma_{R_k}}}\left\{\frac{\Gamma\left(\alpha+\beta+\frac{1}{2}\right) T_2(\alpha+\beta)}{\Gamma\left(\beta+\frac{1}{2}\right) T_2(\beta)}\right\}^{\frac{1}{\alpha}} \tag{34}$$

where $\gamma_{R_k} = Z_R^2/(0.5\lambda_D(k))$ is the *a posteriori* SNR of the real or imaginary part of the DFT coefficient. Unlike the previous *a posteriori* SNR defined in (25), γ_{R_k} has an extra 1/2 term in the denominator. This is because here, γ_{R_k} takes into account the variances of the real and imaginary part of the noisy speech DFT coefficients separately, whereas the *a posteriori* SNR of the spectral magnitude, γ_k, in (25) considers the sum of the variances of the real and imaginary part of the noisy speech DFT coefficients. The constraints in (34) are given by

$$\beta > -0.5, \quad \alpha+\beta > -0.5, \quad \alpha \in Z - \{0\}, \quad \beta \in Z^+ \cup \{0\}. \tag{35}$$

The integer constraint for α and β comes from the fact that for $T_2(\alpha+\beta)$ (and hence $\hat{G}_k$ (34)) to be real, the term $(-1)^{\alpha+\beta}$ in $T_2(\alpha+\beta)$ must be real which happens for all integral values of the sum $\alpha+\beta$. Since $T_2(\beta)$ is also a term in (34), β must be an integer. Therefore, combining the two constraints (i.e., $(\alpha+\beta) \in Z$ and $\beta \in Z$), implies that α must also be an integer. Along with the fact that $\alpha \neq 0$, restricts α to the set $\alpha \in Z - \{0\}$. Moreover, the constraint $\beta > -0.5$ and $\beta \in Z$ implies that $\beta_{\min} = 0$. Therefore, the possible values are $\beta = 0, 1, 2, \ldots$. Therefore, the minimum integer value of α satisfying $\alpha+\beta > -0.5$ is $\alpha_{\min} = -\beta$. Unfortunately, β cannot take on negative values here unlike (24). The DFT MMSE estimator in [[21], (13)] is therefore a special case of the DFT WED Gamma estimator (34) when $\alpha = 1, \beta = 0$.

C. Joint Maximum A Posteriori Estimator Based on Chi Prior

Here, we discuss the joint MAP estimate of speech spectral magnitude X_k and phase $\phi_{X,k}$ for the case of Chi prior in (21). This is given as

$$(\hat{X}_k, \hat{\phi}_{X,k}) = \arg\max_{X_k,\phi_{X,k}} p(X_k, \phi_{X,k} \,|\, Z(\omega_k)) = \arg\max_{X_k,\phi_{X,k}} \frac{p(Z(\omega_k) \,|\, X_k, \phi_{X,k}) p(X_k, \phi_{X,k})}{p(Z(\omega_k))}. \tag{36}$$

The denominator $p(Z(\omega_k))$ can be ignored since it is only a normalization term. For a rotational invariant pdf, and assuming a uniform distribution of phase in $[-\pi, +\pi]$, the relationship between the spectral magnitude and phase is given in (22). Since the natural logarithm function is monotonic increasing, the $\ln(\cdot)$ of (36) could be maximized in order to maximize (36) (note, this may be represented by L). Substituting (21), (22), and (23) into the $\ln(\cdot)$ function of (36) and ignoring $p(Z(\omega_k))$, we obtain

$$L = \ln(p(Z(\omega_k) \,|\, X_k, \phi_{X,k}) p(X_k) p(\phi_{X,k})) = -\frac{|Z(\omega_k) - X(\omega_k)|^2}{\lambda_D k} - \frac{X_k^2}{\lambda_X(k)} + (2a-1)\ln(X_k) + \delta \tag{37}$$

where $\delta = -\ln(\pi^2 \lambda_D(k) \lambda_X^a \Gamma(a))$ is a term-independent of X_k and $\phi_{X,k}$. After partially differentiating L with respect to $\phi_{X,k}$ and setting the derivative to zero, we obtain the optimal phase estimate to be the same as the noisy phase (i.e., $\hat{\phi}_{X,k} = \phi_{Z,k}$). Similarly, partially differentiating L with respect to X_k and setting the derivative to zero yields the quadratic

$$X_k^2 - \left(\frac{Z_k \xi_k}{1+\xi_k}\right) X_k - \left(\frac{(2a-1)\lambda_d(k)\xi_k}{1+\xi_k}\right) = 0. \tag{38}$$

The root of the quadratic is that value of X_k which maximizes (36), and hence this root must be $\hat{X}_k$. Therefore, setting $\hat{X}_k = \hat{G}_k Z_k$ we obtain

$$\hat{G}_k = \frac{\xi_k + \sqrt{\xi_k^2 + 4\left(a-\frac{1}{2}\right)(1+\xi_k)\left(\frac{\xi_k}{\gamma_k}\right)}}{2(1+\xi_k)}. \tag{39}$$

The constraints of (39) are given by

$$a > 0, \quad a \geq \frac{1}{2} - \frac{\nu_k}{4} \tag{40}$$

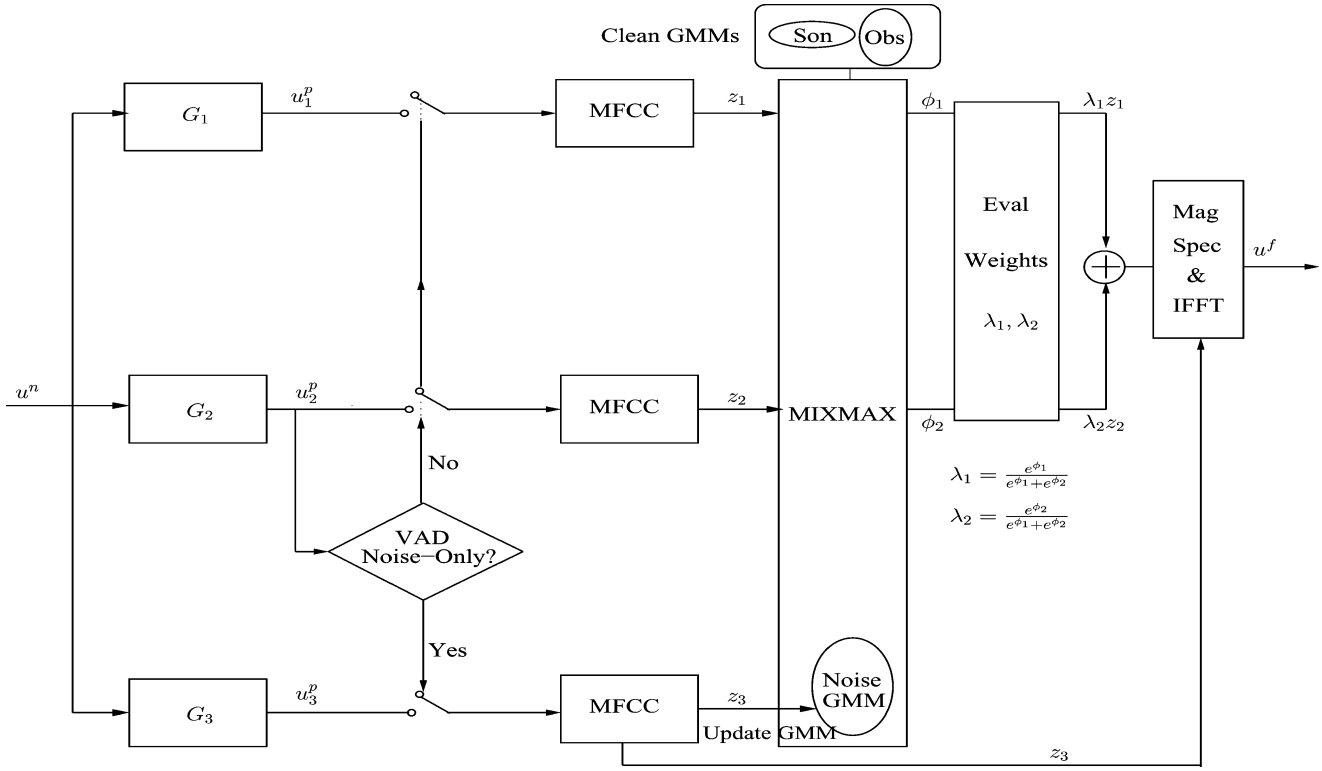


Fig. 1. ROVER enhancement framework using the MIXMAX model.

where ν_k is defined as in (25). The second constraint should be satisfied in order to obtain real values from the square root term in (39). It may be noted that the joint MAP estimate of the Chi prior magnitude in (39) differs from the joint MAP estimate of Wolfe and Godsil [[22], (29)] by the term $4(a-1/2)$ inside the square root instead of 2. Therefore, this represents a unifying solution. By setting $a = 1$ in (39), results in the special case gain function of [22].

