Background Speech Cancellation using a Generalized Subspace Speech Enhancement Method

Radu Mihnea Udrea, Constantin Paleologu, Silviu Ciochina
Telecommunications Department
"Politehnica" University of Bucharest
Bucharest, Romania
mihnea@comm.pub.ro, pale@comm.pub.ro, silviu@comm.pub.ro

***Abstract*— This paper presents a speech enhancement method for reducing undesired babble noise, or background speech, that affects the desired speech, using a generalized subspace approach. The subspace decomposition is obtained with a nonunitary transform based on diagonalization of the clean speech and background distortion covariance matrices. The clean signal is estimating using an optimal subspace estimator that nulls the signal components in the distortion signal subspace and keeps the components in the signal subspace. Objective and subjective measures show a better suppression of background speech that other subspace-based methods that were proposed for white noise.**

Keywords-speech enhancement; colored noise; subspace

I. INTRODUCTION

Over the years, many applications of acoustic noise reduction and speech enhancement require high performance and efficient algorithms. Spectral subtraction [1] is perhaps one of the most popular speech enhancement algorithm due to its low complexity. Even if several methods were proposed [2], [3] to reduce speech distortions and "musical noise" introduced by this algorithm, still there is a compromise to be made between reducing speech distortion and reducing residual noise.

Another approach of more recent speech enhancement algorithms is based on decomposition of the noisy signal in two subspaces: signal subspace and noise subspace. An estimate of the clean signal can be made by nulling the components of the signal in the noise subspace and retaining only the components of the signal in the signal subspace. The subspace decomposition can be done using the eigenvalue decomposition (EVD) [4]-[6] or the singular value decomposition (SVD) [7].

In [4], an optimal estimator that minimizes the speech distortion subject to the constraint that the residual noise fell below a preset threshold is proposed using the eigenvalue decomposition of the covariance matrix. The decomposition of the vector space of the noisy signal into a signal and noise subspace can be obtained by applying the Karhunen–Loéve transform (KLT) to the noisy signal. The KLT components representing the signal subspace were modified by a gain function determined by the estimator, while the remaining KLT components representing the noise subspace were nulled. The enhanced signal was obtained from the inverse KLT of the modified components. This subspace approach was based on the assumption that the input noise was white.

The work in [4] was extended for colored noise. In [5] it is given a proper noise shaping for colored noise without prewhitening, first by classifying the noisy speech frames into speech-dominated and noise-dominated frames and then using a different KLT matrix for these frames to construct the estimator. In [6] a generalized subspace approach with built-in prewhitening for enhancing speech corrupted with colored noise is determined.

In this paper we propose a nonunitary transform, based on the simultaneous diagonalization of the clean speech and the background distortion covariance matrices. No assumptions were made about the covariance matrix of the KLT-transformed noise vectors, hence this estimator is optimal.

This paper is organized as follows. In Section II, the subspace approach using time-domain constraints is presented for white noise and for colored noise (like babble talk). Section II also gives an expression, different that in [6], for the subspace estimator for any type of distortion signal which is uncorrelated to speech. Implementation details are provided in Section III, experimental results are given in Section IV, and the conclusions are given in Section V.

II. SUBSPACE APPROACH FOR SPEECH ENHANCEMENT

The linear model for the clean speech signal assumes that each *K*-dimensional vector *x* can be represented as:

$$\mathbf{x} = \sum_{m=1}^{M} s_m \mathbf{b}_m, \quad M < K \tag{1}$$

where $\{s_1, \ldots, s_M\}$ are zero mean random variables, and $\mathbf{b}_1, \ldots, \mathbf{b}_M$ are *K*-dimensional complex basis vectors, which

 ISBN: 978-1-61208-139-7

are assumed linearly independent. For speech signals, such representation is possible also, when $M<K$.

