

A PERCEPTUALLY MOTIVATED SUBSPACE APPROACH FOR SPEECH ENHANCEMENT

Yi Hu and Philipos C. Loizou

Department of Electrical Engineering
University of Texas at Dallas
Richardson, Texas 75083-0688
{yihuyxy, loizou}@utdallas.edu

ABSTRACT

A perceptually motivated subspace based approach is proposed for enhancement of speech corrupted by colored noise. The proposed approach takes into account the frequency masking properties of the human auditory system and reduces the perceptual effect of the residual noise. Objective measures and informal listening tests demonstrated improvements over other subspace-based methods when tested with TIMIT sentences corrupted with speech-shaped noise.

1. INTRODUCTION

Following the work of Ephraim and Van Trees [1], several subspace-based speech enhancement methods (e.g., [2]-[4]) were proposed. The original paper [1] of Ephraim and Van Trees (EV) dealt primarily with white noise and suggested using an extra prewhitening step to handle colored noise. In [2]-[4], EV's work was extended to colored noise.

In [2], the approach focused on providing proper noise shaping for colored noise without prewhitening. This was done by first classifying the noisy speech frames into speech-dominated and noise-dominated frames and then using a different KLT matrix for those frames to construct the estimator. Rezayee and Gazor [3] extended EV' time domain constrained method to deal with colored noise by approximating the covariance matrix of the KLT-transformed noise vectors with a diagonal matrix. Their approximation, however, led to a sub-optimal estimator. Hu and Loizou [4] derived two generalized optimal linear estimators based on time domain and spectrum domain constraints with built-in prewhitening.

The work in [1], [2] and [4] showed that proper noise shaping is essential to subspace based speech enhancement methods, since noise shaping can exploit to some extent the masking properties of the human auditory system and thereby reduce the perceptual effect of the residual noise. No explicit method was proposed in [1] however to shape

the frequency spectrum of the residual noise. Motivated by the perceptual weighting techniques used in the analysis-by-synthesis speech coders [5], we propose a new subspace based speech enhancement approach, which takes into account the frequency masking properties of the human auditory system.

We also address the issue of obtaining a good estimate of the covariance matrix, an issue which is very important for the performance of subspace-based speech enhancement methods. McWhorter and Scharf's work [6] provided a new viewpoint for the problem of covariance estimation. They showed that the commonly used covariance estimators are special cases of multiwindow estimators and that a special type of window may be used to improve the estimator's performance. Our proposed approach integrated this multiwindow covariance matrix estimator.

This paper is organized as follows. In section 2 the proposed perceptually-motivated subspace speech enhancement approach is presented. In section 3, implementation details are described. Experimental results are presented in section 4, and the conclusions are given in section 5.

2. PERCEPTUALLY MOTIVATED SUBSPACE METHOD FOR SPEECH ENHANCEMENT

In this section, we first introduce the perceptual weighting techniques used extensively in analysis-by-synthesis speech coders, and then we derive the estimator which minimizes the energy of the speech distortion subject to the energy of the perceptually-weighted residual noise falling below a preset threshold level.

2.1. Perceptual weighting

In most low-rate speech coders (e.g., CELP), the excitation used for LPC synthesis is selected in a closed-loop fashion using a perceptually weighted error criterion [5]. This error criterion exploits the masking properties of the auditory system. More specifically, it is based on the fact that the

Research supported in part by Grant No. R01 DC03421 from NINDCD/NIH.

auditory system has a limited ability to detect the quantization noise near the high-energy regions of the spectrum (e.g., near the formant peaks). Quantization noise near the formant peaks is masked by the formant peaks, hence will not be audible. Auditory masking can be exploited by shaping the frequency spectrum of the error so that less emphasis is placed near the formant peaks and more emphasis is placed on the spectral valleys, where any amount of noise present will be audible. The error is shaped using the following perceptual filter:

$$\begin{aligned} P(z) &= \frac{A(z/\gamma_1)}{A(z/\gamma_2)} \\ &= \frac{1 - \sum_{k=1}^p a_k \gamma_1^k z^{-k}}{1 - \sum_{k=1}^p a_k \gamma_2^k z^{-k}} \end{aligned} \quad (1)$$

where $A(z)$ is the LPC polynomial, a_k are the short-term linear prediction coefficients, γ_1 and γ_2 ($0 \leq \gamma_2 \leq \gamma_1 \leq 1$) are parameters that control the energy of the error in the formant regions and p is the prediction order [7]. The above perceptual weighting filter was incorporated in the proposed subspace approach to shape the residual noise.

