SMOOTHED SUBSPACE BASED NOISE SUPPRESSION WITH APPLICATION TO SPEECH ENHANCEMENT

Jesper Jensen, Richard C. Hendriks, and Richard Heusdens

Dept. of Mediamatics
Delft University of Technology
Delft, The Netherlands
E-mail: {J.Jensen, R.C.Hendriks, R.Heusdens}@ewi.tudelft.nl

Søren Holdt Jensen

Dept. of Communication Technology
Aalborg University
Aalborg, Denmark
E-mail: shj@kom.aau.dk

ABSTRACT

Subspace based noise suppression schemes typically rely on eigenvalue estimates of covariance matrices of successive noisy signal frames. We propose in this paper a scheme for improving these estimates, and, consequently, the performance of the noise suppressor. More specifically, the presented scheme aims at combining past and current eigenvalue estimates into approximately stationary time series in order to obtain a smoothed eigenvalue estimator with a reduced variance. The scheme is general in the sense that it is applicable to essentially any subspace-based noise suppression scheme. In simulation experiments with speech signals degraded by additive white Gaussian noise, the proposed scheme shows improvements over the traditional non-smoothed approach for a range of objective quality measures. Further, in a subjective preference test, the proposed method was prefered in more than 90% of the cases.

1. INTRODUCTION

With the steady growth of mobile, digital voice communications systems, there is an increasing demand for such systems to work well in acoustically noisy environments. Since most of these systems are designed for nearly noise-free input speech signals, they do not perform well when the input signal is degraded by acoustic background noise. One solution to this problem is to apply a speech enhancement algorithm as a pre-processor to reduce the background noise in the noisy speech signal.

Classical approaches for single-channel noise reduction include methods based on the short-time Fourier transform (STFT), e.g. [1], and model based methods which attempt to exploit apriori speech production knowledge, e.g. [2, 3]. More recently, the subspace based approach [4] was proposed. This scheme exploits the fact that the covariance matrix of a noisy speech signal frame can be decomposed into two mutually orthogonal vector spaces: a signal (+noise) subspace and a noise subspace. Noise reduction is obtained by discarding the noise subspace completely, while modifying the noisy speech components in the signal (+noise) subspace. Later, extensions where presented to allow for coloured noise, e.g. [5], and to take into account the perceptual effects of the human auditory system, e.g. [6].

The subspace based enhancement scheme relies on eigenvalues of the covariance matrix of the noisy speech signal, e.g. [4, 7, 8]. In practice, covariance matrix estimates and corresponding eigenvalues are computed on a frame-by-frame basis throughout the signal to be enhanced. Unfortunately, due to the variance of the covariance matrix estimates and consequently the corresponding eigenvalues, the resulting enhanced speech signal may contain a significant amount of perceptually disturbing ('musical') residual noise. In [4] this problem was reduced by introducing estimators which minimize the signal distortion subject to energy constraints on the residual noise.

This research was partly supported by Philips Research and the Technology Foundation STW, applied science division of NWO and the technology programme of the ministry of Economics Affairs. S. H. Jensen was supported in part by Julie Damm's Studiefond.

The problem of musical noise is well-known for the class of STFT based spectral-subtraction type algorithms, e.g. [9], where the gain functions applied to the noisy signal rely on estimates of the power spectral density (psd) of the noisy signal; here, the traditional solution is to average psd estimates across time in order to reduce the variance of the estimates. The problem of determining the number of time-domain neighbors to use for computing this smoothed psd estimate is delicate; if too few are used, the variance of the estimator is not reduced enough and the musical noise problem remains, but if too many are used, the signal region across which the average is computed can not be assumed stationary, and the estimate becomes biased, leading to speech distortions in the enhanced signal.