III. Rover Enhancement Using MIXMAX Model

The drawback with the parametric estimators is that a single set of parameters does not exist that can generate reasonably good enhancement levels across all phoneme classes. This will be evident in Section IV. Therefore, the need arises to create different enhanced utterances for each class.

In Section II, we developed parametric estimators whose parameters can be modified to achieve enhancement on a per phoneme class level. We assume that phonemes may be classified into three groups of broad phone classes (BPC)—sonorants, obstruents, and silence. Therefore, as illustrated in Fig. 1, for a given noisy utterance u^n, three utterances u_1^p, u_2^p, u_3^p may be generated from the three parametric estimators G_1, G_2, G_3, respectively, with each estimator having its parameters set to generate improved enhancement levels for a specific BPC. Hence, u_1^p is expected to possess improved enhancement levels for regions dominated by sonorants *only* disregarding the enhancement levels in regions of obstruents and silence. Similarly, u_2^p is expected to possess better enhancement levels for regions of obstruents *only* disregarding the enhancement levels in the regions of sonorants and silence. The superscript p in $u_{(\cdot)}^p$ indicates that the utterances are the outputs of the *parametric* estimators. For this reason, $u_{(\cdot)}^p$ may also be referred to as the *prior* enhanced utterances. It is to be noted that the three estimators G_1, G_2, G_3 are homogeneous, i.e., they have the same gain function with different values of parameters. Therefore, as an example, a homogeneous system using the GSS β-unconstrained estimator will use the same gain function (14) in G_1, G_2, G_3 for the three BPCs but with three different values of the parameter set $\{\alpha, \beta\}$. Furthermore, a voice activity detector (VAD) was employed to detect the speech active and inactive regions. During speech active regions, the top two switches connecting u_1^p, u_2^p in Fig. 1 were enabled whereas the bottom switch connecting u_3^p was disabled. During speech inactive regions, the bottom switch connecting u_3^p was enabled to update the noise GMM whereas the top two switches were disabled.

At this point, we have a homogeneous system generating three prior enhanced utterances for the three BPCs. However, we still need to present the end user with a single composite enhanced utterance constructed from the three prior enhanced utterances. To achieve this, a phoneme class labeler is required. The three prior enhanced utterances represent a preliminary level of the enhancement framework whereas the single composite utterance represent the final level of enhancement framework.

In this section, we present the final level of enhancement framework using the MIXMAX model to combine the outputs of the parametric estimators. The MIXMAX model is a phoneme class labeler using a probabilistic rule that associates a frame belonging to a specific BPC. It was first introduced by Nádas *et al.* [26] especially for noisy speech recognition and

later used for class independent MMSE speech enhancement by Burshtein and Gannot [29]. Our motivation for use of the MIXMAX model is that, for a given segment with a fixed time duration, we can apply the probabilistic rule on each of the three BPC specific utterances to evaluate the possibility of it belonging to a specific BPC. Using this, we can combine the segments from the three utterances to generate a composite utterance. Further, since the three utterances have improved enhancement levels per BPC they are expected to give higher recognition accuracies than the noisy utterance u^n. This justifies the application of the MIXMAX model on the three prior enhanced utterances $u^p_{(.)}$ instead of the noisy utterance u^n.

Here, we modify the notation slightly from the previous sections. Let $\mathbf{X} = [X_1, X_2, \ldots, X_K]^T$ denote the random vector representing the Mel frequency cepstral coefficients (MFCC) of the clean speech sonorants with the kth component being X_k, where $k = 1, \ldots, K$ with K being the size of the MFCC vector. The pdf of $\mathbf{X}$ can be modeled with a Gaussian mixture model (GMM) with each mixture consisting of K components and a diagonal covariance matrix as follows:

$$f(\mathbf{x}) = \sum_i c_x(i) \prod_k f_{i,k}(x) \tag{41}$$

where

$$f_{i,k}(x) = \frac{1}{\sqrt{2\pi\sigma^2_{i,k}}} \exp\left\{ -\frac{(x-\mu_{i,k})^2}{2\sigma^2_{i,k}} \right\} \tag{42}$$

and $c_x(i)$ is the weight of the ith mixture. Similarly, let the MFCC vectors for 1) obstruents in clean speech be represented by $\mathbf{Y}$, 2) noisy speech be $\mathbf{Z}$, and 3) noise be $\mathbf{D}$. As in (42), the kth component of the jth GMM mixture representing $\mathbf{Y}$ can be given by the pdf $t_{j,k}(y) \sim \mathcal{N}(\mu_{j,k}, \sigma_{j,k})$ with mixture weight $c_y(j)$. Similar to [29], we assume the modeling of noise can be achieved using a single mixture K dimensional Gaussian represented by

$$g(\mathbf{d}) = \prod_k g_k(d) \tag{43}$$

where

$$g_k(d) = \frac{1}{\sqrt{2\pi\sigma^2_{D,k}}} \exp\left\{ -\frac{(d-\mu_{D,k})^2}{2\sigma^2_{D,k}} \right\}. \tag{44}$$

The GMMs were trained using clean speech and the details of the training parameters are outlined in Section IV. Assuming zero means and statistical independence between the MFCC components of clean speech and noise, we find the pdf of the noisy speech model using the MIXMAX model

$$\mathbf{Z} \approx \max(\mathbf{X}, \mathbf{Y}, \mathbf{D}). \tag{45}$$

The max-operation is performed separately in each component of the three vectors $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{D}$. Nádas *et al.* [26] assumed that the log-spectral components of the noisy speech can be modeled by the MIXMAX model since the log-energies of the noisy speech computed over the different frequency bands of speech spectrum could be represented by the maximum of the speech signal and noise. In our study, we replaced the log-spectral components by MFCCs. Our motivation for using MFCCs instead of log-spectral components stems from the weak assumption that the log-spectral components are independent of each other and that diagonal covariance matrices can be used to model GMM mixtures. However, with MFCCs, the DCT transform decorrelates the MFCC components justifying the use of diagonal matrices. Also, MFCCs provide a more compact representation of the signal than log-spectral components thereby reducing the number of modeling parameters.

To find the pdf of Z_k in (45), we determine the cumulative distribution function (cdf) $H_{i,j,k}(z)$ of Z_k given by

$$\begin{aligned} &H_{i,j,k}(z) \\ &\quad = P(Z_k < z \mid I = i, J = j) \\ &\quad = P(\max(X_k, Y_k, D_k) < z \mid I = i, J = j) \\ &\quad = P(X_k < z, Y_k < z, D_k < z \mid I = i, J = j) \\ &\quad = P(X_k < z \mid I = i) P(Y_k < z \mid J = j) P(D_k < z) \\ &\quad = F_{i,k}(z) T_{j,k}(z) G_k(z) \\ &\qquad \text{(since mutual independence of } \mathbf{X}, \mathbf{Y}, \mathbf{D}) \end{aligned} \tag{46}$$

where $F_{i,k}(z), T_{j,k}(z), G_k(z)$ represent the cdf's of $f_{i,k}(x), t_{j,k}(y), g_k(d)$ respectively. The pdf of Z_k modeled using (45) is obtained by differentiating (46) with respect to z as follows:

$$\begin{aligned} \frac{dH_{i,j,k}(z)}{dz} = h_{i,j,k}(z) &= f_{i,k}(z) T_{j,k}(z) G_k(z) \\ &+ F_{i,k}(z) t_{j,k}(z) G_k(z) + F_{i,k}(z) T_{j,k}(z) g_k(z). \end{aligned} \tag{47}$$

