EXPLORING THE TIME-FREQUENCY MICROSTRUCTURE OF SPEECH FOR BLIND SOURCE SEPARATION

Hsiao-Chun Wu, Jose C. Principe and Dongxin Xu

Computational Neuro-Engineering Laboratory
Department of Electrical and Computer Engineering
University of Florida, Gainesville, FL 32611
E-mail: {wu, principe, xu} @cnel.ufl.edu

ABSTRACT

This paper explores the different frequency contents in short time segments (temporal microstructure) of speech to identify the mixing matrix in blind source separation. We propose a new method based on the eigenspread in different frequency bands to identify the segments which contain only one of the mixtures. It is much simpler to accurately estimate the mixing matrices from these segments. This short-time subband analysis trains very fast and estimates reliably the column vectors of the linear mixture. Simulation results show that our proposed method outperforms the existing model-based and competitive learning approaches in the identification of the mixing matrix for both sensor-sufficient (as many sensors as sources) and sensor-deficient (less sensors than sources) cases.

1. INTRODUCTION

The previous research in blind source separation of speech signals relied on model-based approaches with different types of assumptions such as: assumption on the underlying source distribution. [1, 2]; joint covariance and cumulant matrices diagonalization [3, 4]. The search for correlation matrices with distinct eigenvalues has been studied in [5] and applied using segmentation in time [6] or time frequency windowing [7].

Instead of the statistical model-based approach, non-parametric linear feature extraction applied to blind source separation has recently drawn some attention. The equi-partition method [9] and the local geometric approach [10] explore the intrinsic local structure of the source signals. However, although these two papers discovered the importance of energy disparity in the data, they still treated all data equally without emphasizing the data segments which provide the most important information about the underlying mixing matrix. In addition, exhaustive competitive learning was used and hence the mixing matrix could not be estimated very precisely all the time.

Neurophysiological evidence shows the important role played by the speech time-frequency microstructure in the way humans recognize and separate different voices (the cocktail party effect [8]). Therefore, in this paper we explore the short-time time-frequency structure to enhance the dominant speaker within each time-frame, which leads to the estimation of the independent component coordinate system. We show that sub-band filtering within each short-time window cleans the scatter plot of the sensed signals and makes the underlying features (independent component bases) easier to estimate; moreover by choosing the "most important features" instead of "averaging after clustering" as done in competitive learning, we may estimate mixing parameters much more precisely and much faster. In this paper we deal with blind source separation of several speech signals in a noise-free environment.

2. PROBLEM STATEMENT

The problem of blind separation of independent sources can be formulated as follows. A vector of m input signals $\boldsymbol{x}(t) \in \mathbb{R}^m$ is composed from a a vector of independent sources $\boldsymbol{s}(t) \in \mathbb{R}^n$ as

$$\boldsymbol{x}(t) = L_A[\boldsymbol{s}(t)] \tag{1}$$

where A parameterizes a family of invertible mixing maps L_A, and t denotes time. The problem is to reconstruct A (or find the inverse of A directly) and the original sources $\boldsymbol{s}(t)$ from the given input $\boldsymbol{x}(t)$.The sources $\boldsymbol{s}(t)$ can at best be reconstructed up to a permutation and scaling. We will assume all the signals to be zero-mean.

If we consider instantaneous mixtures only, we have $\boldsymbol{x}(t) = A\boldsymbol{s}(t)$, or equivalently,

$$x(t) = \sum_{i=1}^{n} a_i s_i(t) \qquad (2)$$

where A is a $m \times n$ matrix. If we would like to recover the signal, the number of sensor m should be larger than or equal to n. However, instead of the source recovery, it is possible to estimate the mixing matrix A even for $m < n$.

