MULTI-CHANNEL SPEECH SIGNAL SEPARATION BY EIGENDECOMPOSITION AND ITS APPLICTION TO CO-TALKER INTERFERNCE REMOVAL

Yuchang Cao, Sridha Sridharan, Miles P. Moody

Signal Processing Research Centre, School of Electrical and Electronic Systems Engineering
Queensland University of Technology, GPO Box 2434, Brisbane Q4001, Australia

ABSTRACT—This paper describes the concept of eigendecomposition for multi-channel signal separation, an alternative method of enhancing the desired signal corrupted by interference. The method uses two observations, which come from a pair of single sensors (or beamformers) both of which contain the desired signal and the undesired signal(s). The method assumes that the desired signal and the undesired signal(s) are uncorrelated and the signal-noise ratios (the ratio of the desired signal to the undesired signal(s)) of each observation are different. The technique has been successfully used to separate speech signal corrupted heavily by ambient noise, co-talker interference and other sources such as background music.

I. INTRODUCTION

Speech signal separation using a multiple microphone system has become an active research area for audio signal enhancement. For example, in a real life acoustic environment, the observations from multiple microphones contain the desired speech with some interfering audio signals which could come from television, radio, automobiles, competing speakers or other background noise sources. The aim is to retrieve the desired speech from the noisy observations. These problems occur, for example, in hands-free mobile phones, teleconferencing and forensic covert recording. Similar problems also occur in multi-sensor probing systems for underwater acoustics, diagnostic and geological survey.

In this paper we consider specifically the case of two-channel observations $y_1(t)$, $y_2(t)$ which are observed from a pair of sensors. To simplify the analysis the undesired signal is assumed to have only one source. The acoustic model is shown as Fig.1, where h_{11} represents the impulse response from the desired signal source $s_d(t)$ to the output $y_1(t)$, h_{22} represents the impulse response from the undesired signal source $s_u(t)$ to the output $y_2(t)$, and h_{12} and h_{21} represent the coupling effects between the channels. It is reasonable to assume that the acoustic system is a 2×2 linear time invariant (LTI) system. In the frequency domain we have

$$\begin{bmatrix} Y_1(\omega) \\ Y_2(\omega) \end{bmatrix} = \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} \begin{bmatrix} S_d(\omega) \\ S_u(\omega) \end{bmatrix} \qquad (1)$$

where H_{11}, H_{12}, H_{21} and H_{22} are the corresponding transfer functions.

The most widely used method of enhancing, or separating the desired signal from the two-channel observations was suggested by B. Widrow *et al.* [1]. Fig.2 shows his adaptive noise cancelling concept. A desired signal s_d is transmitted over a channel h_{11} to a sensor that also receives an undesired signal s_{u12} uncorrelated with s_{d11}. The combination of the received desired signal and the undesired signal y_1 (where $y_1 = s_{d11} + s_{u12}$) forms the primary input to the canceller. A second sensor receives an undesired signal or reference signal y_2 (where $y_2 = s_{u22}$ from source s_u transmitted over channel h_{22}) only, so the reference input is uncorrelated with the desired signal ($h_{21} = 0$). The reference input y_2 is filtered (the adaptive filter is controlled by the error signal e) to produce an output y that is as close a replica as possible of s_{u12}. This output is subtracted from the primary input y_1 to produce the system output $z = s_{d11} + s_{u12} - y$. This method will be referred to as the least squares (LS) method. Recursive and sequential/adaptive

schemes based on the least mean squares (LMS) and the recursive least squares (RLS) algorithms have been proposed in [1] and in, e.g., [2], respectively.

The least squares method has been applied in a wide variety of contexts. However, a critical assumption in this approach is that there is no leakage of the desired signal s_d into the reference sensor (i,e. $h_{21} = 0$). If both the desired signal s_d and the undesired signal s_u are coupled into each sensor, then with the least squares method, the desired signal will be partially cancelled together with the undesired signal. Note that for the method to be succesful, the noise component in the corrupted signal y_1 must be correlated with the component y_2. In many practical situations, this condition can be satisfied only by keeping the two sensors close together. For example, in hands free telephone applications using two microphones, the two microphones have to be kept around 5 cm apart to make the noise components in y_1 correlated to the components in y_2 [3]. This makes it almost impossible to prevent speech signal s_d from being included in the reference y_2. The least squares method fails to provide good performance in this case.

