Perceptual Postfilter Estimation for Low Bit Rate Speech Coders Using Gaussian Mixture Models

Wei Chen *Peter Kabal* *Turaj Z. Shabestary*

Department of Electrical and Computer Engineering
McGill University, Montreal, Quebec H3A 2A7

Abstract

A novel perceptual postfilter is introduced. For each frame, the filter gains, $\mathbf{z}$, are estimated given a vector, $\mathbf{y}$, of the quantized LSFs and the long-term prediction gain of the corresponding frame. The proposed perceptual postfilter is derived from an optimal MMSE estimator, i.e. the estimated gain vector is $\hat{\mathbf{z}} = E\{\mathbf{z}|\mathbf{y}\}$. The MMSE estimator is based on the conditional pdf of $\mathbf{z}$ given $\mathbf{y}$, which is computed from the joint pdf modelled by a GMM. The proposed perceptual postfilter improves the speech naturalness comparing with the conventional adaptive postfilter, while maintaining the property of being an "add-on" postfilter without modification to the current encoder.

1 Introduction

Adaptive postfilters [1] have been widely applied in current Linear Prediction Analysis-by-Synthesis (LPAS) speech coders. Conventional postfiltering improves the decoded speech quality using the information available at the decoder, and is empirically designed based on aspects of human perception. As research furthers in modelling of the human auditory system, better psychoacoustic models [2, 3] have been proposed and applied in speech and audio processing, especially in audio coding. However, only a few improvements (for instance, [4, 5]) have been made to adaptive postfilters despite our better understanding of the human auditory system.

A speech codec usually operates on a frame-by-frame basis. When we have access to the clean speech and its decoded version from a speech codec, a perceptual postfilter can be constructed based on perceptual properties. The perceptual filter gains can be derived from each processing frame and applied to the decoded speech to improve the speech quality. However, in practice we do not have the information about the perceptual postfilter gains at the decoder if they are not sent as side information. In this paper, we focus on the estimation of the perceptual postfilter gains without additional side information.

Assume a given speech frame is coded by a LPAS speech coder, the decoder retrieves the quantized linear prediction (LP) coefficients. The LP coefficients represent the envelope of the short-time power spectrum which is very important for both the quality and intelligibility of coded speech. The perceptual postfilter gains are calculated for the corresponding frame. Since the open-loop prediction gain of the long-term prediction (LTP) in speech signals indicates the degree of voicing of the speech, we also calculate the LTP gain of this frame. We take the LP coefficients and the LTP gain as an "input" vector, and the perceptual postfilter gains as a "target" vector. A feature vector is constructed from "input" and "target" vectors. In order to find a Minimum Mean Square Error (MMSE) estimate of the "target" vector, *a priori* information of the joint probability density function (pdf) of the feature vector is required. A Gaussian Mixture Model (GMM) is used to model the joint density.

We discuss the perceptual postfilter in Section 2. In Section 3 we present the MMSE estimator and the GMM. Simulation results are given in Section 4.

2 Perceptual Postfilter

LPAS speech coding models the human speech production by exciting a time-varying LP all-pole filter by an excitation signal [6]. The coder attempts to minimize a perceptually weighted error signal (exploiting masking properties). Both the quantized information about the excitation and the LP coefficients are transmitted to the receiver. The LP coefficients are usually converted to Line Spectral Frequencies (LSFs) before quantization and transmission.

Masking is an important phenomenon of our auditory system. It means a sound is inaudible in the presence of a stronger sound. A masking threshold from psychoacoustics measures the amount of the allowable distortion. For speech, noise in spectral valleys is more sensitive than that in spectral peaks according to empirical results [1]. Perceptual weighting tries to shape the spectrum of the coding noise following the speech spectrum to some extent to suppress noise in spectral valleys. However, perceptual weighting alone is not enough to make the coding noise inaudible at low encoding rates. In order to make better use of masking, an adaptive postfilter [1] is commonly used to lessen the coding noise in the spectral valleys. The conventional adaptive postfilter is the combination of a pitch postfilter and a formant postfilter to reduce the coding noise in the spectral valley regions.[1] From the decoded information, an adaptive postfilter can be easily built and can act as an independent add-on component to the system. The conventional adaptive postfilter is based on empirical results for low bit rate coders [1]: a) The masking threshold follows to some extent the spectral peaks and valleys of speech spectrum; b) the noise shaping by perceptual weighting filter at encoder makes the coding noise fall below the masking threshold around the spectral peaks but appear above the masking threshold in the valleys.