The model (1) can be expressed as:

$$\mathbf{x} = \mathbf{B} \cdot \mathbf{s} \tag{2}$$

where $\mathbf{B}$ is a $K \times M$ matrix whose rank is M and $\mathbf{s}$ is an M-dimensional vector. The covariance matrix of $\mathbf{x}$ is given by:

$$\mathbf{R}_{\mathbf{x}} \triangleq E\left\{\mathbf{x} \cdot \mathbf{x}^T\right\} = \mathbf{B} \cdot \mathbf{R}_{\mathbf{s}} \cdot \mathbf{B}^T \tag{3}$$

where $\mathbf{R}_{\mathbf{s}}$ is the covariance matrix of the vector $\mathbf{s}$, which is assumed positive definite. Hence, the rank of $\mathbf{R}_{\mathbf{x}}$ is M, and it has $K - M$ zero eigenvalues.

Let $\mathbf{d}$ being the K-dimensional vector of the noise (distortion) signal. Assuming the distortion signal is additive and uncorrelated with the speech signal, we can write the corrupted signal as:

$$\mathbf{y} = \mathbf{B} \cdot \mathbf{s} + \mathbf{d} = \mathbf{x} + \mathbf{d} \tag{4}$$

where $\mathbf{y}$ is the K-dimensional corrupted speech vector.

The clean speech linear estimator will be:

$$\hat{\mathbf{x}} = \mathbf{H} \cdot \mathbf{y} \tag{5}$$

where $\mathbf{H}$ is a $K \times K$ matrix. The error signal resulted from this estimation is given by:

$$\varepsilon = \hat{\mathbf{x}} - \mathbf{x} = (\mathbf{H} - \mathbf{I}) \cdot \mathbf{x} + \mathbf{H} \cdot \mathbf{d} = \varepsilon_x + \varepsilon_d \tag{6}$$

Let

$$\begin{aligned} \overline{\varepsilon_x^2} &= E\left[\varepsilon_x^T \varepsilon_x\right] = tr\left(E\left[\varepsilon_x \varepsilon_x^T\right]\right) \\ \overline{\varepsilon_d^2} &= E\left[\varepsilon_d^T \varepsilon_d\right] = tr\left(E\left[\varepsilon_d \varepsilon_d^T\right]\right) \end{aligned} \tag{7}$$

be the energy of the speech distortion and, respectively, the energy of the residual noise vector. The linear estimator can be obtained [4] by solving the following time-domain constrained (TDC) optimization problem:

$$\begin{aligned} &\min_H \overline{\varepsilon_x^2} \\ &\text{subject to:} \frac{1}{K} \overline{\varepsilon_d^2} \le \alpha \sigma_d^2 \end{aligned} \tag{8}$$

where $0 \le \alpha \le 1$. The estimator derived in this way minimizes the signal distortion over all linear filters which result in the permissible residual noise level. The solution to (8) is given by [4]:

$$\mathbf{H}_{opt} = \mathbf{R}_{\mathbf{x}} \left(\mathbf{R}_{\mathbf{x}} + \mu \mathbf{R}_{\mathbf{d}}\right)^{-1} \tag{9}$$

where $\mathbf{R}_{\mathbf{x}}$ and $\mathbf{R}_{\mathbf{d}}$ are the covariance matrices of the clean speech and noise respectively, and μ is the Lagrange multiplier.

Consider the eigen-decomposition of $\mathbf{R}_{\mathbf{x}}$

$$\mathbf{R}_{\mathbf{x}} = \mathbf{U} \mathbf{\Lambda}_{\mathbf{x}} \mathbf{U}^T \tag{10}$$

where $\mathbf{U}$ is the eigenvector unitary matrix and $\mathbf{\Lambda}_{\mathbf{x}}$ is the diagonal eigenvalue matrix of $\mathbf{R}_{\mathbf{x}}$.

The optimal filter from (9) can be simplified using (10) to:

$$\mathbf{H}_{opt} = \mathbf{U} \mathbf{\Lambda}_{\mathbf{x}} \left(\mathbf{\Lambda}_{\mathbf{x}} + \mu \mathbf{U}^T \mathbf{R}_{\mathbf{d}} \mathbf{U}\right)^{-1} \mathbf{U}^T. \tag{11}$$