An approach to the two-channel signal separation problem, when both the desired signal and the undesired signal are coupled into each sensor, is presented in [4]. In that approach it is assumed that s_d is a Gaussian autoregressive (AR) process, and s_u is uncorrelated white Gaussian noise. The problem is formulated as a maximum likelihood (ML) estimation problem, and the iterative estimate-maximize (EM) algorithm is used for its solution.

Another approach to the signal separation problem is presented in [5]. This approach consists of reconstructing the input signals by assuming that they are statistically uncorrelated and imposing this constraint on the signal estimates. In order to restrict the set of solutions, additional information on the true signal generation and/or on the form of the coupling systems is incorporated.

Methods based on eigendecomposition have also been considered. In [6], the author wishes to establish independence of outputs and thus the zeroness of the off-diagonal elements of the covariance matrix. The eigen analysis technique is used to produce the intermediate variables $y_1'(t)$ and $y_2'(t)$. The next step is to find the rotation that provides the correct independence. This method is called "blind separation" and examples can be found in [7] and [8].

There are limitations when applying the methods proposed in [4], [5] and [6] to real acoustic scenarios because some critical assumptions made therein cannot be satisfied in many real situations. The major reason for the limitations is that the acoustic paths (including the acoustic fields and the recording systems) change the inherent relationships of the original speech signal. In other words, the real coupling functions are much more sophisticated than, for example, that assumed in [5] and [6], where some of the the coupling functions are constants and others are FIR filters with known order. In this paper, we present a new technique for noise cancellation based on eigendecomposition. The method has been successfully applied to separate speech signal corrupted heavily by ambient noise and co-talker inference and other noise sources such as background music.

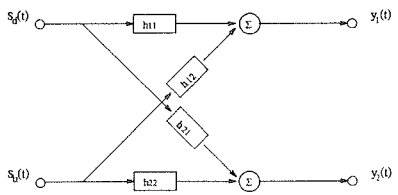

Fig.1 Model of two-channel observations

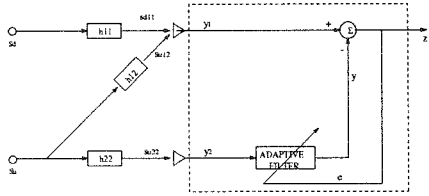

Fig.2, The Least Squares Method.

II. SIGNAL SEPARATION BY EIGENDECOMPOSITION

For simplicity, let us begin by considering two random sequences each consisting of complex exponentials. Let

$$s_i(n) = A_i u_i(n) \qquad for \quad i = 1, 2; \tag{2}$$

where

$$u_i(n) = e^{j\omega_i n} \tag{3}$$

and the complex amplitudes

$$A_i = \mid A_i \mid e^{j\phi_i} \qquad for \quad i = 1, 2; \tag{4}$$

For the acoustic modelling shown in Fig.1, we have

$$y_1 = s_1 \otimes h_{11} + s_2 \otimes h_{12} \tag{5}$$

and

$$y_2 = s_1 \otimes h_{21} + s_2 \otimes h_{22} \tag{6}$$

where the '$\otimes$' indicates linear convolution.

To simplify the analysis we first consider the case in which h_{11}, h_{12}, h_{21} and h_{22} can be expressed as

$$h_{ij} = \mid h_{ij} \mid e^{j\theta_{ij}} \qquad for \quad i, j = 1, 2; \tag{7}$$

where $\theta_{i,j}$ as well as ϕ_i are random variables with uniform distribution. Although this case is somewhat restrictive, it will show some important and interesting characterastics of signal subspace. The more general case will be considered in the sequel. Now, the observations y_i can be expressed as

$$y_1(n) = B_{11} u_1(n) + B_{12} u_2(n) \tag{8}$$

and

$$y_2(n) = B_{21} u_1(n) + B_{22} u_2(n) \tag{9}$$

where

$$B_{ij} = A_i h_{ij} = \mid A_i \mid\mid h_{ij} \mid e^{j(\phi_i + \theta_{ij})} \qquad for \quad i, j = 1, 2; \tag{10}$$