3. SOURCE SEPARATION BY FREQUENCY DECOMPOSITION

It is enlightening to analyze how we can use sub-band information for source separation instead of segmentation in time [6]. For simplicity, we consider a speech mixture with two sources and two sensors. The following properties are well known and easy to prove:

1. The convolutive operator and scalar multiplication are associative, i.e., $x_k(t) \otimes h(t) = (a_{k1}s_1 + a_{k2}s_2) \otimes h(t) = a_{k1}s_1 \otimes h(t) + a_{k2}s_2 \otimes h(t)$. where $x_k(t)$ is the k^{th} element of vector $x(t)$ and a_{kj} is the k^{th} element of vector a_j in Equation (2).
2. Assume two independent sources $s_i(t)$ and $s_j(t)$. If $y_i(t) = s_i(t) \otimes h(t)$, $z_j(t) = s_j(t) \otimes g(t)$ are formed by linear convolution with two different FIR filters the statistical average cross-correlation $E\{y_i(t)\, z_j(t)\} = 0$ since $E\{s_i(t-p)\, s_j(t-q)\} = 0$, $\forall i \neq j$ and $\forall p, q$.

From Property 1 and 2, we can obtain the following :

3. If we define $y(t)$ as $s(t) \otimes h(t)$ (each element $s_i(t)$ convolved by the same filter $h(t)$) and $E\{s(t)s^T(t)\}$ is diagonal, then $E\{y(t)\, y^T(t)\}$ is also diagonal.

Property 3 tells us that we may also achieve blind source separation by extracting different covariance matrices $E\{Ay(t)y^T(t)A^T\}$ in different subbands (i.e. through different filters $h(t)$). Since the unique solution occurs only when we have distinct source energy ratios ($E\{y_i^2(t)\}$ / $E\{s_i^2(t)\}$) [5], how to choose the subband filters still remains an open question.

Instead of the exhaustive search for subband filters, we propose a generalized eigenfilter method to choose the subbands as follows:

If $r_i(\tau)$ is the auto-correlation function of the i^{th} sensed signal, we can construct the correlation matrices,

$$R_i = \begin{bmatrix} r_i(0) & r_i(1) & \dots & r_i(L-1) \\ r_i(1) & r_i(0) & \dots & \dots \\ \dots & \dots & \dots & \dots \\ r_i(L-1) & \dots & \dots & r_i(0) \end{bmatrix} . \qquad (3)$$

The generalized eigenfilters are unit gain filters which maximize or minimize the Rayleigh quotient as,

$$\bar{h} = \arg\max(\arg\min) \frac{\bar{h} R_1 \bar{h}^T}{\bar{h} R_2 \bar{h}^T} \qquad (4)$$

where $\bar{h} = (h(0), \dots, h(L-1))^T$ is the eigenfilter. If we denote the mixing matrix A as $\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$., we obtain

$$\bar{h} R_i \bar{h}^T = \sum_k a_{ik} \int_{-\infty}^{\infty} |H(\omega)|^2 |S(\omega)|^2 d\omega \le \sum_k a_{ik} \int_{-\infty}^{\infty} |S(\omega)|^2 d\omega$$

since $h(t)$ has unit gain. (5)

Equation (5) shows that the eigenfiltering can only "*reduce*" the source energy by eliminating the information in some frequency bands.

The performance surface of the minimax problem in Equation (4) is hard to analyze and it is still under investigation. Alternatively we can use an extreme case of the surface plot to explain the optimization procedure. In Figure 1, we can see that the global optimal points occur in the two vertices with highest and lowest energy ratios. Hence the minimax procedure in Equation (4) will search the maximum or minimum ratios in the confined region by the unit gain filter $h(t)$.

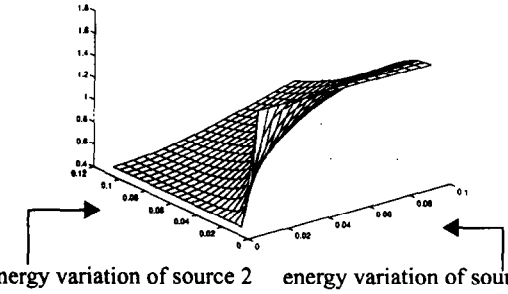

Figure 1. The ratio surface plot for Equation 4.

4. LOCAL GEOMETRICAL STRUCTURE IN TIME

We will discuss below the temporal information which enables the estimation of the underlying mixing matrix. If we consider a nonsingular mixing matrix A in Equation (2), the column vectors a_i's define an independent component coordinate system where the a_i's are in general not orthogonal. For the mixing model of Equation (2), the source signal amplitudes are the components of the input in the independent component coordinate system. We will call disparity the ratio of the amplitudes in the principal component system.