In this paper we propose a smoothed subspace based noise suppression approach for speech enhancement. Along the same lines as for the smoothed spectral subtraction algorithm described above, the eigenvalues of the noisy covariance matrix are computed as an average of the current and possibly several past eigenvalues. We show that the decision as to which and how many of the past eigenvalues to use to compute this average is important to achieve good performance, and solve this problem using a generalized likelihood ratio test. More specifically, we aim at forming averages using time sequences of eigenvalues which can be considered roughly stationary. The proposed method is general and can be applied to essentially any subspace based enhancement algorithm. We show in simulation experiments with speech signals degraded by white noise that the smoothed subspace approach leads to objective as well as subjective performance gains over a traditional, non-smoothed subspace algorithm.

2. THE SIGNAL SUBSPACE APPROACH FOR NOISE REDUCTION

To facilitate our discussion, we review here the classical subspace based approach for noise reduction, see e.g. [4, 7]. We consider a signal model of the form $x = s + w$, where $x \in \mathbb{R}^N$ denotes an observed noisy speech signal vector, s denotes the clean speech signal and w is an additive noise vector. We assume that the clean speech and noise processes are uncorrelated. Further, we restrict ourselves to the white noise case where the noise covariance matrix is of the form $R_w \triangleq E\{ww^T\} = \sigma_w^2 I$, where $E\{\cdot\}$ denotes the statistical expectation operator, $(\cdot)^T$ denotes vector transposition, σ_w^2 is the noise variance, and I is the identity operator in $\mathbb{R}^N$. Let $R_s \in \mathbb{R}^{N\times N}$ denote the covariance matrix of the clean signal, and let $R_s = U \Lambda_s U^T$ denote its eigenvalue decomposition (EVD); the matrix $U \in \mathbb{R}^{N\times N}$ is unitary and contains the eigenvectors as columns (U is also called the Karhunen-Loéve transform matrix), and $\Lambda_s = diag(\lambda_{s_1}, \ldots, \lambda_{s_K}, 0, \ldots, 0), K \leq N$ is a diagonal matrix with the eigenvalues $\lambda_{s_1} \geq \lambda_{s_2} \geq \cdots \lambda_{s_K} \geq 0$ on the main diagonal. Using that w is white and uncorrelated with s we can write the noisy covariance matrix as $R_x = R_s + R_w = U\left(\Lambda_s + I\,\sigma_w^2\right)U^T$. We see that R_x, R_s, and R_w share their eigenvectors, and that the noisy eigenvalues are simply $\lambda_{x_k} = \lambda_{s_k} + \sigma_w^2$, $k = 1, \ldots, N$. It is convenient to partition U as $U = [U_1\ U_2]$, where $U_1 \in \mathbb{R}^{N\times K}$ consti-

tutes an orthonormal basis for the signal (+noise) subspace, while $U_2 \in \mathbb{R}^{N\times(N-K)}$ constitutes a basis for the noise subspace; clearly, we have $U_1^T U_2 = 0$.

Let $\hat{s} = Hx$ be a linear estimator of the clean speech vector s, where $H \in \mathbb{R}^{N\times N}$ is a filtering matrix. We wish to find the matrix H^* which minimizes the power of the estimation error $^2 = E\|s - \hat{s}\|_2^2 = E\|s - Hx\|_2^2 = \text{tr}E\{(s-Hx)(s-Hx)^T\}$. Solving $\frac{^2}{H} = 0$ for H results in the optimal linear estimator given by

$$H^* = U \begin{bmatrix} G & 0 \\ 0 & 0 \end{bmatrix} U^T = U_1 G U_1^T, \tag{1}$$

where $G = diag(\ _{s_1}/\ ^2_w + \ _{s_1}, \cdots, \ _{s_K}/\ ^2_w + \ _{s_K}) \in \mathbb{R}^{K\times K}$. Using that $\ _{x_k} = \ _{s_k} + \ ^2_w$, we can rewrite G as follows

$$G = I - diag(\ ^2_w/\ _{x_1}, \ldots, \ ^2_w/\ _{x_K}). \tag{2}$$

Thus, in practice we compute the matrix G using the eigenvalues $_{x_k}$ of the covariance matrix R_x of the noisy signal and knowledge of the noise level ^2_w. From Eq. (1) we see that the optimal linear estimate $\hat{s} = H^*x$ is found by first applying the KLT to the noisy signal, resulting in $z = U^T x$, then modifying the KLT coefficients using a diagonal matrix, and finally obtaining the enhanced signal vector by back-transforming the modified KLT coefficients using the inverse KLT (pre-multiplication with U).