During testing, for a noisy utterance input, a voice activity detector (VAD) was used to classify regions of voice inactivity. The statistics of the noise probability density function of (43) were updated during regions of voice inactivity [29]. For the voice active regions, the noisy speech component z was substituted in (47) to evaluate the probability of z belonging to sonorant, obstruent, and noise. Examining this further, the first term in (47) can be expanded as

$$\begin{aligned} &f_{i,k}(z) T_{j,k}(z) G_k(z) \\ &\quad = p(X_k = z \mid I = i) p(Y_k < z \mid J = j) p(D_k < z). \end{aligned} \tag{48}$$

This means that for a given combination of ith and jth mixture, $X_k = Z_k = z$, and $Y_k < X_k, D_k < X_k$. Hence, X_k is the maximum and z likely belongs to a sonorant. Similarly, the term $F_{i,k}(z) t_{j,k}(z) G_k(z)$ indicates Y_k is the maximum, and $F_{i,k}(z) T_{j,k}(z) g_k(z)$ indicates D_k is the maximum. The probability that X_k is the maximum value over *all* possible combinations of ith and jth mixtures can be represented by

$$\begin{aligned} &p(X_k = z \mid Z_k = z) \\ &\quad = \sum_{i,j} p(I = i, J = j \mid Z_k = z) \\ &\qquad \times p(X_k = z \mid Z_k = z, I = i, J = j) \\ &\quad = \sum_{i,j} \omega_{i,j,k}(z) \left\{ \frac{f_{i,k}(z) T_{j,k}(z) G_k(z)}{h_{i,j,k}(z)} \right\}. \end{aligned} \tag{49}$$

The numerator term $f_{i,k}(z)T_{j,k}(z)G_k(z)$ in $\{\cdot\}$ indicates the joint probability of $X_k = z$ and $Z_k = z$ given the ith and jth mixture (and hence X_k being the maximum) is the same as (48). The term $h_{i,j,k}(z)$ is a normalization factor and $\omega_{i,j,k}(z)$ is the weight for each combination of i, j mixture considered in $\sum_{i,j}$. In a similar fashion, the probability that Y_k is the maximum may be represented by

$$\begin{aligned} p(Y_k = z \mid Z_k = z) &= \sum_{i,j} p(I = i, J = j \mid Z_k = z) \\ &\quad \times p(Y_k = z \mid Z_k = z, I = i, J = j) \\ &= \sum_{i,j} \omega_{i,j,k}(z) \left\{ \frac{F_{i,k}(z) t_{j,k}(z) G_k(z)}{h_{i,j,k}(z)} \right\}. \end{aligned} \quad (50)$$

Furthermore, $\omega_{i,j,k}(z)$ may be considered a weighting term that is related to the *a posteriori* probability

$$\begin{aligned} \omega_{i,j,k}(z) &= p(I = i, J = j \mid Z_k = z) \\ &= \frac{p(Z_k = z \mid I = i, J = j) p(I = i) p(J = j)}{p(Z_k = z)} \\ &= \frac{p(Z_k = z \mid I = i, J = j) p(I = i) p(J = j)}{\sum_{i,j} p(Z_k = z \mid I = i, J = j) p(I = i) p(J = j)} \\ &= \frac{h_{i,j,k}(z) c_x(i) c_y(j)}{\sum_{i,j} h_{i,j,k}(z) c_x(i) c_y(j)}. \end{aligned} \quad (51)$$

As noted earlier, outputs u_1^p, u_2^p of the parametric estimators are present prior to the MIXMAX evaluation. For a given time period spanning the duration of a frame, let $\mathbf{z_1}$ and $\mathbf{z_2}$ be the MFCC vectors from u_1^p and u_2^p, respectively, as illustrated in Fig. 1. In (49), the probability of the observed noisy feature Z_k belonging to sonorant is evaluated for every dimension. In (52) below, the probability that the observed noisy feature vector $\mathbf{z_1}$ belonging to a sonorant is given as the products of the individual dimensions as

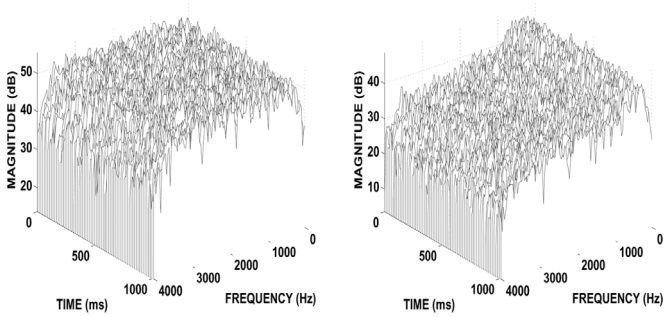
$$p(\mathbf{X} = \mathbf{z_1} \mid \mathbf{Z} = \mathbf{z_1}) = \prod_{k=1}^{K} p(X_k = z_{1,k} \mid Z_k = z_{1,k}). \quad (52)$$

Similarly, the probability that the observed noisy feature vector $\mathbf{z_2}$ belonging to an obstruent is the products of the individual dimensions in (50) and is given as

$$p(\mathbf{Y} = \mathbf{z_2} \mid \mathbf{Z} = \mathbf{z_2}) = \prod_{k=1}^{K} p(Y_k = z_{2,k} \mid Z_k = z_{2,k}) \quad (53)$$

At this point, we have two MFCC vectors $\mathbf{z_1}$ and $\mathbf{z_2}$ of a frame and the probabilities in (52) and (53) represent the scores of these vectors belonging to sonorants or obstruents respectively. Since we do not know the BPC of the frame and that we need to construct a single composite MFCC vector from both $\mathbf{z_1}, \mathbf{z_2}$ we assign weights to the vectors $\mathbf{z_1}, \mathbf{z_2}$ by normalizing the individual scores. We do this by taking the logarithm of (52) and (53) and denoting their log probabilities as ϕ_1 and ϕ_2, respectively. Therefore, the resulting MFCC vector is given by

$$\{\hat{\mathbf{X}} \text{ or } \hat{\mathbf{Y}}\} = \frac{e^{\phi_1}}{e^{\phi_1} + e^{\phi_2}} \mathbf{z_1} + \frac{e^{\phi_2}}{e^{\phi_1} + e^{\phi_2}} \mathbf{z_2}. \quad (54)$$



(a) (b)

Fig. 2. Time and frequency characteristics of (a) flat communications channel noise (FLN) and (b) large crowd noise (LCR).

The MFCC estimate ($\hat{\mathbf{X}}$ or $\hat{\mathbf{Y}}$) is converted back into the magnitude spectrum [30]. Using the magnitude spectrum and noisy phase, the composite speech signal is reconstructed by the standard overlap-add method.

IV. Experimental Results

A set of 32 (16 females, 16 males) phonetically balanced utterances from the TIMIT test corpus, downsampled from 16 kHz to 8 kHz, were used for objective quality evaluations. Test utterances were degraded with two noise types—flat communications channel noise (FLN, mostly stationary), and large crowd noise (LCR, mostly nonstationary). Test utterances were degraded at global SNRs of -5, 0, 5, and 10 dB. The time versus frequency responses of FLN and LCR noises are illustrated in Fig. 2. The quality of enhanced speech was assessed using objective speech quality measures such as the segmental SNR (SegSNR), Itakura–Saito distortion (IS), and DFT distortion. Higher values of SegSNR and lower values of IS or DFT distortion represent better speech quality. Segmental SNR evaluations were limited to the range of -10 dB to 35 dB.