Since speech is composed of nonstationary segments (vowels, consonants and silence intervals) with different amplitudes, disparity over time will naturally occur in speech of different speakers. For instance, within some time-frame, one speaker articulates a vowel (with relatively higher energy) and others may articulates fricatives (with relatively lower energy) or be silent; during that time only one source s_i(t) has large amplitude swings, so the mixed signal will fall predominantly along a line parallel to a_i. We may use subband filters to enhance the energy ratio between the dominant source and the other sources. Figure 2a shows the scattering plot of two mixed speech signals, we can directly observe that the extreme points are nearly coincide with the underlying mixing vectors a_i's (denoted by the two lines in the graph).

5. PROCEDURE FOR ESTIMATING THE MIXING MATRIX

As depicted in Figure 3, our procedure to estimate the mixing matrix A (for two sensors) can be described as follows:

Time-frequency Decomposition:

1. Use short time analysis (sliding window of 20 msec).
2. Compute eigenfilters $h_i(t)$ and $h_j(t)$ ($h_k(t)$ and $h_l(t)$) to maximize and minimize Equation (4) and filter the mixed input through both subbands of each sensor (filter length 30 around taps).
3. Use PCA to estimate the spatial direction of the mixing vectors.

Feature estimation:

4. Running the frame analysis we collect a set of PCA directions from each subband and each sensor. We further cluster the PCA estimates in the number of expected sources using competitive learning. We then choose those estimates corresponding to the largest eigenspread in each cluster.
5. The separation matrix B is computed by $B = \hat{A}^{-1}$.

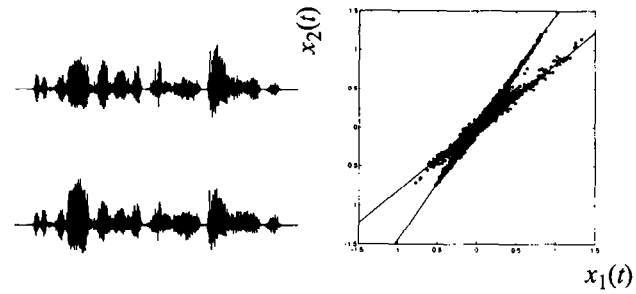

Figure 2a. The waveform (scatters) of two mixed signals

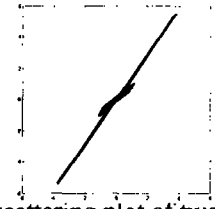

Figure 2b. The scattering plot of two mixed signals (through maximum generalized eigen filtering)

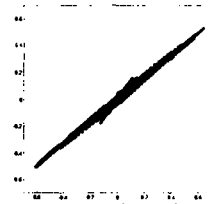

Figure 2c. The scattering plot of two mixed signals (through minimum generalized eigen filtering)

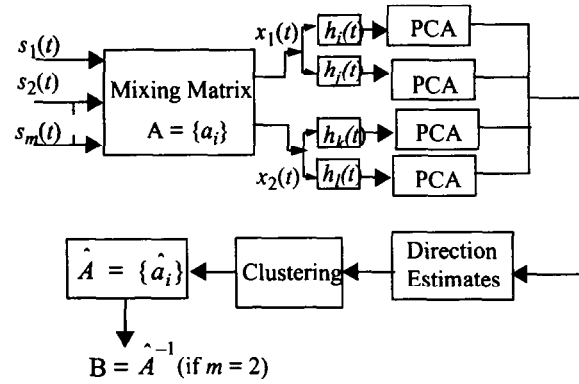

Figure 3. The estimation procedure for mixing matrix A

6. SIMULATION

6.1 The thining effect of subband filtering by generalized eigenfilters

In order to observe the "thining effect" as previously discussed in Section 3 we filter the mixed signals of Figure 2 by both generalized eigenfilters $h_1(t)$ and $h_2(t)$ (with

500 taps for illustration purposes). The corresponding scattering plot is depicted in Figure 2b and Figure 2c. They show that the appropriate subband filtering will enhance the identifiability of the mixing matrix.

