A MULTI-MICROPHONE SIGNAL SUBSPACE APPROACH FOR SPEECH ENHANCEMENT

Firas Jabloun and Benoît Champagne

Department of Electrical & Computer Engineering, McGill University
3480 University Street, Montreal, Canada, H3A 2A7
firas@tsp.ece.mcgill.ca, champagne@ece.mcgil.ca
www.tsp.ece.mcgill.ca/∼firas

ABSTRACT

In this paper, we extend the single microphone signal subspace approach for speech enhancement, to a multi-microphone design. In the single microphone case, the trade-off between speech quality and intelligibility is an handicap which limits its performance. This is because it is based on a linear speech model which does not usually offer enough degrees of freedom for noise reduction. In our method, we show how we can easily, and with comparable computational complexity, get more degrees of freedom by using signals from more than one microphone. Experimental results show that this leads to improvements in the noise reduction performance.

1. INTRODUCTION

Single microphone techniques for noise reduction are widely used in most of today's telecommunication systems. They owe their popularity to their simplicity and ease of implementation. In general, in these methods, the noisy signal is transformed to an appropriate domain and filtered by a usually data dependent filter. This filter uses noise statistics gathered during non-speech activity periods to improve the speech quality while trying to minimize the signal distortion. Finally, an inverse transform is applied to recover the enhanced signal in the time domain.

In the spectral subtraction method [1], noise reduction is performed in the frequency domain using a data independent transform, namely, the Discrete Fourier Transform (DFT). In the subspace decomposition method [2][3][4], on the other hand, a data dependent transform, called the Karhunen Loeve Transform (KLT) is used. The major drawback of these methods is that they introduce a residual noise which has an annoying noticeable tonal characteristic. This processing artifact is usually referred to as *musical noise*.

The multi-microphone approach is another promising class for noise reduction. In this approach, the methods developed improve the speech quality by rejecting interfering signals coming from directions different from a desired look direction [5]. However, to achieve an acceptable performance, we need a large number of microphones. Unfortunately, this is not practical in general in terms of spatial placement and the total cost of the whole system.

To cope with the drawbacks of both classes, combinations of the single and multi-microphone techniques have been proposed recently in which a beamformer is followed by a post-filter. The Post-filter can take the form of Wiener filtering [6] or spectral subtraction [7]. In a similar manner, we show in this paper how to design a multi-microphone system with a post-filter derived from the signal subspace decomposition. The covariance matrices required to design the eigenfilter of [2] is approximated from data gathered from different microphones. Our approach provides more degrees of freedom for noise reduction than the single microphone case, hence a better speech quality while maintaining the signal distortion minimum. These improvements are supported by experimental results

2. NOTATION AND SIGNAL MODEL

Let $\mathbf{x} = [x_1, x_2, \ldots, x_P]^T$ be the noisy signal vector of P samples. Here x_i denote $x(n - i + 1)$ for some discrete time index n which is omitted for simplicity of notation. We assume that the noise is additive and uncorrelated with the speech signal, so that the vector $\mathbf{x}$ can be written

$$\mathbf{x} = \mathbf{s} + \mathbf{w} \tag{1}$$

where $\mathbf{s}$ is the clean speech vector and $\mathbf{w}$ is the noise vector. The noise covariance matrix R_w is considered to be known since it can be approximated during non speech activity. The white noise assumption is considered in this paper, that is $R_w = \sigma^2 I$, since prewhitening can always be used otherwise. The data covariance matrix is then $R_x = R_s + \sigma^2 I$, where R_s is the clean speech covariance matrix. Therefore,

R_x would be a full rank matrix with rank$(R_x) = P$. However, R_s is assumed to be rank deficient with rank$(R_s) = K < P$ because $\mathbf{s}$ is described by a linear model of order K [2].

Therefore both R_s and R_x have the same eigenvectors and their eigenvalues are related as follows:

$$\lambda_{x,i} = \begin{cases} \lambda_{s,i} + \sigma^2 & i = 1, \ldots, K. \\ \sigma^2 & i = K+1, \ldots, P. \end{cases} \quad (2)$$

So if Λ_s is a $K \times K$ diagonal matrix with the eigenvalues of R_s on the diagonal then

$$R_x = [U_1 U_2] \begin{bmatrix} \Lambda_s & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} U_1^T \\ U_2^T \end{bmatrix} + \sigma^2 I \quad (3)$$

where the columns of U_1 span the signal subspace and those of U_2 span the noise subspace. Note that because the covariance matrices involved are real symmetric, all the quantities in (3) are real.

