A METHOD FOR DIRECTIONALLY-DISJOINT SOURCE SEPARATION IN CONVOLUTIVE ENVIRONMENT

Shlomo Dubnov

CRCA
University of California
San Diego
sdubnov@ucsd.edu

Joseph Tabrikian, Miki Arnon-Targan

Dept. of ECE
Ben Gurion University of the Negev
Beer Sheva, Israel
{joseph,arnontar@ee.bgu.ac.il}

ABSTRACT

In this paper we propose a new method for source separation that is based on directionally-disjoint estimation of the transfer functions between microphones and sources at different frequencies and at multiple times. The directions are estimated from eigen-vectors of the microphones correlation matrix. Smoothing and association of transfer function parameters across different frequencies is achieved by simultaneous Kalman filtering of the noisy amplitude and phase estimates. This approach allows estimating transfer functions even in the case where the difference between the sources is in delay only and it can operate both for wideband and narrowband sources. Simulation results show superior performance in comparison to other existing methods.

1. INTRODUCTION

Many audio communication and entertainment applications deal with acoustic signals that contain combinations of several acoustic sources in a mixture that overlaps in time and frequency. In recent years there has been a growing interest in methods that are capable to separate audio signal from microphone arrays using Blind Signal Separation (BSS) techniques [1]. In contrast to most of the research works in BSS that assume multiple microphones, in most practical situations the audio data is limited to stereo recordings. Moreover, the majority of the potential applications of BSS in the audio realm consider separation of simultaneous audio sources in reverberating or echoing environments, such as a room or inside a vehicle. These applications deal with convolutive mixtures [2] that often contain long impulse responses that are difficult to estimate or invert.

In this paper we consider a simpler but still practical and largely overlooked situation where the mixture contains a combination of source signals occuring in relatively non-reverberant environment, such as speech or music recorded with close microphones. The main mixing effect in such a case is the delay effect and possibly a small combination of delays that can be described by a convolution with a relatively short impulse response. Recently, several works proposed separation of multiple signals when the signals are disjoint in the time-frequency(TF), [3]-[4], usually called W-disjoint, i.e. each source occupies separate regions in Short Time Fourier Transform (STFT) representation. In such a case the amplitude and delay estimation of the mixing parameters of each source is possible from the ratio of the STFT's of signals from the two microphones. Moreover, the W-disjoint or approximately W-disjoint situation allows estimation of more sources than microphones. Since the disjoint assumption appears to be too strict for many real-world situations, several improvements have been reported that allow only an approximate disjoint situation. The basic idea in such a case is to use some sort of a detection function that allows to determine the TF areas where each source occurs alone and use these areas only for separation.

2. THE MODEL

In blind source separtaion an N-channel sensor signal, $\mathbf{x}(t)$, arises from M unknown scalar source signals $s_m(t)$, linearly mixed together by an unknown $N \times M$ matrix $\mathbf{A}$, corrupted by a zero-mean, white additive noise $\mathbf{v}(t)$.

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{v}(t)$$

This model has been extensively investigated in the literature. In a convolutive environment, the signals arrive at the array after delays and reflections. We consider the case where each one of the sources is placed at a different location thus having a different tempo-spatial transfer function. Therefore, the signal at the

nth microphone is given by

$$x_n(t) = \sum_{m=1}^{M}\sum_{l=1}^{L} a_{nml}s_m(t-\tau_{nml})+v_n(t)\ ,\ n=1,\ldots,N \tag{1}$$

where τ_{nml} is the delay in the lth path of the speaker signal m received at microphone n. STFT of the (1) gives

$$X_n(t,\omega) = \sum_{m=1}^{M} A_{nm}(\omega)S_m(t,\omega)+V_n(t,\omega)\ ,\ n=1,\ldots,N \tag{2}$$

where $S_m(t,\omega)$ and $V_{t,n}(\omega)$ are the STFT of $s_m(t)$ and $v_n(t)$, respectively, and the temporal transfer function of the mth signal to the sensor n is defined as

$$A_{nm}(\omega) = \sum_{l=1}^{L} a_{nml}e^{-j\omega\tau_{nml}} \tag{3}$$

In matrix notation, the model in (2) can be written in the form:

$$\mathbf{X}(t,\omega) = \mathbf{A}(\omega)\mathbf{S}(t,\omega) + \mathbf{V}(t,\omega) \tag{4}$$

Our goal here is to estimate the signal vector $\mathbf{s}(t)$ from the measurement vector $\mathbf{x}(t)$ where the tempo-spatial transfer function matrix $\mathbf{A}(\omega)$ is unknown. Our solution does not require that the number of sensors be greater or equal than the number of sources, i.e. M may be greater or equal to N.