3. THE SMOOTHED SUBSPACE APPROACH

The subspace based enhancement scheme relies on eigenvalues of covariance matrices R_x of noisy signal frames, see Eq. (2). However, the true, underlying covariance matrices are not known but must be estimated from the available noisy signal, and therefore the eigenvalues of these estimated covariance matrices are only estimates of the true underlying eigenvalues. In this paper we propose a novel technique for improving these eigenvalue estimates using a smoothing approach, where a given eigenvalue of R_x in frame n is estimated as an average of several measured eigenvalues from the current and previous frames $\tilde{}_x(n) = \frac{1}{L} \sum_{l=n-L+1}^{n} \hat{}_x(l)$, where $\hat{}_x(l)$ denotes the eigenvalue of the covariance matrix R_{xx} estimated in frame l. The goal here is to obtain an estimator $\tilde{}_x(n)$ with a smaller variance than that of the measured eigenvalue $\hat{}_x(n)$; for example, if the measured eigenvalues $\hat{}_x(l), l = n-L+1, \ldots, n$ can be assumed to be independently and identically distributed (iid), then the variance of the estimator $\tilde{}_x(n)$ is $1/L$ that of the individual estimates $\hat{}_x(n)$ [10].

3.1 An Initial Experiment

A natural first approach is to compute the smoothed estimator $\tilde{}_{x_k}(n)$ of the k'th eigenvalue in the n'th frame using the k'th eigenvalues of previous frames[1], i.e.,

$$\tilde{}_{x_k}(n) = \sum_{l=n-L+1}^{n} \hat{}_{x_k}(l). \tag{3}$$

We see that this approach bears similarities to traditional smoothed spectral subtraction algorithms [9], where a smoothed power spectral estimate in discrete Fourier transform (DFT) in bin number k is obtained as an average of previous power spectral estimates in DFT bin number k.

We tried this idea out in an initial experiment using the estimator in Eq. (2) applied to succesive speech frames taken with 75% overlap throughout a total of 20 speech signals degraded with additive white noise at a global SNR of 20 dB (a detailed description of the test setup as well as the objective performance measures used can be found in Sec. 4). The value of L was kept constant for each of the eigenvalues and across time. We then evaluated the quality of the enhanced waveforms using a number of objective quality measures averaged across the 20 enhanced speech signals. Specifically, Fig. 1 shows the resulting average SNR, Seg-SNR, COSH measure (symmetrized Itakura-Saito (IS) measure) and Itakura distance (log-likelihood ratio, LLR) [11, 12] for the enhanced waveforms. We see from this figure that rather than improving performance, the smoothing applied in this way degrades objective quality. In other words, the advantages of the smoothing techniques described in [9] for spectral subtraction cannot be directly adopted to subspace based noise suppression schemes.

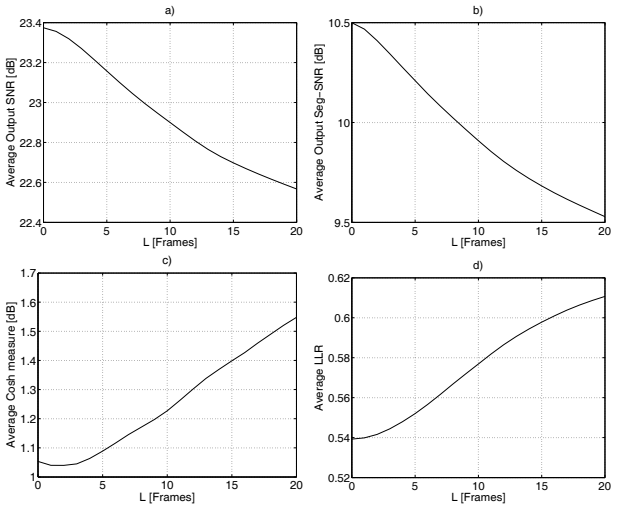

Figure 1: Objective performance as a function of number of past frames used for estimating eigenvalues of R_x.