Applications of masking models usually involve calculating a masking threshold and arranging the coding noise to below the masking threshold. The frequency resolution of our human ear is represented by *critical bands*, which are nonlinearly spaced on the frequency scale (Hz). One Bark spans the width of a critical band. The excitation pattern, which is obtained by frequency and time domain spreading, predicts the physical activity of hair cells along the basilar membrane in the ear. A masking threshold is derived by weighting the excitation intensity. Different

[1] Weighting at the encoder (where the clean speech signal is available) shapes the coding noise. Postfiltering at the decoder affects both the speech and the coding noise.

psychoacoustic models may handle the orders and the function expressions of time and frequency domain spreading differently [2, 3].

In psychoacoustic modelling, a neural excitation called loudness is assumed to directly affect perceived strength. A loudness distribution is predicted from the excitation intensity by a nonlinear transformation. Recent research has considered the loudness model instead of the masking model [7], however, both loudness and masking are directly connected with excitation by operations independent of the signal level. The excitation pattern model is also considered as a loudness representation. Lam and Stewart [8] exploited the excitation pattern model to derive a generalized perceptual audio filter in low rate audio coding. Their filter reduces the audible coding noise by trying to "equalize" the excitation pattern representation of the original signal and the coded signal. Let us denote the original signal as $s(n,i)$, the coded signal as $y(n,i)$, and the perceptually filtered $y(n,i)$ as $z(n,i)$. Their corresponding spectral components are $S_p(k,i)$, $Y_p(k,i)$ and $Z_p(k,i)$. The excitation pattern values are $S_E(z,i)$, $Y_E(z,i)$ and $Z_E(z,i)$, respectively. Here, n is the time domain index, i is the frame index of the signal, k is the frequency domain index, and z is the critical band index. By restoring the excitation patterns, they suppressed the audible quantization noise in low bit rate wideband audio coding [8]

$$Z_E(z,i) = S_E(z,i), \quad 1 \le z \le B \tag{1}$$

where B is the total number of critical bands in the perceptual domain. The gain of the perceptual filter is assumed to be constant within the same critical band, and denoted by $H(z,i)$. The filtered signal in each critical band is given by

$$Z_p(k,i) = H(z,i)Y_p(k,i), \quad 1 \le z \le B, k \in [b_l(z), b_h(z)] \tag{2}$$

where $b_l(z)$ is the lower boundary of critical band z, and $b_h(z)$ is the upper boundary of critical band z. The perceptual filter gains are derived from Eq. (1) with a simplified psychoacoustic model. In the implementation, these filter gains were sent as side information to enhance the decoded signal.

This perceptual filter exploits the properties of the psychoacoustic model which it is based on, and can be directly applied to the frequency domain of the decoded signal to suppress the perceptible noise. It gives us a new perspective on adaptive postfiltering. We build a perceptual postfilter for speech coders based on the ideas of Lam and Stewart [8], but with the estimation done entirely at the decoder.

The psychoacoustic model used by Lam and Stewart [8] is an invertible auditory model (ignoring level dependant effects on the spreading functions and time-domain spreading). In this case, the operation from the critical band intensity to the excitation intensity is unnecessary. The gains of the perceptual filter in [8] can be found directly as the ratio of the critical band intensity of the original signal to that of the coded signal. If the critical band intensity of the coded signal is adjusted to be equal to the same level of the original signal, the filtered signal will have the same loudness representation as that of the original signal. This motivates us to build our perceptual postfilter by equalizing the energy in perceptual domain

$$Z_B(z,i) = S_B(z,i), \quad 1 \le z \le B \tag{3}$$

where $S_B(z,i)$ and $Z_B(z,i)$ are the critical band intensity of the original signal and the perceptually filtered signal, respectively. By using the grouping as in Johnston's model [2], the energy in each critical band of $s(n,i)$ is summed to give the critical band spectrum $S_B(z,i)$:

$$S_B(z,i) = \sum_{k=b_l(z)}^{b_h(z)} S_p(k,i), \quad 1 \le z \le B \tag{4}$$

Apply the grouping to Eq. (2) and combining with Eq. (3), our new perceptual postfilter has the expression

$$H(z,i) = S_B(z,i)/Y_B(z,i), \quad 1 \le z \le B \tag{5}$$

where $Y_B(z,i)$ is the critical band spectrum of the coded signal.