3. THE PROPOSED SOURCE SEPARATION METHOD

The proposed approach seeks for time-frequency cells in which only one source is present. At these cells it is possible to estimate the unstructured spatial transfer function for each frequency. Therefore, the first task is to identify the single source cells, and calculate the spatial transfer functions for those cells. In the second stage, the estimated spatial transfer functions are clustered and tracked via a Gaussian Mixture Model (GMM) and Kalman filter as decsibed in the next Section.

The autocorrelation matrix at a given time-frequency cell is given by

$$\mathbf{R}_x(t,\omega) = \mathbf{A}\mathbf{R}_s(t,\omega)\mathbf{A}^H + \mathbf{R}_v(t,\omega) \tag{5}$$

where $\mathbf{R}_x$, $\mathbf{R}_s$ and $\mathbf{R}_v$ are the correlation matrices of the measuements, source signals and noise, respectively. After averaging over time windows in which the signal can be considered stationary, we obtain:

$$\mathbf{R}_x(\omega) = \mathbf{A}(\boldsymbol{\omega})\mathbf{R}_s(\omega)\mathbf{A}(\boldsymbol{\omega})^H + \mathbf{R}_v(\omega) \tag{6}$$

Here we assume that the noise covariance matrix is known. Therefore, Eq. (6) can be prewhitened by left mutipying by $\mathbf{R}_v^{-1/2}(\omega)$. Since the spatial transfer function matrix is unknown and unstructured, then with no loss of generality, we can assume that the matrix $\mathbf{R}_v(\omega)$ is diagonal: $\mathbf{R}_v(\omega) = \sigma_v^2\mathbf{I}_N,\ \forall\omega$ where $\mathbf{I}_N$ is an identity matrix of size N.

For each time-frequency window for which (6) is estimated, we first test whether there exist only a single signal. In cases of a single signal, the spatial transfer function is estimated. For simplicity of notations we drop the dependence on time-frequency. Let $\lambda_1 \geqslant \cdots \geqslant \lambda_N$ denote the eigenvalues of $\mathbf{R}_x$. In the presence of a single source, $rank(\mathbf{A}\mathbf{R}_s\mathbf{A}^H) = 1$, and therefore, $\lambda_1 = \sigma_s^2 + \sigma_v^2$ and $\lambda_2 = \lambda_3 \cdots = \lambda_N = \sigma_v^2$, where σ_s^2 denotes the signal power. Therefore, in order to identify the case of a single source, the following test is performed:

$$T = \frac{\lambda_1}{\frac{1}{N-1}\sum_{n=2}^{N}\lambda_n} \underset{H_0}{\overset{H_1}{\gtrless}} \gamma\ . \tag{7}$$

Under the hypothesis of a single source, $T = SNR+1$, while in cases of no source, or more than one source we obtain $T < SNR+1$. Alternatively, model order selection methods, such as MDL or AIC, can be performed in order to find the number of the sources present in the time-frequency cell.

If only the mth speaker is present, equation (6) becomes

$$\mathbf{R_x}^m(\omega) = \mathbf{a}_m(\omega)\mathbf{a}_m^H(\omega)\sigma_{s_m}^2 + \sigma_v^2\mathbf{I} \tag{8}$$

where $\mathbf{a}_m(\omega)$ is the mth column of the matrix $\mathbf{A}(\omega)$. Therefore, $\mathbf{a}_m(\omega)$ is proportional to the eigenvector of the autocorrelation matrix $\mathbf{R_x}^m$ associated with the maximum eigenvalue: $\lambda_1^m = \sigma_{s_m}^2 + \sigma_v^2$.

4. TRACKING AND FREQUENCY ASSOCIATION ALGORITHM

A common problem in convolutive blind source separation is that the mixing parameters estimation is performed separately for every frequency. In order to reconstruct the time signal, the separated channels in frequency must be combined together in a consistent manner, i.e. one must provide that different frequency components correspond to same source[1]. Since multiple estimates of amplitude ratios are available at each frequency, we find the means of their clusters by estimating parameters of a GMM. The association of the mixing parameters across frequencies is performed by operating separate Kalman filters for every source.