The clean speech GMMs for BPCs used in the MIXMAX model were trained using 39-dimensional MFCC vectors extracted from 300 separate utterances taken from the TIMIT training corpus. GMMs for sonorants and obstruents were modeled using 16 mixtures, whereas for noise only 1 mixture was used. We found that generating these MFCC vectors from a small frame size of 10 ms and skip rate of 5 ms resulted in the best performance. This is because with a small frame size and skip rate, the likelihood of missing phone-to-phone boundaries is lower than with a larger frame size. Hence, the same frame size and skip rate was maintained during the testing phase. Furthermore, for MFCC extraction, each frame was windowed using a Hamming window and speech spectrum split into 24 triangular filters over the Mel scale. During testing, a statistical model-based VAD [[31], (11.1)] was used to classify regions of voice inactivity. Based on these VAD decisions, the statistics of the noise probability density function of (43) were continuously updated. During construction of the voice inactive regions in the final composite utterance, frames obtained from the prior enhanced utterance customized for silence (i.e., u_3^p) were used. For the remaining regions comprising of sonorants and obstruents, the MIXMAX weighting of (54) was used. The gain functions derived in Section II are functions of *a priori* SNR ξ_k (12) and *a posteriori* SNR γ_k (25). In (12), since clean

TABLE I
SEGMENTAL SNR AND ITAKURA–SAITO DISTORTION FOR SPEECH DEGRADED BY FLN/LCR NOISE AT GLOBAL SNRS OF −5, 0, 5, 10 dB. (a) FLN NOISE, (b) LCR NOISE

(a)

BPC	SegSNR (dB)				IS Distortion			
	-5 dB	0 dB	5 dB	10 dB	-5 dB	0 dB	5 dB	10 dB
Son	-3.13	0.15	4.24	8.72	6.74	6.03	5.33	4.68
Obs	-9.60	-8.85	-7.25	-4.59	8.76	7.90	6.97	6.04
Sil	-9.99	-9.97	-9.94	-9.72	11.38	10.46	9.51	8.44
Ovl	-5.93	-3.74	-0.74	2.87	7.82	7.04	6.25	5.47

(b)

BPC	SegSNR (dB)				IS Distortion			
	-5 dB	0 dB	5 dB	10 dB	-5 dB	0 dB	5 dB	10 dB
Son	-1.58	0.96	5.08	9.66	5.00	4.80	4.19	3.73
Obs	-9.02	-8.55	-6.80	-3.91	6.99	6.63	5.82	5.05
Sil	-9.99	-9.97	-9.91	-9.68	9.66	8.96	7.91	6.89
Ovl	-4.79	-3.13	-0.08	3.67	6.09	5.77	5.06	4.45

speech is unknown $\lambda_X(k)$ cannot be determined. Hence, we adopt the "decision-directed" approach of [[11], (51)] to predict the *a priori* SNR. Calculation of *a posteriori* SNR is straight forward since it involves taking square of the magnitude of the noisy speech in (25).

From this point onward, the following naming conventions are used to identify the parametric estimators—GSS β-unconstrained (GBU) (14), GSS β-constrained (GBC) (19), WED Chi (WC) (24), and JMAP Chi (JC) (39). Their corresponding ROVER implementations are denoted by—ROVER GSS β-unconstrained (RGBU), ROVER GSS β-constrained (RGBC), ROVER WED Chi (RWC), and ROVER JMAP Chi (RJC). Similarly, the baseline estimators are denoted by—GSS unconstrained (GU), GSS constrained (GC), MMSE, and JMAP. The MATLAB programs in [31] were used to execute the MMSE and JMAP algorithms.

The *improvement in segmental SNR* is a measure of increase in SNR over noisy speech, whereas the *IS improvement* indicates a decrease in IS distortion from noisy speech. These metrics can be represented by

$$\begin{aligned}&\text{SegSNR Improve}\\&\quad=\text{SegSNR Enhanced}-\text{SegSNR Noisy},\\&\text{IS Improve}\\&\quad=\text{IS Distortion Noisy}-\text{IS Distortion Enhanced}.\end{aligned}\tag{55}$$

Hence, in both metrics, the greater the improvement, the larger the objective quality of speech. The segmental SNR and IS distortion values of noisy speech in FLN and LCR noises at global SNRs of −5, 0, 5, 10 dB are given in Table I(a) and (b), respectively.

A. Performance Across Phoneme Classes: ROVER Versus Parametric Estimators Using Best Configurable Parameters

A comparison of the improvement metrics defined in (55) for ROVER enhancement versus its corresponding parametric and baseline estimators across different BPCs is tabulated in Table II(a)–(d) for the case of speech degraded by FLN noise at a global SNR of 0 dB. In addition, these tables include the values of the tunable parameters used in the parametric estimators of GBU, GBC, WC, JC to generate the best objective quality for a particular BPC. The tunable parameters used in the algorithms are indicated as $\langle a, \alpha, \beta\rangle$. A "-" in place of a tunable parameter indicates that the corresponding parameter is not applicable for that estimator. The subscripts—S(sonorants), O(obstruents), N(silence), Ovl(overall)—denote that the parameters $\langle a, \alpha, \beta\rangle$ were tuned to generate the best enhancement quality for a single BPC while overlooking the quality of all other BPCs. For example, in Table II(a), the improvement in segmental SNR of sonorants for $\mathrm{WC}_S\langle 0.10, 2.50, 1.00\rangle$ (4.58) is better than $\mathrm{WC}_O\langle 0.10, 0.50, 0.00\rangle$ (2.61), $\mathrm{WC}_N\langle 0.05, 0.50, 0.00\rangle$ (2.39), or $\mathrm{WC}_{\mathrm{Ovl}}\langle 0.10, 0.50, 0.00\rangle$ (2.61). However, the quality of other BPCs in $\mathrm{WC}_S\langle 0.10, 2.50, 1.00\rangle$ (i.e., obstruents and silence) is compromised at the expense of generating the best quality of sonorants. It can be inferred that the segmental SNR of sonorants enhanced by $\mathrm{WC}_S\langle 0.10, 2.50, 1.00\rangle$ is 4.73 since segmental SNR of sonorants degraded in FLN noise at 0-dB SNR is 0.15 [from Table I(a)] and segmental SNR improvement is 4.58 [from Table II(a)]. The three prior enhanced utterances used in ROVER enhancement algorithms do not use the same parameters as those in the best BPC specific configurations. The enhancement parameters used for RGBC, RGBU, RWC, and RJC are shown in Table III. These parameters are used for the remaining experiments discussed in the paper. The parameters used for the DFT WED Gamma estimator are given separately in Section IV-C.

In Table II(a), the segmental SNR improvement for sonorants in WC_S, obstruents in WC_O, silence regions in WC_N, and overall in $\mathrm{WC}_{\mathrm{Ovl}}$ are greater than those for baseline $\mathrm{MMSE}\langle 1.00, 1.00, 0.00\rangle$. However, none of the WC_S, WC_O, or WC_N configurations provide improved levels of enhancement across all BPCs. WC_O and WC_N perform well in obstruents and silence regions, but the segmental SNR of their sonorants is lower by approximately 2 dB versus the baseline MMSE estimator. Similarly, obstruents and silence regions in WC_S are lower than those in $\mathrm{WC}_{\mathrm{Ovl}}$ by about 4 dB and 9 dB, respectively. This is most likely the cause for a lower improvement in segmental SNR in the overall utterance of WC_S (4.55) compared to WC_O (5.35), WC_N (5.27), or $\mathrm{WC}_{\mathrm{Ovl}}$ (5.35). However, in RWC, all BPCs are jointly enhanced versus MMSE, or any BPC specific enhancement configuration (i.e., WC_S, WC_O, WC_N), or even $\mathrm{WC}_{\mathrm{Ovl}}$. It is noted that the score for sonorants in RWC (4.51) is marginally greater than $\mathrm{MMSE}\langle 1.00, 1.00, 0.00\rangle$ (4.47) and marginally lower than WC_S (4.58). Similarly, the score of obstruents in RWC (8.52) is smaller than obstruents in WC_O (9.16) by 0.64 dB, and the score of silence regions in RWC (6.79) is smaller than silence regions in WC_N (9.89) by 3.1 dB. In general, the objective quality of a BPC in ROVER enhancement may undergo some distortion when compared to the same BPC in its BPC specific best enhancement configuration. However, when all BPCs are jointly considered, ROVER is expected to achieve the best overall score for the utterance. A similar effect can be observed in the IS distortions in Table II(a). The IS improvement for the overall utterance in RWC (4.62) is marginally greater than WC_S (4.57), and WC_O (4.59), and marginally lower than $\mathrm{WC}_{\mathrm{Ovl}}$ (4.70). Similarly, overall IS improvements for RJC

TABLE II
SEGMENTAL SNR AND ITAKURA–SAITO IMPROVEMENT ACROSS BPCS CORRUPTED BY FLN NOISE AT SNR OF 0 dB FOR BASELINE VERSUS PARAMETRIC (USING BEST CONFIGURABLE PARAMETERS PER BPC) VERSUS ROVER ALGORITHMS. (a) MMSE VERSUS WC VERSUS RWC, (b) JMAP VERSUS JC VERSUS RJC, (c) GU VERSUS GBU VERSUS RGBU, (d) GC VERSUS GBC VERSUS RGBC