A. White Noise Subspace Estimator

For white noise with variance σ_d^2, $R_d = \sigma_d^2 I$ and the estimator from (11) reduces to White Noise Subspace Estimator (WNSE) [4]:

$$\mathbf{H}_{WNSE} = \mathbf{U} \mathbf{\Lambda}_{\mathbf{x}} \left(\mathbf{\Lambda}_{\mathbf{x}} + \mu \sigma_d^2 \mathbf{I}\right)^{-1} \mathbf{U}^T = \mathbf{U} \cdot \mathbf{G}_{WNSE} \cdot \mathbf{U}^T. \tag{12}$$

where

$$\mathbf{G}_{WNSE} = \mathbf{\Lambda}_{\mathbf{x}} \left(\mathbf{\Lambda}_{\mathbf{x}} + \mu \sigma_d^2 \mathbf{I}\right)^{-1}. \tag{13}$$

The gain matrix $\mathbf{G}_{WNSE}$ is diagonal with elements (gains):

$$g_{WNSE}(m) = \frac{\lambda_x(m)}{\lambda_x(m) + \mu \sigma_d^2}. \tag{14}$$

Hence, the signal estimate is obtained by applying the Karhunen-Loève transform (KLT) to the noisy signal, then modify the components of the KLT by a gain function and finally, by inverse KLT of the modified components. A block diagram of this estimator is shown in Fig. 1.

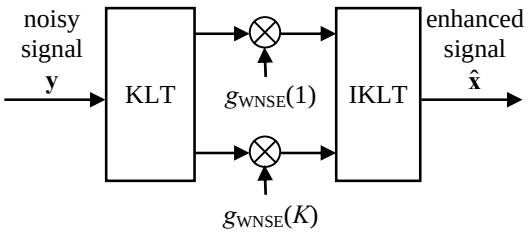

Figure 1. Signal subspace linear estimator

B. Colored Noise Subspace Estimator

If the distortion is not white noise, the matrix $\mathbf{U}^T\mathbf{R_d}\mathbf{U}$ is not diagonal since **U**, being the eigenvector matrix of the symmetric matrix $\mathbf{R_x}$, diagonalizes $\mathbf{R_x}$ and not $\mathbf{R_d}$. There is a matrix **V** that simultaneously diagonalize $\mathbf{R_x}$ and $\mathbf{R_d}$.

As stated in [6], consider the basis matrix $\mathbf{\Sigma}$ satisfying the following equations:

$$\begin{aligned}\mathbf{\Sigma} &= \mathbf{R_d^{-1}R_x}\\ \mathbf{\Sigma V} &= \mathbf{V\Lambda}\end{aligned} \tag{15}$$

where $\mathbf{\Lambda}$ and **V** are the eigenvalue matrix and eigenvector matrix respectively of $\mathbf{\Sigma}$. Applying diagonalizing matrix **V** will result the fully diagonal eigenvalues matrices of $\mathbf{R_x}$ and $\mathbf{R_d}$:

$$\begin{aligned}\mathbf{V}^T\mathbf{R_x V} &= \mathbf{\Lambda_x} \neq \mathbf{\Lambda}\\ \mathbf{V}^T\mathbf{R_d V} &= \mathbf{\Lambda_d}\end{aligned} \tag{16}$$

where $\mathbf{\Lambda_x}$ and $\mathbf{\Lambda_d}$ are the eigenvalue matrices of $\mathbf{R_x}$ and $\mathbf{R_d}$, respectively.

The resulted equations (16) are more general than the relations in [6] (where it is considered that $\mathbf{\Lambda_x} = \mathbf{\Lambda}$ and $\mathbf{\Lambda_d} = \mathbf{I}$). The approach proposed in (16) allows applying the subspace method to any type of distortion signal which is uncorrelated to speech signal.

Applying the eigen-decomposition of $\mathbf{\Sigma}$ from (15) and using (16), the optimal linear Colored Noise Subspace Estimator (CNSE) can be expressed as:

$$\begin{aligned}\mathbf{H}_{CNSE} &= \mathbf{R_d V\Lambda_x}\left(\mathbf{\Lambda_x} + \mu\mathbf{\Lambda_d}\right)^{-1}\mathbf{V}^T =\\ &= \mathbf{V}^{-T}\cdot\mathbf{G}_{CNSE}\cdot\mathbf{V}^T\end{aligned}. \tag{17}$$

where

$$\mathbf{G}_{CNSE} = \mathbf{\Lambda_x}\left(\mathbf{\Lambda_x} + \mu\mathbf{\Lambda_d}\right)^{-1}. \tag{18}$$

In case of the colored noise, the corrupted signal is decorrelated with the non-KLT matrix $\mathbf{V}^T$, then it is modified by the signal subspace gain matrix $\mathbf{G}_{CNSE}$, and, finally, the enhanced signal estimate is obtain by the inverse non-KLT matrix $\mathbf{V}^{-T}$.