To understand the reason for this result we note that any signal frame enhanced with the subspace approach can be described in terms of the following orthogonal subspace decomposition:

$$\hat{s} = H^*x = \sum_{k=1}^{K} g_k x_k \text{ with } x_k = u_k u_k^T x,$$

where u_k is the k'th column in U, x_k is the orthogonal projection of x onto the subspace spanned by u_k, and g_k is the k'th diagonal element of the matrix G. Observing that for stationary processes the eigenfunctions u_k are typically sinusoidal-like[2] [13],we conclude that the subspace signals x_k reside in limited frequency regions. Since the eigenvalues of R_x estimated in a given frame are ordered according to their *magnitude* and not according to the frequency content of their corresponding subspace signal, we see that the frequency content of the k'th subspace may change across time, eventhough the noisy signal is stationary. Thus, smoothing of the k'th eigenvalue across time corresponds, in fact, to making the unreasonable assumption that spectrally remote events should be combined and treated as part of the same stationary time series.

3.2 A Generalized Scheme for Eigenvalue Smoothing

In order to develop a subspace based scheme where smoothing of eigenvalues across time is advantageous, we generalize the scheme above. Here we allow for smoothing across subspaces with similar *frequency* content; in Sec. 4 we show that this generalization leads

[1] We assume that eigenvalues of a given estimate of R_x appear in decreasing order. Thus, $\hat{}_{x_k}(l)$ refers to the k'th largest eigenvalue in frame l.

[2] This observation is especially clear in the case of circulant or Toeplitz structured covariance matrices which are known to be diagonalizable and asymptotically ($N \rightarrow$) diagonalizable, respectively, by the Fourier transform matrix.

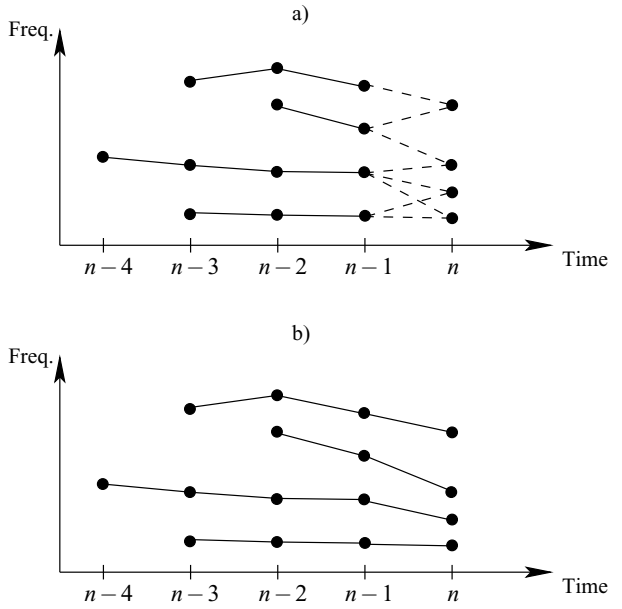

Figure 2: Time-frequency representation of subspace signals x_k (represented by dots) for successive frames. Solid lines indicate which subspaces/eigenvalues are combined to form smoothed estimates. a) Smoothed eigenvalue estimates have been computed for frame $n-1$. Dotted lines indicate possibilities of extending sequences from frames $n-1$ to include eigenvalue estimates $\hat{\ }_x(n)$ from the current frame. b) Solid lines indicate stationary eigenvalue sequences used for computing smoothed eigenvalue estimates $\tilde{\ }_x(n)$ of the current frame.

to performance gains over the traditional, non-smoothed case. More specifically, instead of requiring the smoothed estimate $\tilde{\ }_{x_k}(n)$ to be computed using eigenvalue number k in each of the $L-1$ previous frames, we now choose from *several* eigenvalues in each of the previous frames. Further, the value of L is no longer fixed, but may vary for each eigenvalue estimate $\tilde{\ }_{x_k}(n)$. The task at hand for a given frame is now for each eigenvalue of R_x to decide which and how many (if any) of the eigenvalues in the previous frames should contribute to the smoothed estimate $\tilde{\ }_{x_k}(n)$. In order to approach the iid assumption mentioned above, we require that eigenvalues contributing to a smoothed estimate form an approximately stationary sequence.