The correlation matrixes of the observed sequences are [9, 10]

$$R_{Y_i} = E\{Y_i Y_i^{*\mathrm{T}}\} = P_{i1} U_1 U_1^{*\mathrm{T}} + P_{i2} U_2 U_2^{*\mathrm{T}} \qquad for \quad i = 1, 2; \tag{11}$$

where Y_i and U_i are the data matrices which consist of N consecutive samples of the observations $y_i(n)$ and $u_i(n)$ respectively,

$$P_{ij} \stackrel{def}{=} E\{B_{ij} B_{ij}^*\} = E\{\mid B_{ij} \mid^2\} \qquad for \quad i, j = 1, 2; \tag{12}$$

and '$E\{\cdot\}$' stands for expectation, '$*$' denotes the complex conjugate, and '$*\mathrm{T}$' denotes the conjugate transpose, $u_1(n)$ and $u_2(n)$ have been assumed to be independent. Then, a new matrix R_{ratio} can be built from R_{Y_1} and R_{Y_2}

$$R_{ratio} = R_{Y_1}^{-1} R_{Y_2} = (P_{11} U_1 U_1^{*\mathrm{T}} + P_{12} U_2 U_2^{*\mathrm{T}})^{-1} (P_{21} U_1 U_1^{*\mathrm{T}} + P_{22} U_2 U_2^{*\mathrm{T}}) \tag{13}$$

The matrix R_{ratio} is called a ratio matrix of the two observed sequences. R_{Y_1} is invertible as it is a positive definite matrix.

Now consider the eigenvectors of matrices R_{Y_1}, R_{Y_2} and R_{ratio} . If $u_1(n)$ and $u_2(n)$ are statistically independent (as they are when $\omega_1 \neq \omega_2$) and zero-mean, we have

$$E\{s_1(t) s_2^*(t + \tau)\} = 0 \qquad \forall \tau$$

or

$$E\{u_1(t)u_2^*(t+\tau)\} = 0 \qquad \forall\tau \qquad (14)$$

as well as

$$E\{E\{u_i(t)u_i^*(t+\tau)\}u_j^*(\tau+\gamma)\} = 0 \qquad \forall\tau,\gamma, \quad for \quad i \neq j \qquad (15)$$

where $t = T_0 n$. When we perform an eigendecomposition of the matrixes R_{Y_1}, R_{Y_2}, it is possible to find the eigenvalues

$$\lambda_{i1} \quad \lambda_{i2} \quad \cdots \quad \lambda_{iM}$$

and the corresponding eigenvectors

$$v_{i1} \quad v_{i2} \quad \cdots \quad v_{iM} \quad for \quad i = 1, 2;$$

For each R_{Y_1} and R_{Y_2} there are two eigenvectors (although they are not normalized) which are the signal vectors u_1 and u_2, where

$$u_k = \left[1 \;\; e^{j\omega_k} \;\; \cdots \;\; e^{j(M-1)\omega_k}\right]^T \qquad for \quad k = 1, 2; \qquad (16)$$

It can be shown that u_1 and u_2 are also eigenvectors of R_{ratio}. For example, to show that u_1 is an eigenvector of R_{ratio}, we right-multiply (13) by u_1.

$$R_{ratio}u_1 = R_{Y_1}^{-1}R_{Y_2}u_1 = (P_{11}U_1U_1^{*T} + P_{12}U_2U_2^{*T})^{-1}(P_{21}U_1U_1^{*T} + P_{22}U_2U_2^{*T})u_1$$

$$= (P_{21}/P_{11})u_1 = \lambda_{ratio_1}u_1 \qquad (17)$$

where the third step follows from (14) or (15) and is based on (18), a formula for inverting matrices.