(a)

Enhancement$\langle a, \alpha, \beta\rangle$	Rise in SegSNR (dB)				Enhancement$\langle a, \alpha, \beta\rangle$	Reduction in IS			
	Son	Obs	Sil	Ovl		Son	Obs	Sil	Ovl
MMSE$\langle 1.00, 1.00, 0.00\rangle$	4.47	4.14	0.53	4.16	MMSE$\langle 1.00, 1.00, 0.00\rangle$	2.47	3.01	3.25	2.72
WC$_S\langle 0.10, 2.50, 1.00\rangle$	4.58	5.02	0.88	4.55	WC$_S\langle 0.05, 1.50, 0.50\rangle$	4.00	5.25	5.73	4.57
WC$_O\langle 0.10, 0.50, 0.00\rangle$	2.61	9.16	9.33	5.35	WC$_O\langle 0.05, 0.50, 0.50\rangle$	3.90	5.52	5.88	4.59
WC$_N\langle 0.05, 0.50, 0.00\rangle$	2.39	9.09	9.89	5.27	WC$_N\langle 0.05, 1.00, 0.00\rangle$	-3.09	0.11	6.65	-0.72
WC$_{Ovl}\langle 0.10, 0.50, 0.00\rangle$	2.61	9.16	9.33	5.35	WC$_{Ovl}\langle 0.05, 1.00, 0.50\rangle$	3.96	5.47	5.76	4.70
RWC	4.51	8.52	6.79	6.55	RWC	3.86	5.55	6.35	4.62

(b)

Enhancement$\langle a, \alpha, \beta\rangle$	Rise in SegSNR (dB)				Enhancement$\langle a, \alpha, \beta\rangle$	Reduction in IS			
	Son	Obs	Sil	Ovl		Son	Obs	Sil	Ovl
JMAP$\langle 1.00, -, -\rangle$	4.69	4.81	0.90	4.54	JMAP$\langle 1.00, -, -\rangle$	2.78	3.51	3.85	3.12
JC$_S\langle 0.80, -, -\rangle$	4.72	6.36	2.08	5.22	JC$_S\langle 0.50, -, -\rangle$	3.88	5.37	6.18	4.54
JC$_O\langle 0.07, -, -\rangle$	3.33	8.93	8.22	5.72	JC$_O\langle 0.50, -, -\rangle$	3.88	5.37	6.18	4.54
JC$_N\langle 0.05, -, -\rangle$	3.31	8.93	8.23	5.71	JC$_N\langle 0.25, -, -\rangle$	2.19	3.58	6.48	3.01
JC$_{Ovl}\langle 0.25, -, -\rangle$	3.61	8.93	8.16	5.87	JC$_{Ovl}\langle 0.50, -, -\rangle$	3.88	5.37	6.18	4.54
RJC	4.91	8.79	7.90	6.89	RJC	3.75	5.32	6.47	4.46

(c)

Enhancement$\langle a, \alpha, \beta\rangle$	Rise in SegSNR (dB)				Enhancement$\langle a, \alpha, \beta\rangle$	Reduction in IS			
	Son	Obs	Sil	Ovl		Son	Obs	Sil	Ovl
GU$\langle -, 1.00, 0.00\rangle$	4.39	3.36	0.20	3.85	GU$\langle -, 1.00, 0.00\rangle$	2.51	2.93	3.11	2.71
GBU$_S\langle -, 1.50, -1.00\rangle$	4.85	5.25	0.88	4.79	GBU$_S\langle -, 0.50, -1.25\rangle$	4.28	5.36	5.82	4.77
GBU$_O\langle -, 1.00, -1.75\rangle$	3.94	7.77	4.23	5.41	GBU$_O\langle -, 1.50, -1.75\rangle$	3.74	5.63	6.19	4.55
GBU$_N\langle -, 2.00, -1.75\rangle$	-2.70	7.12	7.45	1.45	GBU$_N\langle -, 1.75, -1.75\rangle$	1.97	4.73	6.89	3.34
GBU$_{Ovl}\langle -, 1.25, -1.75\rangle$	3.98	7.76	4.37	5.44	GBU$_{Ovl}\langle -, 0.5, -1.50\rangle$	4.17	5.59	5.92	4.82
RGBU	5.18	7.89	4.97	6.58	RGBU	4.32	5.79	6.54	5.01

(d)

Enhancement$\langle a, \alpha, \beta\rangle$	Rise in SegSNR (dB)				Enhancement$\langle a, \alpha, \beta\rangle$	Reduction in IS			
	Son	Obs	Sil	Ovl		Son	Obs	Sil	Ovl
GC$\langle -, 2.00, 0.00\rangle$	4.75	4.19	0.51	4.35	GC$\langle -, 2.00, 0.00\rangle$	2.79	3.46	3.63	3.08
GBC$_S\langle -, 1.25, -0.75\rangle$	4.95	4.93	0.73	4.74	GBC$_S\langle -, 0.50, -1.75\rangle$	3.92	4.46	4.74	4.18
GBC$_O\langle -, 1.75, -1.75\rangle$	4.02	6.09	1.60	4.68	GBC$_O\langle -, 0.25, 0.00\rangle$	3.77	4.48	4.78	4.09
GBC$_N\langle -, 1.25, -1.75\rangle$	3.34	5.81	1.68	4.18	GBC$_N\langle -, 0.25, 0.00\rangle$	3.77	4.48	4.78	4.09
GBC$_{Ovl}\langle -, 1.25, -1.25\rangle$	4.68	5.61	1.12	4.91	GBC$_{Ovl}\langle -, 0.50, -1.75\rangle$	3.92	4.46	4.74	4.18
RGBC	5.11	6.35	1.89	5.75	RGBC	3.87	4.37	4.64	4.12

TABLE III
ENHANCEMENT PARAMETERS USED IN DIFFERENT ROVER ALGORITHMS

Algorithm, BPC	Parameters			Algorithm, BPC	Parameters		
	a	α	β		a	α	β
RGBU, Son	-	1.25	-1.25	RWC, Son	1.00	2.50	-1.00
RGBU, Obs	-	1.25	-1.50	RWC, Obs	1.00	0.50	-1.50
RGBU, Sil	-	1.25	-1.75	RWC, Sil	1.00	0.50	-1.75
RGBC, Son	-	1.50	-1.00	RJC, Son	0.75	-	-
RGBC, Obs	-	1.50	-1.50	RJC, Obs	0.25	-	-
RGBC, Sil	-	2.00	-1.75	RJC, Son	0.05	-	-

(4.46) in Table II(b) and RGBC (4.12) in Table II(d) undergo a marginal degradation in the range 0.06–0.08 compared to JC$_{\text{Ovl}}$ (4.54) and GBC$_{\text{Ovl}}$ (4.18) respectively. This might prompt us to choose the *Ovl* configuration over ROVER. However, there are two drawbacks to note. First, the tunable parameters for WC$_{\text{Ovl}}$/JC$_{\text{Ovl}}$/GBC$_{\text{Ovl}}$ vary across a wide range of SNRs and noise types. This requires an exhaustive search over $\langle a, \alpha, \beta\rangle$ to find the *Ovl* parameters which makes it practically infeasible to implement in real-time scenarios (it is noted that this would be acceptable for offline enhancement processing). Second, it is not warranted that the parameters used in the *Ovl* configuration will always increase objective quality over all BPCs up to the levels achieved in ROVER. From our experiments, we found this to be true during segmental SNR evaluations of the overall utterance in Table II(a)–(d) where RWC/RJC/RGBU/RGBC always outperformed WC$_{\text{Ovl}}$/JC$_{\text{Ovl}}$/GBU$_{\text{Ovl}}$/GBC$_{\text{Ovl}}$, respectively.