Since we have no access to the covariance matrix $\mathbf{R_x}$ of the clean speech signal, the matrix $\mathbf{\Sigma}$ is estimated from the noisy speech signal. Assuming that speech is uncorrelated with noise, we have

$$\mathbf{R_y} = \mathbf{R_x} + \mathbf{R_d}. \tag{19}$$

and so

$$\mathbf{\Sigma} = \mathbf{R_d^{-1}R_x} = \mathbf{R_d^{-1}R_y} - \mathbf{I}. \tag{20}$$

The estimation of μ in the gain function (14) or (18) affects the quality of speech. A large value of μ would reduce the residual noise but would introduce speech distortion. A small value of μ would minimize the speech distortion at the expense of higher values of residual noise. A trade-off between residual noise and speech distortion can be obtained by making μ dependent on the short-time SNR:

$$\mu = \mu_0 - \frac{SNR_{dB}}{s_0}. \tag{21}$$

where μ_0 and s_0 are constants chosen experimentally [6] as explained in the implementation section.

III. ALGORITHM IMPLEMENTATION

The proposed algorithm can be implemented, for each speech frame, as follows:

- The distortion covariance matrix $\mathbf{R_d}$ is computed prior to the starting of the speech signal during speech-absent frames.
- The matrix $\mathbf{\Sigma}$ is estimated using (20) from the noisy signal covariance matrix $\mathbf{R_y}$ and the inverse of $\mathbf{R_d}$.
- The eigen-decomposition of $\mathbf{\Sigma}$ is performed using (15). Extract the eigenvector matrix **V** and eigenvalue matrix $\mathbf{\Lambda}$.
- The dimension of the speech signal subspace is estimated, considering that the eigenvalues of $\mathbf{\Sigma}$ are ordered $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_K$, from:

$$M = \arg\max_{1\leq k\leq K}\left\{\lambda_k > 0\right\}. \tag{22}$$

- The μ factor is computed as a linear function of SNR [6]:

$$\mu = \begin{cases} 5 & SNR_{dB} < -5\\ \mu_0 - \dfrac{SNR_{dB}}{s_0} & -5 < SNR_{dB} < 20\\ 1 & SNR_{dB} \geq 20\end{cases}. \tag{23}$$

  where $\mu_0 = 4.2, s_0 = 6.25, SNR_{dB} = 10\log_{10} SNR$.
- SNR can be computed directly from the eigenvalues λ_k of $\mathbf{\Sigma}$ using the following equation [6]:

$$SNR = \frac{tr\left(\mathbf{V}^T\mathbf{R_x V}\right)}{tr\left(\mathbf{V}^T\mathbf{R_d V}\right)} = \frac{\sum_{k=1}^{M}\lambda_k}{K}. \tag{24}$$

 ISBN: 978-1-61208-139-7

- Compute the optimal estimator $\mathbf{G}_{CNSE}$ using (18) and estimate the desired signal using (5).

The covariance matrices $\mathbf{R}_x$ and $\mathbf{R}_d$ were estimated as Toeplitz matrices using K samples of the unbiased autorcorrelation sequence, without using future or past frames. We choose $K = 40$ samples for speech sampled at 8 kHz. The estimators were applied to frames of the corrupted signal 50% overlapped each other. The covariance matrices were estimated by windowing with rectangular windows. The enhanced speech signal estimation was obtained using overlap and add approach with Hamming windowing.

IV. EXPERIMENTAL RESULTS

We used 20 sentences produced by 10 male and 10 female speakers. For distortion signal we used other speech signals added as babble noise to the clean speech at SNR = 5 dB. For comparative purposes, we also evaluated the algorithm performances applying as a distortion signal white noise at the same SNR.