3. THE SIGNAL SUBSPACE EIGENFILTER

In this section we briefly describe the signal subspace approach for speech enhancement presented in [2]. In this approach a linear estimator of the clean signal from the noisy observations is designed so as to minimize the signal distortion subject to forcing the residual noise to be below a desired threshold. The time domain constraint solution to this problem is given by

$$H = R_s(R_s + \mu\sigma^2 I)^{-1} \quad (4)$$

The parameter μ (Lagrange multiplier) controls the tradeoff between the residual noise and the signal distortion levels desired. Using (3), H can then be written as

$$H = U_1 G_\mu U_1^T \quad (5)$$

where the gain function G_μ is given by

$$G_\mu = \Lambda_s(\Lambda_s + \mu\sigma^2 I)^{-1} \quad (6)$$

The matrix U_1^T is called the KLT transform and its effect on the noisy signal vector $\mathbf{x}$ is to calculate the coefficients of its projection onto the signal subspace. These coefficients are modified using the gain function and finally the enhanced signal is reconstructed in the signal subspace using the inverse KLT, i.e. U_1. When we set $\mu = 0$ in (6) we obtain the Least Squares (LS) estimator $H = U_1 U_1^T$ which is just the projector of the noisy signal onto the signal subspace. This estimator does not result in any signal distortion but has the highest possible residual noise.

For better results using this filter, the dimension of the signal subspace, K (i.e the rank of R_s) should be accurately estimated. This process can be complex and it increases the computation cost of the whole filter. So usually K is set to some fixed value. Choosing K smaller than the actual subspace dimension results in a loss of some of the signal information like the formants, which can seriously affect the signal perception. In the case of consonants for example, K should be chosen very close to P for least signal distortion. Even trying to estimate K from the given observations can be neither easy nor accurate because it is hard to detect any gap in the eigenvalues of R_x.

4. THE MULTI-MICROPHONE APPROACH

The difference P-K represents the degrees of freedom for noise reduction offered by the signal model. So it would be desirable to get more degrees of freedom without introducing much distortion to the signal. In this section, we show how this can be done using a multi-microphone approach without much increasing the computational load.

Suppose we have M microphones for signal acquisition followed by a time delay compensation module to ensure that all microphone signals are correctly synchronized. Under these conditions, we have

$$\mathbf{y}_i = \mathbf{s} + \mathbf{w}_i \quad i = 1 \ldots M. \quad (7)$$

Now as in [6], we assume that the noise and reverberation form a diffuse acoustic field. Therefore these perturbations, in addition to being uncorrelated with the direct path signal, are considered to be incoherent at different microphones. These assumptions coincide with real life applications when the microphones are close to the speaker relative to the interfering sources like a car engine or air conditioning noise inside a room. Hence the covariance matrix of the input signal at two particular microphones with index i and j is given by

$$R_{ij} = E\{\mathbf{y}_i \mathbf{y}_j^T\} = R_s + \sigma^2 \delta(i-j) I \quad (8)$$

Now, define a combined vector Y of length $N = MP$, by stacking the individual input vectors of every microphone in the following way

$$Y = [\mathbf{y}_1^T, \ldots, \mathbf{y}_M^T]^T$$

Then the overall covariance matrix $R = E\{YY^T\}$ can be written as

$$\begin{aligned} R &= \begin{bmatrix} R_{11} & R_{12} & \cdots & R_{1M} \\ R_{21} & R_{22} & \cdots & R_{2M} \\ \vdots & & \ddots & \vdots \\ R_{M1} & \cdots & \cdots & R_{MM} \end{bmatrix} \\ &= \begin{bmatrix} R_s & \cdots & R_s \\ \vdots & \ddots & \vdots \\ R_s & \cdots & R_s \end{bmatrix} + \sigma^2 I_N \end{aligned} \quad (9)$$

where I_N is an $N \times N$ identity matrix.

Proposition 4.1. *Suppose that* $U = [\mathbf{u}_1^T, \ldots, \mathbf{u}_M^T]^T$*, where* $\mathbf{u}_i$*'s* $(i = 1, \ldots M)$ *are* P*-dimensional vectors, is a unit norm eigenvector of* R *with corresponding eigenvalue* λ*, where* $\lambda > \sigma^2$*, i.e. it is one of the first* P *eigenvalues of* R*. Then we have*

$$\mathbf{u}_1 = \mathbf{u}_2 = \ldots = \mathbf{u}_M \tag{10}$$

and $\mathbf{u} = \sqrt{M}\mathbf{u}_1$*, is a unit norm eigenvector of* R_s *with corresponding eigenvalue* $\frac{\lambda - \sigma^2}{M}$*.*

Proof. Since U is an eigenvector of R we have that $RU = \lambda U$, so

$$R_s \sum_{i=1}^{M} \mathbf{u}_i = (\lambda - \sigma^2)\mathbf{u}_j \quad j = 1, \ldots, M. \tag{11}$$