B. Performance Across Phoneme Classes: Rover Versus Baseline Estimators

Segmental SNR and IS distortion performance at global SNRs of -5, 0, 5, and 10 dB are plotted across all baseline algorithms and their corresponding ROVER based versions in Figs. 3–6. In Fig. 3(a) and (b), segmental SNR improvement is evaluated for speech degraded by FLN noise. To gauge which ROVER algorithm had the greatest improvement over its corresponding baseline, we calculated the difference in

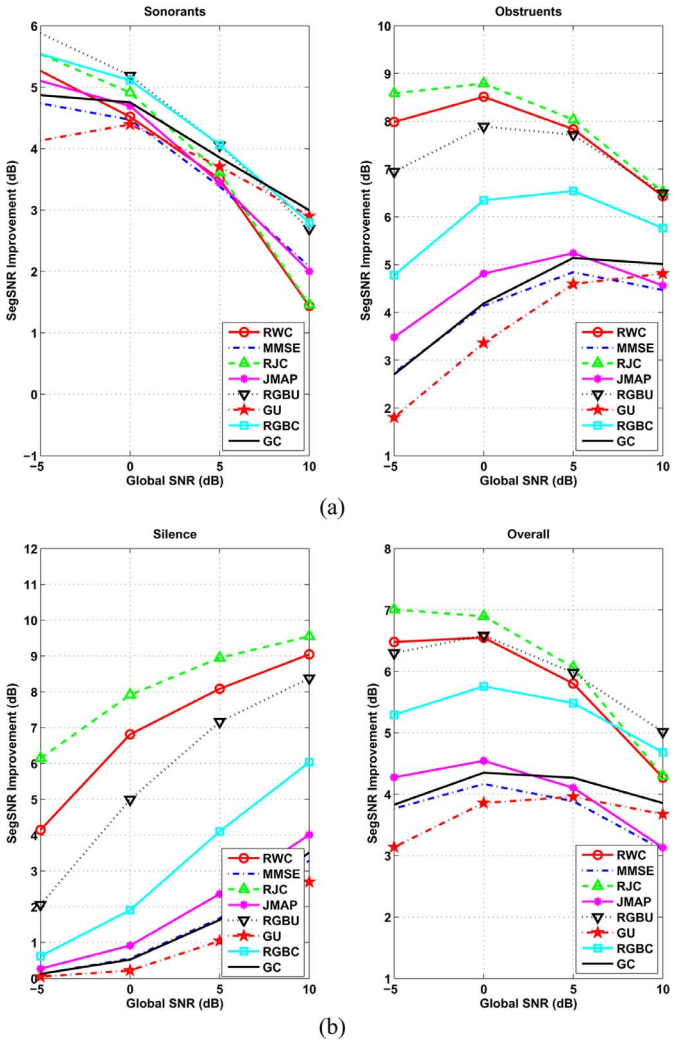


Fig. 3. Segmental SNR improvement of various BPCs in FLN noise for ROVER versus baseline enhancement algorithms.

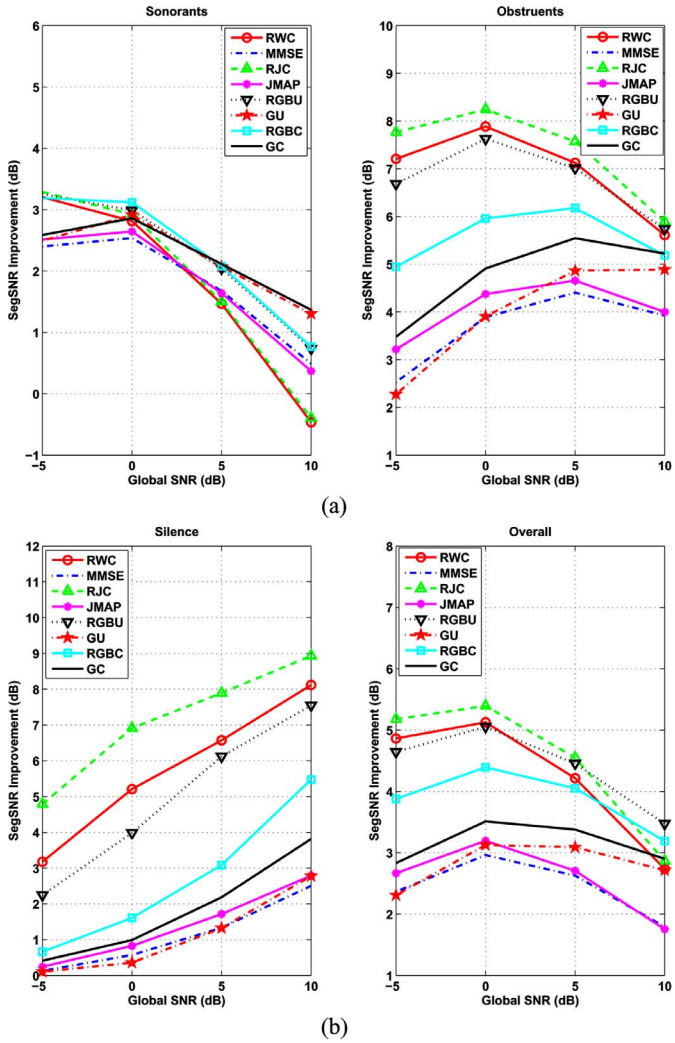


Fig. 4. Segmental SNR improvement of various BPCs in LCR noise for ROVER versus baseline enhancement algorithms.

segmental SNR improvements of ROVER versions over their matching baselines. In the subplot for sonorants, all ROVER based algorithms perform better than their corresponding baseline algorithms at global SNRs of −5, 0, and 5 dB with the largest improvements observed for RGBU over GU. However, at global SNR of 10 dB, the baselines performed better than the corresponding ROVERs. The minimum loss in segmental SNR of the ROVER algorithm over its corresponding baseline was 0.21 dB for RGBU and the maximum was 0.64 dB for RWC. Comparing across all the ROVER algorithms at an SNR of 10 dB, RGBC had the greatest improvement in segmental SNR. In the subplots for obstruents and silence, the ROVER algorithms consistently outperform their baseline counterparts across all SNRs with RJC demonstrating the greatest increase in segmental SNR. In the overall case, RJC at SNRs of −5, 0, and 5 dB and RGBU at an SNR of 10 dB are the best performers. Comparing ROVER versions versus corresponding baselines, it was observed that RGBU had the highest rise in segmental SNR over GU, followed by RWC over MMSE, RJC over JMAP, and RGBC over GC.

In Fig. 4(a) and (b), segmental SNR improvement is evaluated for speech degraded by LCR noise. For the case of sonorants, at global SNR of −5 dB, RJC had the greatest segmental SNR increase and marginally (approximately 0.1 dB) higher than RGBC. At 0-dB SNR, RGBC had the highest segmental SNR. At higher SNRs, GC and GU exhibited the best performances. Therefore, similar to FLN for sonorants, at higher SNRs, all baselines outperformed their ROVER counterparts. The maximum degradation in segmental SNR in ROVER algorithms was no more than 0.95 dB compared to their baselines at high global SNRs of 5–10 dB. In the subplot for obstruents and silence, all ROVER algorithms performed significantly better than their baselines with RJC exhibiting the greatest improvements. In the overall case, the behavior was identical to those in the FLN case (i.e., RJC at SNRs of −5 dB to +5 dB, and RGBU at SNR of 10 dB were the best performers). Also, comparing the ROVERs over their baselines for the overall utterance, RJC had the greatest improvement in segmental SNR over its baseline JMAP followed by RWC, RGBU, and RGBC.