The Perceptual Evaluation for Speech Quality (PESQ) distance measure and the overall (global) SNR [8] measures were adopted for evaluation of the proposed algorithms. We used the ITU-T Recommendation P.862 (PESQ) [9] to obtain a perceptual evaluation of the enhanced speech quality. The Mean Opinion Score (MOS) obtained in the evaluation process is between 0 and 5 where 0 represents a very annoying distortion of the perceived signal and 5 represents imperceptible quality degradation.

TABLE I. MEAN PESQ AND MEAN GLOBAL SNR FOR WHITE NOISE DISTORTION AT 5dB

	Male Speakers		**Female Speakers**	
	SNR	*PESQ*	*SNR*	*PESQ*
Noisy Speech	4.6 dB	1.78	4.8 dB	1.71
WNSE	11.1 dB	2.36	10.9 dB	2.21
CNSE	11.3 dB	2.47	10.8 dB	2.22

TABLE II. MEAN PESQ MEAN GLOBAL SNR FOR BABBLE SPEECH DISTORTION AT 5dB

	Male Speakers		**Female Speakers**	
	SNR	*PESQ*	*SNR*	*PESQ*
Noisy Speech	5.3 dB	0.72	5.2 dB	0.69
WNSE	6.7 dB	1.06	6.9 dB	0.81
CNSE	7.3 dB	1.45	7.1 dB	1.42

Tables I and II give the mean results for 20 TIMIT sentences corrupted by speechshaped noise at 5 dB. The results are given separately for male and female speakers. As can be seen from Tables I and II, in case of speech corrupted of white noise at 5dB SNR, the proposed approach reduces to Ephraim and Van Trees approach [4]. In case of speech corrupted by background babble talk distortion, our proposed approach (CNSE) outperformed Ephraim and Trees approach [4] for both male and female speakers.

Subjective listening tests confirmed the results in Tables I and II and that with the proposed method, the background noise was imperceptible. Since in our experiments, no voice detection algorithm (VAD) was used to update the noise covariance matrix, we expect further improvements in performance if we use a reliable VAD algorithm to update the noise covariance matrix.

V. CONCLUSIONS

A speech enhancement for reducing undesired babble noise, or background speech, that affects the desired speech, using a generalized subspace approach was proposed. The proposed approach is based on the simultaneous diagonalization of the covariance matrices of the speech signal and the colored noise signal. In case of the colored noise, the corrupted signal is decorrelated with the non-KLT matrix, it is modified by a gain matrix, and, finally, the enhanced speech is estimated by inverse non-KLT matrix. Better SNR and perceptual scores were obtained that applying the standard KLT decomposition used by Ephraim and Van Trees [3] for enhancing speech corrupted by white noise.

ACKNOWLEDGMENT

The work has been funded by the Sectoral Operational Programme Human Resources Development 2007-2013 of the Romanian Ministry of Labour, Family and Social Protection through the Financial Agreement POSDRU/89/1.5/S/62557.

REFERENCES

[1] M. Berouti, R. Schwartz, and J. Makhoul, “Enhancement of speech corrupted by acoustic noise,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. 208–211, 1979.

[2] R. Martin, “Spectral subtraction based on minimum statistics”, Proc. Eur. Signal Process., pp. 1182-1185, 1994.

[3] R. M. Udrea, N. Vizireanu, S. Ciochina, and S. Halunga “Nonlinear spectral subtraction method for colored noise reduction using multi-band Bark scale”, Signal Processing, Volume 88 Issue 5, Elsevier North-Holland, Inc., pp. 1299-1303, May 2008.

[4] 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.

[5] 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.

[6] Y. Hu and P. C. Loizou, “A generalized subspace approach for enhancing speech corrupted by colored noise,” IEEE Transactions on Speech and Audio Processing, vol. 11, no. 4, pp. 334-341, 2003.

[7] S. H. Jensen, P. C. Hansen, S. D. Hansen, and J. A. Sørensen, “Reduction of broad-band noise in speech by truncated QSVD,” IEEE Trans. Speech Audio Processing, vol. 3, pp. 439–448, Nov. 1995.

[8] J. R. Deller, J. Hansen, and J. G. Proakis, Discrete-Time Processing of Speech Signals. New York: IEEE Press, 2000.

[9] ITU-T, Perceptual evaluation of speech quality PESQ, an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs, ITU-T Recommendation P.862, 2000.

 ISBN: 978-1-61208-139-7