$$(A - CC^{*T})^{-1} = A^{-1} - A^{-1}C(I - C^{*T}A^{-1}C)^{-1}C^{*T}A^{-1} \qquad (18)$$

In the same way, we can show that

$$R_{ratio}u_2 = \lambda_{ratio_2}u_2 \qquad (19)$$

where $\lambda_{ratio_2} = (P_{22}/P_{12})$. From (17) and (19) we deduce that

- the signal subspace spanned by the eigenvectors of R_{ratio} is coincident with that spanned by the eigenvectors of R_{Y_1} and R_{Y_2};
- the eigenvalues of R_{ratio} are the ratios of the corresponding power densities (i.e., variances of the complex amplitudes) for each signal component of the two observations.

The above results can be used to separate the components $s_1(t)$ and $s_2(t)$ if the signal-to-noise ratios (one of s_1 and s_2 is assigned as signal and the other as noise) in the two observations are different or P_{21}/P_{11} is different from P_{22}/P_{12}. The signal separation can be achieved by building an eigen filter based on the eigenvectors of the matrix R_{ratio}.

III. APPLICATION TO SPEECH SIGNAL SEPARATION

The eigendecomposition analysis discussed in section II which is based on the matrix ratio $R_{Y_1}^{-1}R_{Y_2}$ of the autocorrelation matrices R_{Y_1} and R_{Y_2} at the two microphone inputs can be utilized to remove co-talker interference. Assume that frames of speech of the two talkers can be modelled as a linear combination of random exponential signals. This assumption can be justified for voiced segments. However, it is well known that unvoiced speech cannot be modeled as a linear combination of a set of random complex exponential signals. If we assume that unvoiced

segments can be modeled as band limited random white signals then the proposed analysis can be extended to the unvoiced signals as well.

Based on the eigendecomposition described above, an algorithm has been developed to separate multichannel signals with a pair of observation sequences. This algorithm, called the MRSS-I algorithm (MRSS stands for Matrix Ratio SubSpace) is as follows:

1. Segment the observation sequences and construct a pair of data matrices for each segment respectively.
2. Construct correlation matrices:

$$R_{Y_i} = Y_i Y_i^{\mathrm{T}} \qquad for \quad i = 1,\ 2; \tag{20}$$

3. Build a ratio matrix using

$$R_{ratio} = R_{Y_1}^{-1} R_{Y_2} \tag{21}$$

4. Compute the eigenvectors and eigenvalues by the decomposition of R_{ratio}. It is possible to rearrange the eigenvalues as (22).

$$\underbrace{\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_{p_1}}_{\textit{desired signal subspace}} \leq \underbrace{\lambda_{p_1+1} \leq \cdots \leq \lambda_{p_2}}_{\textit{noise subspace}} \leq \underbrace{\lambda_{p_2+1} \leq \lambda_{p_2+2} \leq \cdots \leq \lambda_p}_{\textit{undesired signal subspace}} \tag{22}$$

5. Find the zeros for each eigenvector corresponding to $\lambda_1,\ \lambda_2,\ \cdots,\ \lambda_{p_1}$, and the common zeros for all of them. If a set O_i represents the zeros for the ith eigenvector, this step is to find

$$O = O_1 \cap O_2 \cap \cdots \cap O_{p_1} \tag{23}$$

6. Build a filter to compress the frequency components around the zeros of the set O.

An illustration of performance for speech signal separation is demonstrated as follows: $s_1(t)$ considered as the desired signal and $s_2(t)$ as the undesired signal, are shown in Fig.3. After the coupling system, the average SNR (the ratio of average power of the desired signal to average power of the undesired signal) of the observation $y_1(t)$ is -0.57dB, and the average SNR of the another observation $y_2(t)$ is -6.59dB. The output of our multichannel signal separation system is $z(t)$ and the SNR of $z(t)$ is approximately 10.44 dB. $y_1(t)$ and $z(t)$ are shown in Fig.4.