To do so we represent each subspace signal x_k by two quantities, i) a frequency parameter computed as

$$_k = \arg\max |X_k(e^j \ )|^2,$$

where $X_k(\cdot)$ denotes the discrete Fourier transform of x_k, and ii) the power of x_k estimated as $E\{\|x_k\|^2\} \approx \hat{\ }_{x_k}$. By doing so, we can plot the subspace signals in the time-frequency plane as shown in Fig. 2a, where each subspace signal is represented as a dot located at its corresponding frequency $_k$.

Fig. 2 illustrates how the eigenvalue sequences used for computing smoothed eigenvalue estimates $\tilde{\ }_{x_k}(n)$ are formed. The solid lines in Fig. 2a indicate which previous eigenvalues contribute to the smoothed eigenvalue estimates in frame $n-1$. For example, the smoothed estimate of the eigenvalue corresponding to the highest frequency content is computed using two previous eigenvalues. In order to form stationary eigenvalue sequences for computing $\tilde{\ }_{x_k}(n)$, we re-use the eigenvalue sequences found for frame $n-1$. In principle, a valid eigenvalue sequence may consist of any eigenvalue $\hat{\ }_{x_k}(n)$ measured in frame n concatenated with eigenvalues from any of the sequences already formed in frame $n-1$ (the solid lines). We require, though, that the frequency content of subspaces corresponding to successive eigenvalues in a sequence does not change drastically. More specifically, we require that $_j(n)/\ _k(n-1) < \ $, where $_k(n-1)$ and $_j(n)$ describe the frequency content in any pair of subspaces taken from frames $n-1$ and n, respectively, and $\ $ is a threshold value. Imposing this constraint limits the number of possible subspace/eigenvalue concatenations to the combinations marked with dashed lines in Fig. 2a.

To explain how stationary eigenvalue sequences used for computing $\tilde{\ }_x(n)$ are formed, let $[\hat{\ }_x(n-L'+1)\cdots\hat{\ }_x(n-1)]$ represent any of the stationary eigenvalue sequences found for frame $n-1$, and let $\hat{\ }_x(n)$ be any of the eigenvalues measured in frame n. For each of the possible subspace/eigenvalue concatenations marked with dashed lines in Fig. 2a, we simply build a stationary eigenvalue sequence by considering sequences of increasing length, starting with the shortest possible sequence $[\hat{\ }_x(n-l)\cdots\hat{\ }_x(n)]$ with $l=1$. Assuming that the elements of the sequence are Gaussian distributed, we apply a generalized likelihood ratio test to determine if the sequence can be considered stationary. If this is not the case, the procedure is terminated, and the resulting stationary sequence simply consists of one element, namely $\hat{\ }_x(n)$. If the two-element sequence is found stationary, we increase l by one and repeat the process with the extended sequence, until either the sequence fails the stationarity test or the end of the sequence is reached. Having generated in this way a stationary eigenvalue sequence for each of the dashed lines in Fig. 2a, we are finally in a position to select a subset of these sequences such that each eigenvalue in frame n is assigned exactly one sequence, and such that no sequence computed for frame $n-1$ is used more than once[3]. Motivated by the fact that the sequence length is roughly inversely proportional to the variance of the corresponding smoothed estimator, we select in this work the subset of eigenvalue sequences having the largest total length. The problem of finding this subset is a so-called linear assignment problem for which several optimal and computationally efficient algorithms exist, see e.g. [14]. Fig. 2b shows a possible outcome of this procedure for frame n. The smoothed estimators $\tilde{\ }_{x_k}(n)$ of each eigenvalue of R_x in frame n can now be computed as the average value within each sequence. Finally, the smoothed estimators $\tilde{\ }_{x_k}(n)$ are inserted into Eq. (2) to form a (modified) filtering matrix G for frame n. The enhancement algorithm as such remains unchanged; in fact, the proposed scheme merely improves the quality of the noisy eigenvalue estimates and can therefore be used in combination with essentially any subspace based enhancement scheme.