3 Estimation of The Perceptual Filter Gains

3.1 MMSE Estimation

We want a perceptual postfilter to act as a "true" add-on component at the receiver without increasing the bit rate of the original speech coder. We want to make use the information available at the decoder to derive the perceptual postfilter. An MMSE estimator can be constructed to estimate the perceptual postfilter gains. Assume a d-dimensional feature vector $\mathbf{s}$ is composed of a k-dimensional "input" subvector, $\mathbf{y}$, of some information at the decoder and an l-dimensional "target" subvector, $\mathbf{z}$, of the perceptual postfilter gains, i.e. $\mathbf{s} = [\mathbf{y}; \mathbf{z}]$. The MMSE estimator gives the estimate of $\mathbf{z}$ by the conditional expectation

$$\hat{\mathbf{z}} = E\{\mathbf{z}|\mathbf{y}\} \tag{6}$$

We use the LTP gain from the decoded speech and the quantized LSFs as the subvector $\mathbf{y}$ in the feature vector. Knowing the joint pdf of the feature vector is essential in MMSE estimation (6), we use a mixture of Gaussian components to model the pdf.

3.2 Gaussian Mixture Models

A GMM is commonly used to approximate a pdf with relatively small number of parameters. Its ability to represent general speech parameters (spectral shapes) by sums of Gaussian components makes it popular in speech recognition and speaker identification [9]. GMMs are also used for vector quantization of LSFs [10]. Qian and Kabal [11] used GMMs for bandwidth extension by estimating the missing high band information from the low band LSFs.

We approximate the joint pdf of the feature vector $\mathbf{s}$ by a GMM with M Gaussian components [10]

$$p_{\mathbf{s}|\boldsymbol{\Theta}}(\mathbf{s}|\boldsymbol{\Theta}) = \sum_{i=1}^{M} \alpha_i \, b(\mathbf{s}|\boldsymbol{\theta}_i) \tag{7}$$

$$\boldsymbol{\Theta} = \{\alpha_1, \cdots, \alpha_M, \boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_M\}. \tag{8}$$

Here $b(\mathbf{s}|\boldsymbol{\theta}_i)$ is a multivariate Gaussian density parameterized by $\boldsymbol{\theta}_i = \{\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}$ with mean vectors $\boldsymbol{\mu}_i$ and covariance matrices $\boldsymbol{\Sigma}_i$. The value α_i denotes the *a priori* probability of the ith mixture component $b(\mathbf{s}|\boldsymbol{\theta}_i)$ with $\sum_{i=1}^{M} \alpha_i = 1$.

The parameter set $\boldsymbol{\Theta}$ can be estimated by the maximum likelihood (ML) method. Expectation-Maximization (EM) algorithm is a widely used approach for ML estimation in cases where a closed-form analytical expression for the optimal parameters is hard to derive. EM is an iterative algorithm where in each iteration over a given database a monotonic increase in the log-likelihood, L, is guaranteed [10], i.e., $L(\boldsymbol{\Theta}^{(k+1)}) \ge L(\boldsymbol{\Theta}^{(k)})$, where $\boldsymbol{\Theta}^{(k)}$ is the value of the parameter set $\boldsymbol{\Theta}$ at iteration k.

We choose diagonal covariance matrices in GMM rather than full covariance matrices to reduce the parameters to be estimated. The number of parameters to be estimated during training is $M(d+d(d+1)/2+1)$ for full covariance Gaussians, and $M(2d+1)$ for diagonal covariance Gaussians. The larger the number of parameters, the greater the possibility to describe the fine structure of the underlying data distribution. On the other hand, with a high degree of freedom in the modelling, there is an risk of overfit.

3.3 MMSE Estimation of The Perceptual Postfilter Using GMMs

While using the MMSE estimator of Eq. (6) with a GMM pdf, we need the conditional pdf of "target" postfilter gain vector $\mathbf{z}$ given the "input" vector $\mathbf{y}$. The individual Gaussian density parameter $\boldsymbol{\theta}_i$ can be written as follows

$$\boldsymbol{\mu}_i = \begin{bmatrix}\boldsymbol{\mu}_{\mathbf{y}i}\\ \boldsymbol{\mu}_{\mathbf{z}i}\end{bmatrix} \tag{9}$$

$$\boldsymbol{\Sigma}_i = \begin{bmatrix}\boldsymbol{\Sigma}_{\mathbf{yy}i} & \boldsymbol{\Sigma}_{\mathbf{yz}i}\\ \boldsymbol{\Sigma}_{\mathbf{zy}i} & \boldsymbol{\Sigma}_{\mathbf{zz}i}\end{bmatrix} \tag{10}$$