6.2 Equal number of Sensors and Sources

The mixing matrix for data depicted in Figure 2a is

$$A = \begin{bmatrix} 0.9586 & 0.6801 \\ 0.7757 & 0.9797 \end{bmatrix} \quad (6)$$

First we follow the procedure in Section 5 to obtain a set of PCA estimates. We may either use a running average of eigenspread values to threshold those PCA estimates with a "large enough" eigenspread, or just sort the whole set of estimates to obtain "the best ones". In Figure 4 we can see that the estimates $\hat{a}_i$'s (denoted by the circles) are located at the directions of the true underlying vectors a_i's (their directions are denoted by the lines). After clustering and using the "best estimates" (with largest eigen-spreads) the final separation-mixing matrix product is

$$BA = \begin{bmatrix} -0.0066 & 1.1930 \\ -1.2381 & 0.0003 \end{bmatrix} \quad (7)$$

and the average SNR is 58.12 dB. Compared to the existing gradient and numeric algorithms, the performance is remarkable.

6.3 More sources than sensors

With the two sensors, if we consider more sources in the linear mixture (for instance, four sources)

$$A = \begin{bmatrix} 0.9501 & 0.6068 & 0.8913 & 0.4565 \\ 0.2311 & 0.4860 & 0.7621 & 0.0185 \end{bmatrix}, \quad (8)$$

we may still identify all the four mixing vectors a_i's . The result is shown in Figure 5, where circles denote the estimated directions and lines denote the true directions. The four angular errors lie between 0.0583 ~ 0.6575 degrees.

7. CONCLUSION

We explored the time-frequency structure of speech for source separation. The angular estimation of mixing vectors is strikingly precise. However the choice of subband filters still remains an open area for research.

Acknowledgment: This work was partially supported by ONR N00014-94-1-0858 and NSF ECS-9510715.

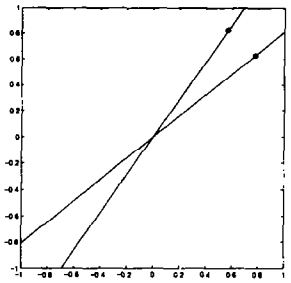

Figure 4. The estimated directions and true mixing vectors for 2-sensor-2-source case.

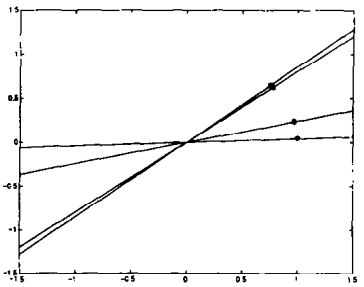

Figure 5. The estimated directions and true mixing vectors for 2-sensor-4-source case.

REFERENCE

[1] A. J. Bell and T. J. Sejnowski, "An information maximization approach to blind separation and blind deconvolution," *Neural Computation*, 7, pp. 1129-1159, 1995.

[2] J.-F. Cardoso and B. Laheld, "Equivariant adaptive source separation," *IEEE Transactions on Signal Processing*, vol. 44, no. 12, December 1996.

[3] H.-C. Wu and J. C. Principe, "A unifying criterion for blind source separation and decorrelation: simultaneous diagonalization of correlation matrices. in *Proc. NNSP*, pp. 496 -505, 1997.

[4] D. Obradovic and G. Deco, "Unsupervised learning for blind source separation: an information-theoretic approach," in *Proc. ICASSP*, pp. 127-130, 1997.

[5] L. Tong, R. Liu, V. Soon and Y. Huang, "Indeterminacy and identifiability of blind identification," *IEEE Trans. on Circuits and Systems*, vol. 38, no. 5, pp. 499-509, May 1991.

[6] A. Souloumiac, "Blind source detection and separation using second order nonstationarity," in *Proc. ICASSP*, pp. 1912-1915, 1995.

[7] A. Belouchrani and Moeness, "Blind source separation using time-frequency distributions: algorithm and asymptotic performance," in *Proc. ICASSP*, pp. 3469-3472, 1997.

[8] J. Blauert, *Spatial theory: the psycho-physics of human sound localization*, MIT Press, 1983.

[9] J. K. Lin, D. G. Grier and J. D. Cowan, "Faithful representation of separable distributions," *Neural Computation*, **9**. pp. 1303-1318, 1997.

[10] J. K. Lin, D. G. Grier and J. D. Cowan, "Feature extraction approach to blind source separation," in *Proc. NNSP*, pp. 398-405, 1997.