2.2. Principles of perceptually based subspace approach

The model used in the subspace approach assumes that the noise signal is additive and uncorrelated with the speech signal, i.e.,

$$\mathbf{y} = \mathbf{x} + \mathbf{n} \quad (2)$$

where \mathbf{y} , \mathbf{x} and \mathbf{n} are the K -dimensional noisy speech, clean speech and noise vectors respectively. Let $\hat{\mathbf{x}} = H \cdot \mathbf{y}$ be a linear estimator of the clean speech \mathbf{x} , where H is a $K \times K$ matrix. The error signal $\boldsymbol{\varepsilon}$ obtained by this estimation is given by:

$$\boldsymbol{\varepsilon} = \hat{\mathbf{x}} - \mathbf{x} = (H - I) \cdot \mathbf{x} + H \cdot \mathbf{n} = \boldsymbol{\varepsilon}_x + \boldsymbol{\varepsilon}_n \quad (3)$$

where $\boldsymbol{\varepsilon}_x$ represents the speech distortion and $\boldsymbol{\varepsilon}_n$ represents the residual noise [1]. Defining the energy of the speech distortion $\overline{\boldsymbol{\varepsilon}_x^2}$ and the energy of the residual noise $\overline{\boldsymbol{\varepsilon}_n^2}$ as:

$$\overline{\boldsymbol{\varepsilon}_x^2} = E[\boldsymbol{\varepsilon}_x^T \boldsymbol{\varepsilon}_x] = \text{tr}(E[\boldsymbol{\varepsilon}_x \boldsymbol{\varepsilon}_x^T]) \quad (4)$$

$$\overline{\boldsymbol{\varepsilon}_n^2} = E[\boldsymbol{\varepsilon}_n^T \boldsymbol{\varepsilon}_n] = \text{tr}(E[\boldsymbol{\varepsilon}_n \boldsymbol{\varepsilon}_n^T]) \quad (5)$$

the optimum linear estimator can be obtained by solving the following time domain constrained optimization problem:

$$\begin{aligned} \min_H \overline{\boldsymbol{\varepsilon}_x^2} \\ \text{subject to : } \frac{1}{K} \overline{\boldsymbol{\varepsilon}_n^2} \leq \sigma^2 \end{aligned} \quad (6)$$

where σ^2 is a positive constant. Such an approach was proposed in [1] for white noise and later extended to colored noise in [4]. An attempt was made in [1] to extend this

approach to the KLT domain using spectral domain constraints. No explicit method was proposed in [1] however to directly shape the spectrum of the residual noise in the frequency domain.

In this paper, we propose the use of a perceptually weighted residual noise in place of the constraint in Eq. (6). The main motivation is to shape the spectrum of the residual noise so that it is not perceptually audible. The perceptually weighted residual noise $\boldsymbol{\varepsilon}_{wn}$ can be obtained as follows:

$$\boldsymbol{\varepsilon}_{wn} = W \cdot \boldsymbol{\varepsilon}_n \quad (7)$$

where W is the $K \times K$ perceptual weighting matrix. W is a lower triangular matrix and has the form:

$$W = \begin{bmatrix} h_0 & 0 & \cdots & 0 \\ h_1 & h_0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ h_{K-1} & h_{K-2} & \cdots & h_0 \end{bmatrix} \quad (8)$$

where h_n , $n = 0, \dots, K-1$, is the truncated impulse response of the perceptual weighting filter $P(z)$ given in Eq. (1).

Defining the energy of the perceptually weighted residual noise as:

$$\begin{aligned} \overline{\boldsymbol{\varepsilon}_{wn}^2} &= E[\boldsymbol{\varepsilon}_{wn}^T \boldsymbol{\varepsilon}_{wn}] = \text{tr}(E[\boldsymbol{\varepsilon}_{wn} \boldsymbol{\varepsilon}_{wn}^T]) \\ &= \text{tr}(W E[\boldsymbol{\varepsilon}_n \boldsymbol{\varepsilon}_n^T] W^T) \end{aligned} \quad (9)$$

the proposed optimal linear estimator can be obtained by solving the following time-domain constrained optimization problem:

$$\begin{aligned} \min_H \overline{\boldsymbol{\varepsilon}_x^2} \\ \text{subject to : } \frac{1}{K} \overline{\boldsymbol{\varepsilon}_{wn}^2} \leq \sigma^2 \end{aligned} \quad (10)$$

It can be shown, using the method of Lagrange multipliers, that the solution to Eq. (10) satisfies the following equation:

$$\mu(W^T W)H + H R_x R_n^{-1} = R_x R_n^{-1} \quad (11)$$

where μ is the Lagrange multiplier.