The left-hand side of (11) is constant for all j, and since $\lambda \neq \sigma^2$, then we have (10). Using this fact we have

$$R_s \mathbf{u}_1 = \frac{(\lambda - \sigma^2)}{M}\mathbf{u}_1 \tag{12}$$

So $\mathbf{u}_1$ and $(\lambda - \sigma^2)/M$ are an eigenvector and the corresponding eigenvalue of R_s, respectively. Now we need to find the unit norm eigenvector. We have

$$U^T U = \sum_{i=1}^{M} \mathbf{u}_i^T \mathbf{u}_i = M\mathbf{u}_1^T \mathbf{u}_1 = 1 \tag{13}$$

So $\mathbf{u} = \sqrt{M}\mathbf{u}_1$ is a unit norm eigenvector of R_s. □

To simplify the notation , let us define the $P \times N$ matrix C as

$$C = \frac{1}{M}[I_P, \ldots, I_P] \tag{14}$$

where I_P is a $P \times P$ identity matrix. For example the effect of C on U above is $CU = \frac{1}{M}\sum_{i=1}^{M} \mathbf{u}_i$.

So if the eigen-decomposition of R_s is given by $R_s = U_s \Lambda_s U_s^T$ and that of R by

$$R = [U_1 U_2] \begin{bmatrix} \Lambda_1 & \mathbf{0} \\ \mathbf{0} & \Lambda_2 \end{bmatrix} \begin{bmatrix} U_1^T \\ U_2^T \end{bmatrix} \tag{15}$$

where Λ_1 and U_1 are $P \times P$ and $N \times P$ matrices respectively. Then we can get an approximation for Λ_s as

$$\hat{\Lambda}_s = \frac{1}{M}(\Lambda_1 - \sigma^2 I_P) \tag{16}$$

and for U_s

$$\hat{U}_s = \sqrt{M} C U_1. \tag{17}$$

where in practice, Λ_1 and U_1 are obtained from the eigen-decomposition of the estimated covariance matrix. These quantities are then used in (6) to get a better approximation of the signal subspace basis without the need to choose or approximate a subspace dimension. So for every new input vector Y obtained from the M microphones, the enhanced vector is obtained as

$$\hat{\mathbf{s}} = \hat{U_s} G_\mu \hat{U_s}^T (CY). \tag{18}$$

Note that that in effect, the term (CY) above is a beamforming operation followed by the signal subspace post-filter.

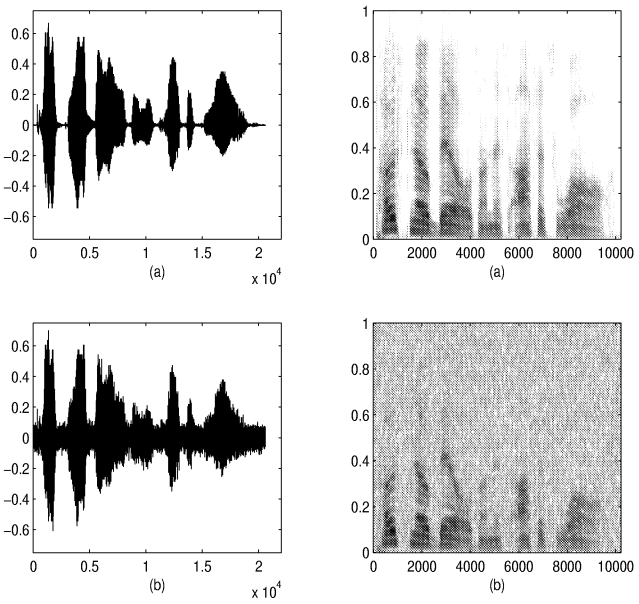

Fig. 1. The clean (a) and the noisy signal (b) at one of the microphones (SNR=10 dB) and their respective spectrograms.

5. IMPLEMENTATION AND EXPERIMENTAL RESULTS

In this section we describe how the algorithm presented above is implemented and and also discuss the evaluation of its performance in comparison with the single microphone version in [2]

To find the eigen-decomposition of the covariance matrices, the singular value decomposition of a $N \times 2N$ data matrix is used. The input vectors are speech frames of length N and with an overlap of $P/2$. In the single microphone approach P is chosen to be 80 and the signal model order $K = 40$. In the multi-microphone approach, $M = 4$ microphones are used and P is chosen to be 20 so that $N = MP = 80$. With these values the computation cost of the two methods are close to each other, since the singular value decomposition will be computed for matrices of the same size. The Lagrange multiplier $\mu = 2$ was chosen for both cases for least signal distortion.