In Fig. 5(a) and (b), the IS improvement is evaluated for speech degraded by FLN noise. In the subplot for sonorants, RGBU demonstrated the highest reduction in IS distortion across all global SNRs. In addition, all ROVER algorithms

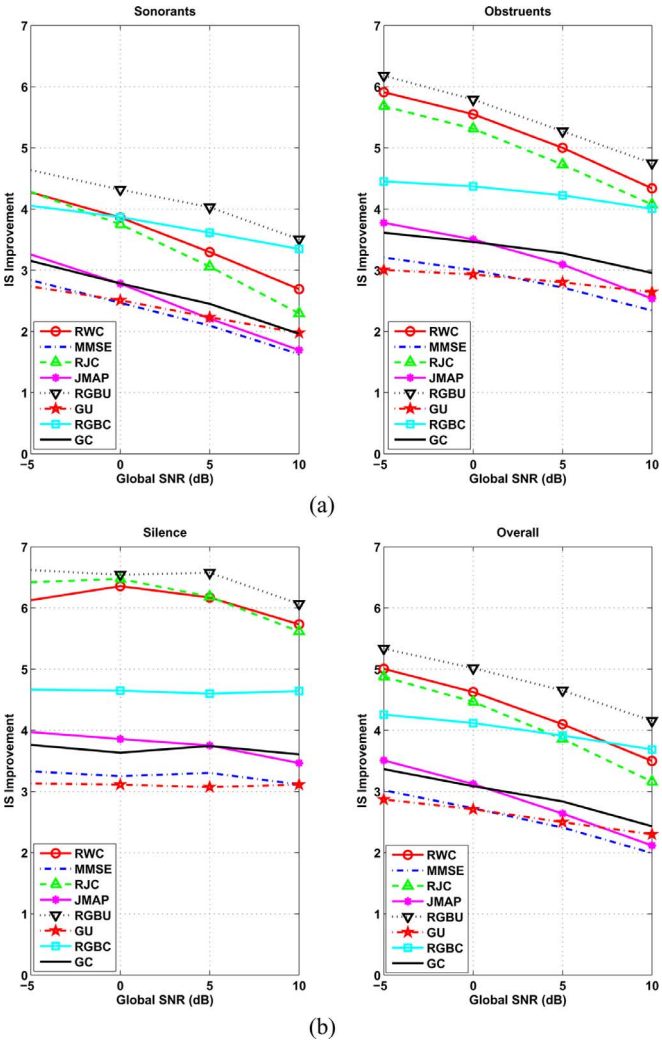


Fig. 5. Itakura–Saito improvement of various BPCs in FLN noise for ROVER versus baseline enhancement algorithms.

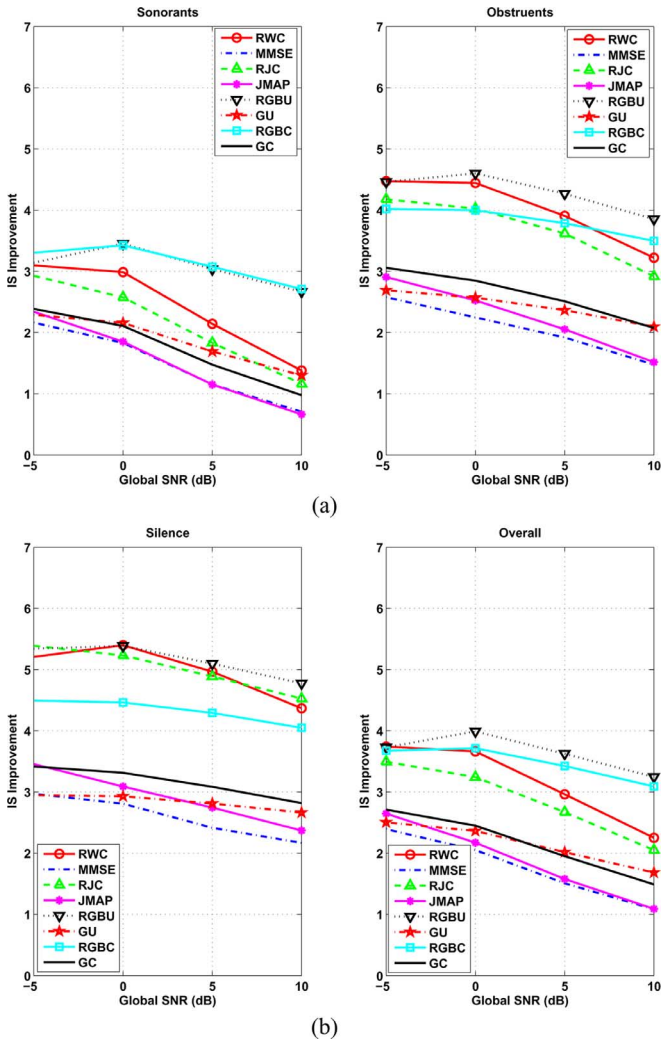


Fig. 6. Itakura–Saito improvement of various BPCs in LCR noise for ROVER versus baseline enhancement algorithms.

outperformed their baselines. Similar to the results observed in segmental SNR evaluations, RGBU/RGBC performed better than RWC/RJC at higher global SNRs in sonorants. For obstruents, silence and overall cases, RGBU continued to have the highest reduction in IS distortion. Again, comparing the ROVERs over their baselines for the overall utterance, RGBU demonstrated the highest improvement in IS scores over its baseline GU followed by RWC, RJC, and RGBC.

In Fig. 6(a) and (b), IS improvement is evaluated for speech degraded by LCR noise. In the subplot for sonorants, RGBC demonstrated the highest reduction in IS distortion for all global SNR whereas it was RGBU for obstruents and silence. In the overall case, RGBU was the best performer. Comparing the ROVERs over their baselines for the overall utterance, RGBU demonstrated the highest improvement in IS scores over its baseline GU, followed by RWC or RGBC, and finally RJC. Ranks of RWC and RGBC were not consistent across SNRs. The margin of improvement of RWC over MMSE was higher at SNRs of -5 dB to 5 dB, whereas at 10 dB RGBC had a higher margin.

To summarize the results, the best and second best enhancement algorithms for each BPC degraded by FLN and LCR noises are outlined in Tables IV(a) and (b), respectively.

C. DFT Distortion Evaluation of the WED Gamma Estimator

The performance of the WED solution (34) of the two-sided Gamma prior is explored in this section. Since (34) is optimized based on the DFT coefficients, we evaluate its performance using D_{DFT} as the root mean square of the DFT distortion [[32], (27)] normalized to 100. For a given frame, this can be represented as

$$D_{\mathrm{DFT}} = \frac{1}{100}\sqrt{\frac{\sum_{k=1}^{K}\left|X(\omega_k) - \hat{X}(\omega_k)\right|^2}{K}} \tag{56}$$

where k represents the frequency bin and K the length of the DFT of the frame. A lower D_{DFT} indicates better enhancement. The plot in Fig. 7 depicts the DFT distortion performance of the overall utterance using the four best values of (α, β) in (34). In both noise types, the estimator with $\alpha = 1, \beta = 0$ was the best performer. This holds true for sonorants, obstruents, and silence.

TABLE IV
TOP TWO ALGORITHMS ACROSS BPCS CORRUPTED BY FLN/LCR NOISE AT GLOBAL SNRS OF −5, 0, 5, 10 dB. BEST INDICATED BY 1ST IN FIRST ROW, SECOND BEST INDICATED BY 2ND IN SECOND ROW. (a) FLN NOISE, (b) LCR NOISE

(a)

BPC	SegSNR				IS Distortion			
	-5 dB	0 dB	5 dB	10 dB	-5 dB	0 dB	5 dB	10 dB
Son 1st	RGBU	RGBU	RGBU	GC	RGBU	RGBU	RGBU	RGBU
Son 2nd	RGBC	RGBC	RGBC	GU	RJC	RGBC	RGBC	RGBC
Obs 1st	RJC	RJC	RJC	RJC	RGBU	RGBU	RGBU	RGBU
Obs 2nd	RWC	RWC	RWC	RWC	RWC	RWC	RWC	RWC
Sil 1st	RJC	RJC	RJC	RJC	RGBU	RGBU	RGBU	RGBU
Sil 2nd	RWC	RWC	RWC	RWC	RJC	RJC	RJC	RWC
Ovl 1st	RJC	RJC	RJC	RGBU	RGBU	RGBU	RGBU	RGBU
Ovl 2nd	RWC	RWC	RGBU	RGBC	RWC	RWC	RWC	RGBC

(b)

BPC	SegSNR				IS Distortion			
	-5 dB	0 dB	5 dB	10 dB	-5 dB	0 dB	5 dB	10 dB
Son 1st	RJC	RGBC	GC	GC	RGBC	RGBC	RGBC	RGBC
Son 2nd	RGBC	RGBU	GU	GU	RGBU	RGBU	RGBU	RGBU
Obs 1st	RJC	RJC	RJC	RJC	RGBU	RGBU	RGBU	RGBU
Obs 2nd	RWC	RWC	RWC	RGBU	RWC	RWC	RWC	RGBC
Sil 1st	RJC	RJC	RJC	RJC	RGBU	RGBU	RGBU	RGBU
Sil 2nd	RWC	RWC	RWC	RWC	RJC	RWC	RWC	RJC
Ovl 1st	RJC	RJC	RJC	RGBU	RGBU	RGBU	RGBU	RGBU
Ovl 2nd	RWC	RWC	RGBU	RGBC	RWC	RGBC	RGBC	RGBC