The conditional pdf and any marginal pdf of jointly Gaussian random variables are Gaussian densities [12]. The joint Gaussian pdf components $b(\mathbf{s}|\boldsymbol{\theta}_i) = b(\mathbf{s}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ in Eq. (7) can be factored into a conditional Gaussian pdf $b(\mathbf{z}|\boldsymbol{\mu}_{\mathbf{z}|\mathbf{y}_i}, \boldsymbol{\Sigma}_{\mathbf{z}|\mathbf{y}_i})$ and a marginal Gaussian pdf $b(\mathbf{y}|\boldsymbol{\mu}_{\mathbf{y}_i}, \boldsymbol{\Sigma}_{\mathbf{yy}i})$

$$b(\mathbf{s}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = b(\mathbf{z}|\boldsymbol{\mu}_{\mathbf{z}|\mathbf{y}_i}, \boldsymbol{\Sigma}_{\mathbf{z}|\mathbf{y}_i})\, b(\mathbf{y}|\boldsymbol{\mu}_{\mathbf{y}_i}, \boldsymbol{\Sigma}_{\mathbf{yy}i}) \tag{11}$$

Therefore, the conditional pdf of $\mathbf{y}$ and $\mathbf{z}$ is expressed in terms of a GMM

$$p(\mathbf{z}|\mathbf{y}) = \frac{\sum_{i=1}^{M} \alpha_i\, b(\mathbf{s}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{k=1}^{M} \alpha_k\, b(\mathbf{y}|\boldsymbol{\mu}_{\mathbf{y}_k}, \boldsymbol{\Sigma}_{\mathbf{yy}k})}$$

$$= \frac{\sum_{i=1}^{M} \alpha_i\, b(\mathbf{z}|\boldsymbol{\mu}_{\mathbf{z}|\mathbf{y}_i}, \boldsymbol{\Sigma}_{\mathbf{z}|\mathbf{y}_i})\, b(\mathbf{y}|\boldsymbol{\mu}_{\mathbf{y}_i}, \boldsymbol{\Sigma}_{\mathbf{yy}i})}{\sum_{k=1}^{M} \alpha_k\, b(\mathbf{y}|\boldsymbol{\mu}_{\mathbf{y}_k}, \boldsymbol{\Sigma}_{\mathbf{yy}k})} \tag{12a}$$

$$= \sum_{i=1}^{M} h_i(\mathbf{y})\, b(\mathbf{z}|\boldsymbol{\mu}_{\mathbf{z}|\mathbf{y}_i}, \boldsymbol{\Sigma}_{\mathbf{z}|\mathbf{y}_i}). \tag{12b}$$

From [12]

$$\boldsymbol{\mu}_{\mathbf{z}|\mathbf{y}_i} = \boldsymbol{\mu}_{\mathbf{z}i} + \boldsymbol{\Sigma}_{\mathbf{yz}i}(\boldsymbol{\Sigma}_{\mathbf{yy}i})^{-1}(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{y}_i}) \tag{13}$$

$$\boldsymbol{\Sigma}_{\mathbf{z}|\mathbf{y}_i} = \boldsymbol{\Sigma}_{\mathbf{zz}i} - \boldsymbol{\Sigma}_{\mathbf{yz}i}(\boldsymbol{\Sigma}_{\mathbf{yy}i})^{-1}\boldsymbol{\Sigma}_{\mathbf{zy}i} \tag{14}$$

and

$$h_i(\mathbf{y}) = \frac{\alpha_i\, b(\mathbf{y}|\boldsymbol{\mu}_{\mathbf{y}_i}, \boldsymbol{\Sigma}_{\mathbf{yy}i})}{\sum_{k=1}^{M} \alpha_k\, b(\mathbf{y}|\boldsymbol{\mu}_{\mathbf{y}_k}, \boldsymbol{\Sigma}_{\mathbf{yy}k})} \tag{15}$$

A MMSE estimate of $\mathbf{z}$ is derived with Eqs. (6), (12b) and (13)

$$\hat{\mathbf{z}} = \sum_{i=1}^{M} h_i(\mathbf{y})\boldsymbol{\mu}_{\mathbf{z}|\mathbf{y}_i} \tag{16}$$