[1]This problem is also sometimes known as frequency permutation or association problem

4.1. GMM and Kalman filters

The GMM assumes that the observations, z are distributed according to the following density function

$$p(z) = \sum_{m=1}^{M} \pi_m N(z|\Theta_m) \ , \tag{9}$$

where π_m's are the relative weights of the Gaussian distributions $N(\cdot|\Theta_m)$ and $\Theta_m = \{\mu_m, \Sigma_m\}$ are the mean and the covariance matrix parameters of the Gaussians. In our case, the observations, z, are the estimates of transfer function at every frequency (see previous section). The parameters of the GMM are obtained using an EM procedure. The resulting mean values are the elements of the estimated transfer function vector $[\mathbf{a}_m(\omega = k)]$ where k is the frequency index are input into a Kalman filter. The state vector $\mathbf{s}[k]$ of the Kalman filter consists of a two dimensional space of {mean values, derivative (speed)}. The dynamics across neighboring frequencies (frequency smoothness constraint) were modeled as

$$\begin{aligned} \mathbf{s}[k] &= \mathbf{T}\mathbf{s}[k-1] + \mathbf{w}[k] \\ \mu[k] &= \mathbf{C}\mathbf{s}[k] + \mathbf{u}[k] \end{aligned} \tag{10}$$

with frequency transition matrix $\mathbf{T} = [1\ 1\ ;\ 0\ 1]$, and observations vector $\mathbf{C} = [1\ 0]$. The vectors $\mathbf{u}, \mathbf{w}$ represent the observation noise and model errors, respectively.

4.2. The separation algorithm

The various steps of the algorithm can be summarized as follows:

- Given a two channel recording perform a separate STFT analysis for every channel, resulting in signal model of equation (4).
- Perform an eigenvalue analysis of the cross-channels correlation matrix at every frequency, as described in Section 3, equations (6) and (7) and determine the transfer function.
- At every frequency determine cluster centers of the set of amplitude ratio measurements using GMM.
- Perform Kalman tracking of the cluster means across frequencies for each source to obtain an estimate of the mixing matrix as a function of frequency.
- Perform an "un-mixing" of the sources by multiplying the STFT channels at each frequency by an appropriate inverse matrix.
- Perform an inverse STFT using associated frequencies for each of the sources.

Since the mixing matrix can be determined only up to a scaling factor, we assume a unit relative magnitude for one of the sources and use the amplitude ratios to determine the mixing parameters of the remaining source[2].

5. EXPERIMENTAL RESULTS

Separation experiments were held for simulated mixing conditions at different geometrical setups, such as varying source locations, relative amplitudes of the sources, angles and amplitudes of the multipath reflections, the relative distance between the microphones and different types of sound sources. Figure 1 shows the measured vs. smoothed spatial transfer functions for a difficult case of two female speaker sources with equal amplitude mixing conditions. The separation is possible due to different phase behavior of the signals, which is properly detected using the Kalman tracking.

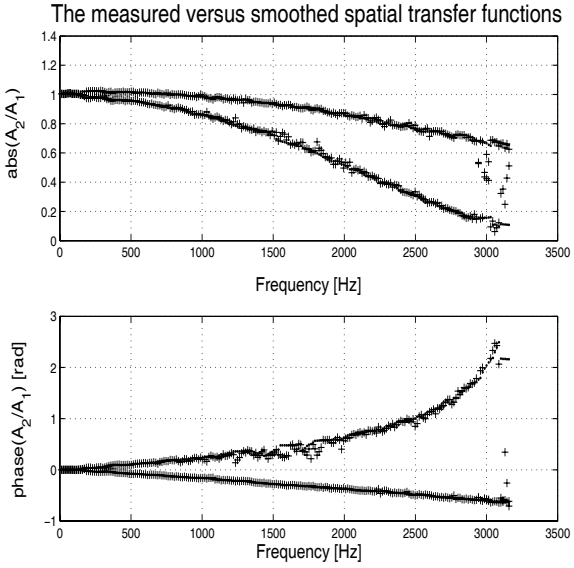

Figure 1: Amplitude and phase of the measured and smoothed transfer functions.