In Figure 1, clean and noisy signal waveforms (and their spectrograms) for a male spoken test sentence (*"Post no*

bills on this office wall"), are shown. The noise is a computer generated white noise at 10 dB average SNR. Three methods for speech enhancement are tested; the conventional delay-and-sum beamformer (BF), the single microphone signal subspace method (SMSS) and the proposed multi-microphone signal subspace method (MMSS). Their time domain signal and spectrograms are shown in Figure 2. It can be seen that the BF and SMSS have slightly enhanced the signal quality. It can also be seen from the corresponding spectrogram that the SMSS method has introduced the musical noise artifact to the signal. In the bottom two plots of Figure 2, the time domain and correspondent spectrogram of the MMSS method are shown. It can easily be seen that the SNR improvement is significant and that not much residual noise is left. Listening tests support these results; listeners generally find that the MMSS enhanced speech signal has a better quality than with the other methods, and that no musical noise is perceived. (Demo files are available in our web site.)

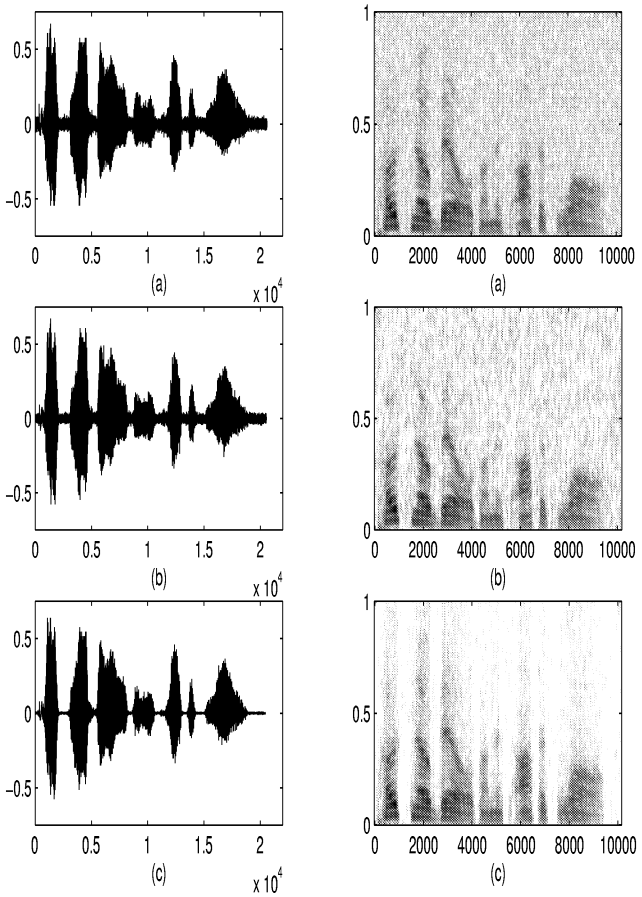

Fig. 2. Enhanced signals and their spectrograms using three different methods: (a) delay-and-sum beamformer, (b) the single microphone approach, (c) the proposed multi-microphone approach.

6. CONCLUSION

In this paper we presented a multi-microphone signal subspace approach for noise reduction. In this method, we use the data obtained from more than one microphone to estimate the clean signal subspace without relying on any signal model. Listening tests showed that, under comparable computational complexity, our method outperforms the original single microphone method. The musical noise artifact, from which the latter suffers, is significantly diminished while maintaining good intelligibility.

7. REFERENCES

[1] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 27, pp. 113–120, April 1979.

[2] Y. Ephraim and H. L. Van Tress, "A signal subspace approach for speech enhancement," *IEEE trans. on Speech and Audio Processing*, vol. 3, no. 4, pp. 251–266, July 1995.

[3] U Mittal and N. Phamdo, "Signal/Noise KLT based approach for enhancing speech degraded by colored noise," *IEEE Trans. on Speech and Audio Processing*, vol. 8, no. 2, pp. 159–167, March 2000.

[4] S. Gazor and A. Rezayee, "An adaptive subspace approach for speech enhancement," *Proc. IEEE ICASSP00*, pp. 1839 –1842, 2000.

[5] J.L. Flanagan, J.D. Johnston, R. Zahn, and G.W. Eiko, "Computer-steered microphone arrays for sound transduction in large rooms," *J. Acoust. Soc. Am.*, vol. 78, no. 5, pp. 1508–1518, November 1985.

[6] C. Marro, Y. Mahieux, and K. U. Simmer, "Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering," *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 3, pp. 240–259, May 1998.

[7] M. Dahl, I. Claesson, and S. Nordebo, "Simultaneous echo cancellation and car noise suppression employing a microphone array," *Proc. IEEE ICASSP97*, pp. 239–242, 1997.