4. SIMULATION EXPERIMENTS

We evaluate the presented algorithm in simulation experiments with six different speech signal excerpts, three female and three male, sampled at 8 kHz and with a duration of 4-5 seconds each. We construct noisy speech signals by adding white Gaussian noise at different SNR levels. As in [4] we choose the dimension of the covariance matrices R_x to $N=40$ samples; these matrices are estimated from segments of 160 samples. We perform an eigenvalue-decomposition of R_x to find the eigenvalues $\hat{\ }_{x_k}(n)$. We consider two ways of determining the filter matrix G in Eq. (2), either by inserting the smoothed eigenvalues $\tilde{\ }_{x_k}(n)$ found with the proposed algorithm, or, as is traditionally done, by inserting directly the eigenvalues $\hat{\ }_{x_k}(n)$ of the covariance matrix R_x estimated from the segment in question. We use the latter approach as a reference method to which we compare the proposed scheme. In this work, the dimension K of the signal subspace in segment n is determined as the number of estimated eigenvalues $\tilde{\ }_{x_k}(n)$ above the noise floor

[3] Enforcing this one-to-one correspondence between past and present subspaces is reasonable when for example the subspace signals represent the harmonics in a voiced speech region.

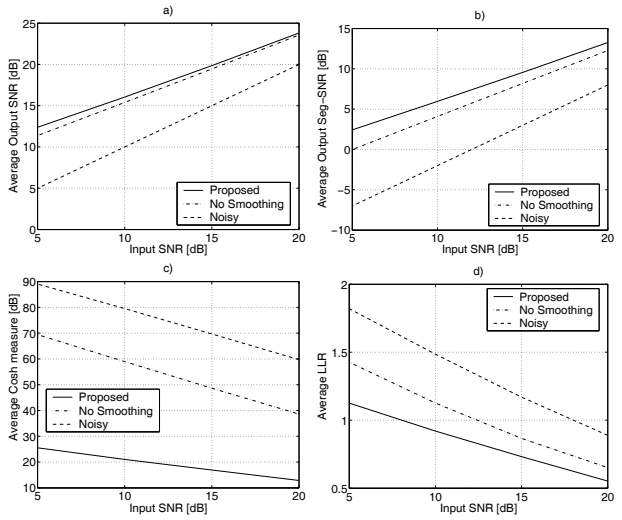

Figure 3: Objective performance scores as a function of input SNR, averaged across six speech signals. a) SNR, b) Seg-SNR, c) COSH measure, and d) Itakura distance (log-likelihood ratio, LLR).

$\frac{2}{w}$: $K = |\{\tilde{}_{x_k}(n) : \tilde{}_{x_k}(n) > \frac{2}{w}, k = 1,\ldots,N\}|$; for the reference method, K is given by a similar expression with $\hat{}_{x_k}(n)$ replacing $\tilde{}_{x_k}(n)$. In both cases, the enhanced signal is constructed by overlap-adding enhanced frames using 75% overlap between successive frames.

In order to evaluate the quality of the enhanced waveforms we apply a number of objective speech quality measures. These include the signal-to-noise ratio (SNR) defined as $\mathrm{SNR}(s', s') = \|s'\|^2/\|\hat{s}' - s'\|^2$, where s' and $\hat{s}'$ denote the clean and enhanced signal, respectively, and the segmental SNR (Seg-SNR) defined as the average SNR computed across signal segments taken troughout s' and $\hat{s}'$, respectively. Also we apply spectral distortion measures including the COSH measure (symmetrized Itakura-Saito (IS) measure) [11] and the Itakura distance measure (log-likelihood ratio, LLR), e.g. [12].