When diagonal covariance matrices are used for $\boldsymbol{\Sigma}_i$, the MMSE estimator is reduced to

$$\hat{\mathbf{z}} = \sum_{i=1}^{M} h_i(\mathbf{y})\boldsymbol{\mu}_{\mathbf{z}i} \tag{17}$$

4 Experimental Results

In the experiments, all speech is sampled at 8 kHz with 16-bit PCM resolution. The database is composed of speech of 23 speakers (12 females and 11 males). One section of the database with 10 female and 9 male speech is used for GMM training, and the other 2 females and 2 males are used for performance evaluation. We do the experiments with the ITU-T G.723.1 speech codec [6] at rate of 5.3 kbps. The ITU-T G.723.1 speech codec operates on frames of 240 samples. Each frame is divided into four subframes of 60 samples each. For each subframe, 10th order LP analysis is used on a Hamming windowed 180 samples centered on the subframe. The LP coefficients for the last subframe are converted to LSFs and quantized. The excitation signal is coded with pitch period and algebraic-code-excitation (ACELP) for each subframe.

For training the GMM, the ITU-T G.723.1 speech coder encodes the corresponding information about excitation and LSFs, as shown in the top part of Fig. 1. A feature vector of dimension 28 for GMM training is composed of a 10 quantized LSFs, a LTP prediction gain and 17 bark-scale perceptual postfilter gains. In each frame, the block of 180 samples centered on the last subframe is used to generate a training vector. For each decoded block, the LSFs are retrieved and an LTP gain is calculated. To get the perceptual postfilter gains, a sine-squared window is applied to the first 60 and the last 60 samples of the processing blocks of the original and decoded speech as shown in Fig. 2. An FFT of length 256 are used on each windowed block. GMM with diagonal covariance matrices is trained at the encoder as indicated in Fig. 1. The training set consists of 348,955 vectors.

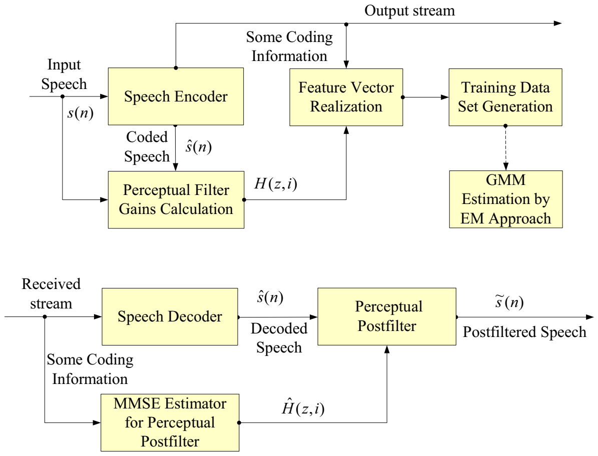

Fig. 1 System Diagrams. Top: GMM training at the encoder; Bottom: Perceptual postfiltering by MMSE estimation at the decoder.

Our proposed perceptual postfilter works at the decoder end, as shown in the bottom part of Fig. 1. The postfiltering is performed on windowed blocks of 180 samples, with 60 sample overlaps, see Fig. 2. The same window and length of FFT are used as in GMM training. For each frame of the decoded speech, a MMSE estimate gives the postfilter gains for the block center on the last subframe by the decoded LSFs and the calculated LTP gain, while the postfilter gains from the block centered on the second subframe are given by the mean of the estimated postfilter gains for the previous and the next blocks. To check the correlation of "input" and "target" vectors, Table 1

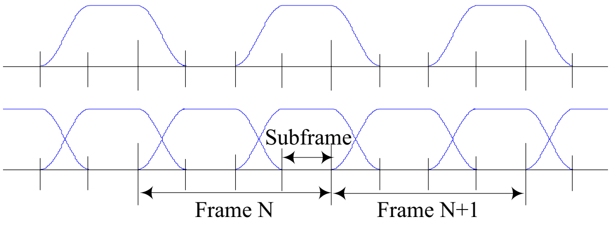

Fig. 2 Windowing for the training (upper plot) and perceptual postfiltering (lower plot).

gives the mutual information from GMM pdf with M=16, 32, 64, and 128 mixture components. It shows that the "input" and "target" are correlated, and more mixtures should give better estimates. We use M=128 for our experiments.

Table 1 Mutual information.