The techniques proposed in [4] can be used to further simplify the above equation. In [4], a non-unitary matrix V was constructed that simultaneously diagonalized the two matrices R_x and R_n as follows [8]:

$$\begin{aligned} V^T R_x V &= \Lambda_x \\ V^T R_n V &= I \end{aligned} \quad (12)$$

where Λ_x and V are the eigenvalue matrix and eigenvector matrix respectively of $\Sigma = R_n^{-1} R_x$, i.e.,

$$\Sigma V = V \Lambda_x \quad (13)$$

It can be shown that Λ_x is a real matrix [9]. Plugging Eq. (12) into Eq. (11), we get the following:

$$\mu(W^T W)H + HV^{-T} \Lambda_x V^T = V^{-T} \Lambda_x V^T \quad (14)$$

The above equation has the form of the well known Lyapunov equation encountered frequently in control theory. Unfortunately, this equation does not have a closed-form solution, however, techniques developed in [10] can be used to obtain a numerical solution. After solving for H in Eq. (14), the enhanced speech signal can be obtained by: $\hat{x} = H \cdot y$.

3. IMPLEMENTATION

3.1. Covariance matrix estimation

From Eq. (11), it can be seen that the linear estimator H depends on the accurate estimation of the covariance matrices R_x and R_n . Informal listening tests showed that the quality of the enhanced speech is greatly influenced by the estimation of the covariance matrix. In this paper, we use a multiwindow estimator of the covariance matrix [6]. Compared with the commonly used covariance matrix estimators in [1], [2] and [4], the multiwindow estimator reduces the estimation variance greatly, and it can be used in place of the commonly used covariance estimators to improve performance.

To obtain the $K \times K$ multiwindow covariance matrix estimator \hat{R} from the incoming N -dimensional data vector ($K < N$), we first form a $K \times (N - K + 1)$ Hankel matrix S from the incoming data vector [6]. Let $S = [s_1, s_2, \dots, s_{N-K+1}]$, where s_j , $j = 1, \dots, N - K + 1$ is the j -th column of S . Then for each column s_j , \hat{R}_j is computed as:

$$\hat{R}_j = \sum_{i=1}^L (E_i \cdot s_j) \cdot (E_i \cdot s_j)^T \quad (15)$$

where $E_i = \text{diag}(e_i)$ is a $K \times K$ diagonal matrix having in its diagonal the i -th discrete prolate spheroidal sequence (Slepian sequence) vector e_i [11], and L is the number of the Slepian sequences used (L was set to 4 in this paper). The final \hat{R} was obtained as follows:

$$\hat{R} = \frac{1}{N - K + 1} \sum_{j=1}^{N-K+1} \hat{R}_j \quad (16)$$

3.2. Proposed algorithm

The proposed approach can be formulated using the following six steps. For each speech frame:

Step 1: Compute the covariance matrix R_y of the noisy signal, and estimate the matrix $\Sigma = R_n^{-1} R_x$. The noise

covariance matrix R_n can be computed using noise samples collected during speech-absent frames.

Step 2: Perform the eigendecomposition of Σ :

$$\Sigma V = V \Lambda_x$$

Step 3: Assuming that the eigenvalues of Σ are ordered as $\lambda_x(1) \geq \lambda_x(2) \geq \dots \geq \lambda_x(K)$, estimate the dimension of the speech signal subspace as follows:

$$M = \arg \max_{1 \leq m \leq K} \{\lambda_x(m) > 0\}$$

Step 4: Compute the μ value according to:

$$\mu = \begin{cases} \mu_0 - (SNR_{dB})/s, & -5 < SNR_{dB} < 20 \\ 1 & SNR_{dB} \geq 20 \\ 20 & SNR_{dB} \leq -5 \end{cases} \quad (17)$$

where $\mu_0 = 16.2$, $s = 1.32$, $SNR_{dB} = 10 \log_{10} SNR$ and SNR is computed as:

$$SNR = \frac{\text{tr}(V^T R_x V)}{\text{tr}(V^T R_n V)} = \frac{\sum_{k=1}^M \lambda_x^{(k)}}{K} \quad (18)$$

As discussed in [1], the value of μ controls the tradeoff between residual noise and speech distortion. A larger value of μ will yield more speech distortion with less residual noise, and vice versa. To better control this tradeoff, we propose the above method for selecting the value of μ according to the estimated SNR .

Step 5: Form the perceptual weighting matrix W from $P(z)$ (Eq. (1)) and solve the Lyapunov equation (14) for H .

Step 6: Estimate the enhanced speech signal by: $\hat{x} = H_{\text{opt}} \cdot y$.

The error weighting filter $P(z)$ was obtained directly from the noisy speech. This is because we found that, at least for 5 dB speech-shaped noise, the estimated formant frequencies of the corrupted speech were close to those of the clean speech [12]. The values of γ_1 and γ_2 were set to 1 and 0.9 respectively.