Fig. 3 shows objective quality scores averaged across the six test signals as a function of input SNR for the proposed method, the non-smoothed method, as well as for the unprocessed noisy input signal. We see that the proposed method succeeds in improving the objective scores consistently across the range of input SNRs.

To study the performance of the proposed method further we conducted an OAB subjective preference test using two female and two male speech signals degraded with additive white Gaussian noise at SNRs 20 and 10 dB. The noisy signals were enhanced with the non-smoothed as well as the smoothed enhancement method (the proposed scheme). We presented signal triplets consisting of the original noise-free signal, followed by the two enhanced versions in randomized order. The task of the listener was for each signal triplet to decide which of the two enhanced signals had the highest subjective quality. Each such triplet was repeated four times. Ten listeners participated in the test (the authors not included), and the listeners were allowed to listen to each triplet using headphones as many times as needed to make a decision. Table 1 shows the relative preference for the proposed method for the female speakers, f1 and f2, and for the male speakers m1 and m2. The subjective test reveals a clear preference for the proposed method.

5. CONCLUSION

We have presented a novel scheme for improving the eigenvalue estimates of successive noisy covariance matrices. The presented scheme aims at finding stationary eigenvalue sequences across time,

SNR	f1	f2	m1	m2
20	95	90	100	100
10	93	95	83	93

Table 1: Relative preference [%] for proposed method over non-smoothed method.

in order for a smoothed estimator to approach the underlying expected value. The scheme is general in the sense that it is applicable to essentially any subspace-based noise suppression scheme. When applied to the problem of enhancing speech signals degraded by additive white Gaussian noise, the proposed scheme shows objective as well as subjective improvements over the traditional, non-smoothed, approach. More specifically, in a subjective preference test, the proposed method was prefered on average in more than 90% of the cases.

REFERENCES

[1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, no. 6, pp. 1109–1121, December 1984.

[2] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proc. IEEE*, vol. 80, no. 10, pp. :1526 – 1555, Oct 1993.

[3] J. H. L. Hansen and M. A. Clements, "Constrained iterative speech enhancement with application to speech recognition," *IEEE Trans. Speech, Audio Processing*, vol. 39, no. 4, pp. 795–805, April 1991.

[4] Y. Ephraim and H. L Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech, Audio Processing*, vol. 3, no. 4, pp. 251–265, July 1995.

[5] H. Lev-Ari and Y. Ephraim, "Extension of the signal subspace speech enhancement approach to colored noise," *IEEE Signal Processing Lett.*, vol. 10, no. 4, pp. 104–106, April 2003.

[6] F. Jabloun and B. Champagne, "Incorporating the human hearing properties in the signal subspace approach for speech enhancement," *IEEE Trans. Speech, Audio Processing*, vol. 11, no. 6, pp. 700–708, November 2003.

[7] B. de Moor, "The singular value decomposition and long and short spaces of noisy matrices," *IEEE Trans. Signal Processing*, vol. 41, no. 9, pp. 2826–2838, September 1993.

[8] S. H. Jensen et al., "Reduction of broad-band noise in speech by truncated qsvd," *IEEE Trans. Speech, Audio Processing*, vol. 3, no. 6, pp. 439–448, November 1995.

[9] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Speech, Audio Processing*, vol. ASSP-27, no. 2, pp. 113–120, April 1979.

[10] A. Leon-Garcia, *Probability and Random Processes for Electrical Engineering*, Addison-Wesley Publishing Company, 1994.

[11] A. H. Gray, Jr. and J. D. Markel, "Distance measures for speech processing," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, no. 5, pp. 380–391, October 1976.

[12] J. R. Deller, Jr., J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*, IEEE Press, 2000.

[13] C. W. Therrien, *Discrete Random Signals and Statistical Signal Processing*, Prentice-Hall International, Inc., 1992.

[14] D. B. West, *Introduction to Graph Theory*, Prentice Hall, Inc., Upper Saddle River, NJ, 1996.