Gaussian Mixtures	Mutual Information
16	2.53
32	2.88
64	3.12
128	3.33

Fig. 3 shows the spectrograms of clean speech, ITU-T G.723.1 coded speech with standard postfiltering, and ITU-T G.723.1 coded speech with the new perceptual postfilter, respectively. Low bit rate coding emphasizes the high energy parts (generally formants at low frequencies) and loses some naturalness at high frequencies. From Fig. 3, it can be seen that the perceptual postfilter recovers some of this loss. Informal listening test shows the proposed postfilter gives more natural speech than the conventional postfilter, while maintaining intelligibility.

5 Conclusions

A novel perceptual postfilter for low bit-rate LPAS speech coders has been introduced in this paper. The LSFs and LTP gains from the decoder are used to estimate the perceptual postfilter gains by a MMSE estimator using a GMM. The proposed postfilter is perceptually based and is an add-on part at the receiver just as for a conventional adaptive postfilter. Informal listening tests show an improved speech quality with a more natural sound.

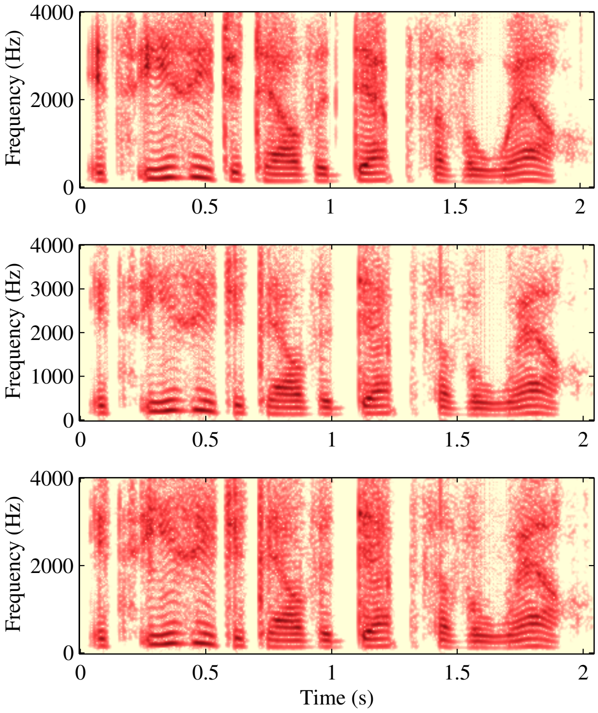

Fig. 3 Spectrograms. Top: Original speech; Middle: ITU-T G.723.1 coded speech with the standard postfiltering; Bottom: ITU-T G.723.1 coded speech with the perceptual postfiltering.

References

[1] J.-H. Chen and A. Gersho, "Adaptive postfiltering for quality enhancement of coded speech," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 59–71, Feb. 1988.

[2] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE J. Selected Areas Commun.*, vol. 6, pp. 314–323, Feb 1988.

[3] ITU-R, *Method for Objective Measurements of Perceived Audio Quality*. ITU-R Recommendation BS.1387, International Telecommunication Union, Dec. 1998.

[4] A. Mustapha and S. Yeldener, "An adaptive post-filtering technique based on a least squares approach," in *Proc. IEEE Speech Coding Workshop*, pp. 156–158, 1999.

[5] W. B. Kleijn, "Enhancement of coded speech by constrained optimization," in *Proc. IEEE Speech Coding Workshop*, pp. 163–165, Oct. 2002.

[6] ITU-T, *Dual Rate Speech Coder for Multimedia Communications Transmitting at 5.3 and 6.3 kbit/s*. ITU-T Recommendation G.723.1, International Telecommunication Union, Mar. 1996.

[7] R. Der, P. Kabal, and W.-Y. Chan, "Towards a new perceptual coding paradigm for audio signals," in *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing*, pp. 457–460, 2003.

[8] Y. Lam and R. Stewart, "Perceptual suppression of quantization noise in low bitrate audio coding," in *Conf. Rec. 31st Asilomar Conf. Signals, Systems, Computers*, pp. 49–53, 1997.

[9] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 72–83, Jan. 1995.

[10] P. Hedelin and J. Skoglund, "Vector quantization based on gaussian mixture models," *IEEE Trans. Speech Audio Processing*, vol. 8, pp. 385–401, Jul. 2000.

[11] Y. Qian and P. Kabal, "Combining equalization and estimation for bandwidth extension of narrowband speech," in *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing*, pp. 713–716, May 2004.

[12] S. Kotz, N. Balakrishnan, and N. Johnson, *Continuous Multivariate Distribution*, vol. 1. John Wiley & Sons, 2000.