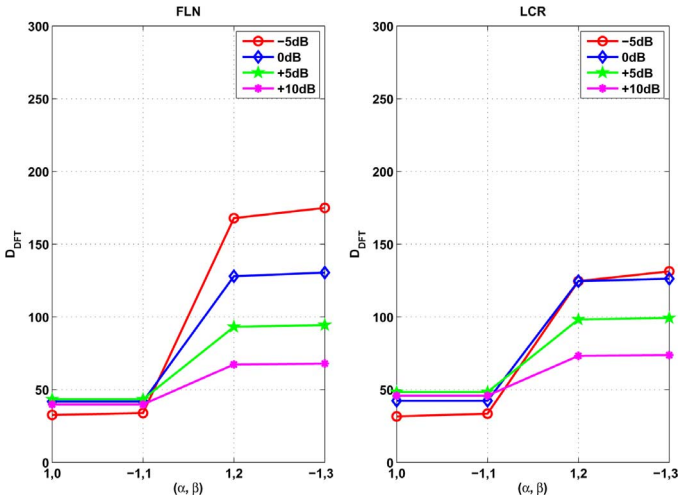


Fig. 7. DFT distortion performance of the overall utterance enhanced by DFT WED Gamma estimator in FLN and LCR noises for different values of (α, β) in the x-axis.

With $\alpha = 1, \beta = 0$, the WED solution (34) collapses to the special case DFT MMSE estimator [21]. Intuitively, we expected that lowering α or β would achieve better performance than the special case estimator. However, such an estimator could not be identified from our experiments. We evaluated performance from other combinations of α, β, but their performance was worse than those shown in the plot. Since $\alpha = 1, \beta = 0$ is the best estimator for all BPCs over all noise types considered, it removes the need to improve it further using the ROVER algorithm.

D. Listener Evaluations

A perceptual evaluation of the proposed enhancement algorithms was performed using listener tests. A group of ten listeners with normal hearing capabilities were asked to indicate their preference in the paired-comparison (AB preference type) tests. For each listener, a set of four sessions were conducted with a total of eight paired-comparison tests per session. For example, session 1 comprised of eight paired-comparison tests using RWC versus MMSE. Hence, in each paired-comparison test, enhanced speech obtained from the proposed RWC algorithm was compared with the enhanced speech obtained from the baseline MMSE algorithm. Similarly, other sessions were grouped according to the enhancement algorithms as RJC versus JMAP, RGBU versus GU, and RGBC versus GC. The utterances were degraded using the FLN and LCR noises. The order of the algorithms presented in the paired-comparison tests was randomized to eliminate any biasing towards a particular algorithm. Overall, in 75% of the cases, the listeners preferred the proposed ROVER based algorithms. The individual breakdown per algorithm was 78% (RWC), 74% (RJC), 73% (RGBU), and 73% (RGBC). The breakdown based on noise type was 71% (FLN) and 78% (LCR). These results indicate that the listeners demonstrated a stronger preference for the ROVER based algorithms over the baselines across both the noise types.

V. CONCLUSION

A ROVER-based speech enhancement algorithm was proposed in this study to achieve improved enhancement at the phoneme class level. This was accomplished in two stages. In stage one, short-time spectral magnitude generalized spectral subtraction β-unconstrained and β-constrained parametric estimators were derived using coefficients that minimize the weighted Euclidean distortion between the clean and estimated speech spectral magnitudes. This idea was extended to the minimum mean square error and joint maximum *a posteriori* algorithms. Using super-Gaussian priors (Chi and two-sided Gamma) to model the clean speech spectral magnitudes or discrete Fourier transform coefficients, a class of Bayesian spectral magnitude estimators were derived using the weighted Euclidean distortion as an overall cost function. Furthermore, the joint maximum *a posteriori* estimator with Chi distributed clean speech spectral magnitude and uniform phase was also proposed. The behavior of all five estimators were investigated using a wide range of constrained configurable parameters over two noise types and four SNR levels.

In the second stage, three prior enhanced utterances from these parametric estimators, each customized for a specific phoneme class, were generated. Using the mixture maximum model, phoneme classification was performed using probabilistic decisions and these decisions were used as weights to combine the phoneme segments from the prior enhanced utterances. This resulted in a composite utterance that produces better levels of speech quality in all adverse conditions. This

also confirms the versatility of the ROVER paradigm across different enhancement algorithms.

ACKNOWLEDGMENT

The authors would like to thank P. Ankititrakul and V. Prakash of CRSS at the University of Texas at Dallas for sharing their feature extraction and model building routines. The authors would also like to thank the anonymous reviewers for their helpful and constructive comments which significantly improved the quality of the manuscript.

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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. Signal Process., Spec. Iss. Signal Process. for Hearing Aids and Cochlear Implants*, pp. 1425–1428, Oct. 2005.

[6] S. Nandkumar 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.

[7] J. H. L. Hansen and S. Nandkumar, "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.

[8] J. H. L. Hansen, V. Radhakrishnan, and K. Arehart, "Speech enhancement based on generalized minimum mean square error estimators and masking properties of the auditory system," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 2049–2063, Nov. 2006.

[9] 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.

[10] 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.

[11] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 443–445, Dec. 1984.

[12] 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. 857–869, Sep. 2005.

[13] 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 Recognition Understand.*, 2003, pp. 321–326.

[14] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, Apr. 1979.

[15] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 1979, pp. 208–211.

[16] S. Kamath and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2002, p. 4164.

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

[18] J. Deller, J. H. L. Hansen, and J. Proakis, *Discrete Time Processing of Speech Signals*. New York: Prentice-Hall, 2000, ISBN 0-7803-5386-2.

[19] J. H. L. Hansen and L. M. 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.

[20] B. L. Sim, Y. C. Tong, J. S. Chang, and C. T. Tan, "A parametric formulation of the generalized spectral subtraction method," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 4, pp. 328–337, Jul. 1998.

[21] R. Martin, "Speech enhancement using MMSE short time spectral estimation with gamma distributed speech priors," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2002, vol. 1, pp. 253–256.

[22] P. J. Wolfe and S. J. Godsill, "Efficient alternatives to the Ephraim and Malah suppression rule for audio signal enhancement," *EURASIP J. App. Signal Process.*, vol. 10.

[23] A. Das and J. H. L. Hansen, "Broad phoneme class based speech enhancement using mixture maximum model," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2010, pp. 4762–4765.

[24] A. Das and J. H. Hansen, "Phoneme selective speech enhancement using the generalized parametric spectral subtraction estimator," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2011, pp. 4648–4651.

[25] 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 Understand.*, pp. 347–354.

[26] A. Nádas, D. Nahamoo, and M. A. Picheny, "Speech recognition using noise-adaptive prototype," *IEEE Trans. Speech Audio Process.*, vol. 37, no. 10, pp. 1495–1505, Oct. 1989.

[27] I. Andrianakis and P. R. White, "MMSE speech spectral amlitude estimators with Chi and Gamma speech priors," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2006, vol. 3, pp. 1068–1071.

[28] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 4th ed. New York: Academic, 1980.

[29] D. Burshtein and S. Gannot, "Speech enhancement using a mixture-maximum model," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 6, pp. 341–351, Sep.. 2002.

[30] D. P. W. Ellis, PLP and RASTA (and MFCC, and Inversion) in Matlab, 2005 [Online]. Available: http://www.ee.columbia.edu/dpwe/resources/matlab/rastamat.

[31] P. C. Loizou, *Speech Enhancement Theory and Practice*. Boca Raton, FL: CRC, 2007.

[32] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete Fourier coefficients with generalized Gamma priors," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 6, pp. 1741–1752, Aug. 2007.

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

He is a currently 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, 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, respectively.

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 59 (27 Ph.D., 32 M.S./M.A.) thesis candidates. He is author/coauthor of 427 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 has served as Technical Advisor to a U.S. Delegate for NATO (IST/TG-01), IEEE Signal Processing Society Distinguished Lecturer (2005/06), 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.