The estimator was applied to rectangular-windowed frames of the noisy signal which overlapped each other by 50%. One extra forward frame was used to estimate the covariance matrices and K was set to 40 samples (assuming an $8kH_z$ sampling frequency). The enhanced speech signal was Hamming windowed and combined using the overlap and add approach [7].

4. EXPERIMENTAL RESULTS

For evaluation purposes, we used 20 sentences from the TIMIT database downsampled to $8kH_z$. The sentences were produced by 10 male and 10 female speakers. For colored noise, we used speech shaped noise added to the clean

speech file at $SNR = 5$ dB. The speech shaped noise, included in the HINT database [4], was computed by filtering white noise through an FIR filter with frequency response that matched the long-term spectrum of the sentences in the HINT database [13].

The modified bark spectral distortion (MBSD) measure [14] and the segmental SNR [7] measure were used for evaluation of the proposed perceptually-weighted (PW) subspace approach. The MBSD measure was found to be highly correlated with speech quality [14]. For comparative purposes, we also implemented and evaluated a version of the spectral domain constrained (SDC) approach in [4].

	Male		Female	
	MBSD	s-SNR	MBSD	s-SNR
Noisy speech	1.05	-2.20	1.12	-2.10
SDC approach in [4]	0.79	-3.89	0.89	-4.17
PW approach	0.40	2.01	0.40	1.91

Table 1. Comparative performance for speech-shaped noise at 5 dB, in terms of mean MBSD measure and segmental SNR measure (s-SNR), for 20 TIMIT sentences produced by ten male speakers and ten female speakers

Table 1 gives the mean results in terms of the two objective measures for 20 TIMIT sentences corrupted by speech-shaped noise at 5 dB. The results are given separately for male and female speakers. As can be seen from Table 1, our proposed approach (PW) outperformed the SDC approach in [4] for both male and female speakers. Particularly large improvements were noted for the segmental SNR measure. Informal listening tests also demonstrated significant improvements in speech quality with the proposed method.

5. CONCLUSIONS

A perceptually-based subspace approach for enhancing speech degraded by colored additive noise was proposed in this paper. The proposed approach exploits the frequency masking properties of the human auditory system and obtains better noise shaping. Objective measures and informal listening tests showed significant improvements over other subspace based speech enhancement approaches.

6. REFERENCES

- [1] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 251–266, 1995.
- [2] U. Mittal and N. Phamdo, "Signal/noise KLT based approach for enhancing speech degraded by colored noise," *IEEE Trans. Speech Audio Processing*, vol. 8, pp. 159–167, Mar. 2000.
- [3] A. Rezaeey and S. Gazor, "An adaptive KLT approach for speech enhancement," *IEEE Trans. Speech Audio Processing*, vol. 9, pp. 87–95, Feb. 2001.
- [4] Y. Hu and P. C. Loizou, "A subspace approach for enhancing speech corrupted by colored noise," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, May 2002, Orlando.
- [5] P. Kroon and B. S. Atal, "Predictive coding of speech using analysis-by-synthesis techniques," in *Advances in speech signal processing*, S. Furui and M. Sondhi, Eds., pp. 141–164. Marcel Dekker, NY, 1992.
- [6] L. T. McWhorter and L. L. Scharf, "Multiwindow estimators of correlation," *IEEE Trans. Signal Processing*, vol. 46, pp. 440–448, Feb. 1998.
- [7] Jr. J. R. Deller, J. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*, IEEE Press, NY, 2000.
- [8] S. B. Searle, *Matrix Algebra Useful for Statistics*, John Wiley & Sons, 1982.
- [9] G. Strang, *Linear Algebra and its Applications*, Harcourt Brace Jovanovich, Inc., third edition, 1988.
- [10] R. H. Bartels and G. W. Stewart, "Solution of the matrix equation $AX+XB=C$," *Communication of the ACM*, vol. 15, no. 9, pp. 820–822, 1972.
- [11] D. J. Thomson, "Spectrum estimation and harmonic analysis," *Proceedings of the IEEE*, vol. 70, no. 9, pp. 1055–1096, Sept. 1982.
- [12] G. Parikh, P. Loizou, and Y. Hu, "The effect of noise on the spectrum of vowels: Implications for speech enhancement," submitted to *Int. Conf. Spoken Language Processing*, 2002.
- [13] M. Nilsson, S. Soli, and J. Sullivan, "Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise," *J. Acoust. Soc. Am.*, vol. 95, pp. 1085–1099, 1994.
- [14] W. Yang, M. Benbouchta, and R. Yantorno, "Performance of the modified bark spectral distortion as an objective speech quality measure," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1998, pp. 541–544.