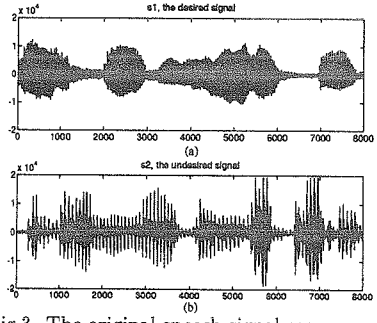

Fig.3. The original speech signal sequences.

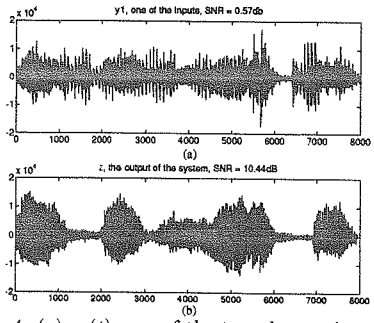

Fig.4: (a) $y_1(t)$, one of the two observation sequences which are the inputs of the signal separation system; (b) $z(t)$, the output of the system.

IV. CONCLUSIONS

We have described a novel method of multichannel signal separation based on eigenanalysis. The method assumes that two observations, both of which contain the desired and undesired signals, are available. This method can be applied directly to the two-sensor signal acquisition system. Multiple sensor systems can be converted to two observations by forming two beams and processing the outputs of the beamformers. The analysis is based on the eigendecomposition of a matrix R_{ratio} derived from the correlation matrices of the signals at the two sensors or beamformers. It has been shown that the ratio matrix has a set of eigenvalues which correspond to the power spectral density ratio of the corresponding frequency components in the observations $y_1(t)$ and $y_2(t)$. By choosing the proper eigenvalue(s) and the corresponding eigenvectors, the desired signal subspace can be separated from the undesired signal subspace as well as the noise subspace. The MRSS algorithms perform satisfactorily even if the desired signal is much weaker in both observations. By using two microphones and employing the proposed technique it is possible to significantly improve the SNR and the corresponding intelligibility.

ACKNOWLEDGEMENT

The authors acknowledge Dr. Dawei Huang for his comments on this research work. This work is supported by grants from the Australian National Institute of Forensic Science, the Queensland Police Service and a Queensland University of Technology Research Encouragement Award.

REFERENCE:

[1] B. Widrow *et al.*, "Adaptive Noise Cancelling: Principles and Applications", *Proc. IEEE*, Vol.63, No.12, pp.1692-1716, Dec. 1975

[2] S. Haykin, *Adaptive Filter Theory*, 2nd ed. New Jersey: Prentice-Hall 1991

[3] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs, NJ 1993

[4] M. Feder, A. Oppenheim, and E. Weinstein, "Maximum Likelihood Noise Cancellation Using the EM algorithm", *IEEE Trans. ASSP*, Vol.37, No.2, pp.204-216, Feb. 1989

[5] E. Weinstein, M. Feder, and A.V. Oppenheim, "Multi-Channel Signal Separation by Decorrelation", *IEEE Trans. on Speech and Audio Processing*, Vol. 1, No.4, pp405-413, Oct.1993

[6] R.E. Bogner, "Blind Separation of Sources", *Technical Report, Defence Research Agency*, Malvern, May 1992

[7] J. Cardoso, "An efficient Technique for the Blind separation of Complex Sources", In *IEEE Signal Processing Workshop on High-order Statistics*, Lake Tahoe, California, June 1993

[8] V.C. Soon, L. Tong, Y.F. Huang and R. Liu, "An Extended Fourth Order Blind Identification Algorithm in Spatially Correlated Noise", In *ICASSP'90*, pp1365-1367, Albuquerque, New Mexico, 1990

[9] C.W. Therrien, *Discrete Random Signals and Statitistical Signal Processing*, Prentice Hall, Englewood Cliffs, NJ, 1992

[10] J.G. Proakis, C.M. Rader, F. Ling and C.L. Nikias, *Advanced Digtal Signal Processing* Macmillan Publishing Co., NY, 1992

[11] J.Makhoul, "On the Eigenvectors of Symmetric Toeplitz Matrices", *IEEE Trans. ASSP*, Vol.29, No.4, pp868-872, August 1981