Figure 2 shows the results of an experiment where we estimated an improvement in SNR for different relative positions of the sources with different relative amplitudes and including multipath reflections. One of the sources was held constant at angle 0 while the

[2]This problem of scale invariance may cause a "coloration" of the recovered signal and might be one of the possible sources of error. This problem is common to many convolutional source separation methods

other source was shifted relatively to it from −40 to 40 degrees. The relative amplitudes of the sources varied from 0.5 to equal amplitude ratios. The multipath reflections occurred at constant angles of 60 and −40 degrees with relative amplitudes of few percent of the original.

For equal amplitudes, we achieve up to 10dB improvement when the sources are 40 degrees apart. The angle sensitivity disappears when sufficient amplitude difference exists between the sources. For amplitude ratio 0.5 (i.e. each microphone receives its main source at amplitude 1 and the interfering source at amplitude 0.5) we achieve between 20 and 30 dB improvement. One should note that the above results contain weak multipath components. Even better improvement (50 dB or more) can be achieved for cases when no multipath is present. Another difficulty with multipath is that it might cause a failure in tracking in the direction of the multipath.

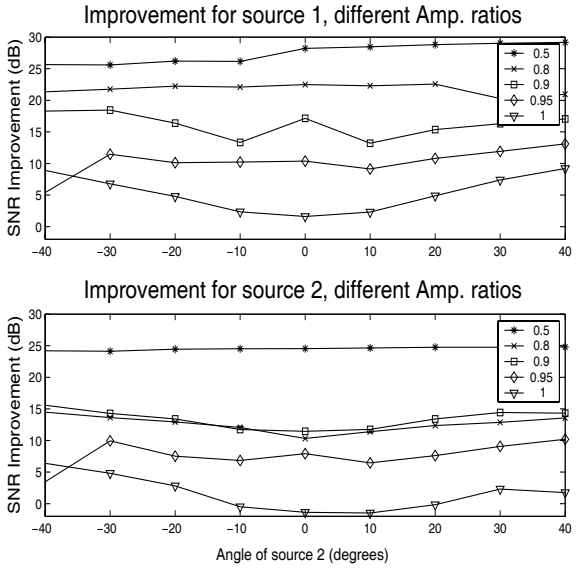

Figure 2: Improvement in snr for different relative positions of the sources and different relative amplitudes. See text for more details.

Figure 3 shows the amplitude of a spatial transfer function of the inverse mixing matrix for each frequency, for the case of two sources, one around the center and the other source around 60 degrees, with no multipath. One can see that the inverse matrix puts a notch at the direction of every source, separately for each microphone.

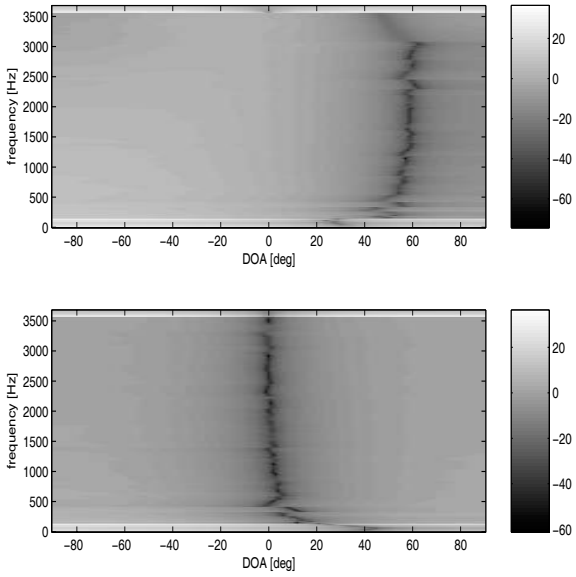

Figure 3: Spatial pattern obtained by of the inverse of the mixing matrix for each frequency in case of two sources at 0° and 60°.

Acknowledgment

This work was partially supported by the Israeli Science Foundation (ISF).

6. REFERENCES

[1] K. Torkkola, "Blind separation for audio signals – are we there yet," Proc. 1st Int. Workshop Indep. Compon. Anal. Signal Sep., Aussois, France, pp. 239-244, Jan. 1999.

[2] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Transactions Speech and Audio Processing*, pp. 320–7, 2000.

[3] A. Jourjine, S. Rickard, and O. Yilmaz, "Blind separation of disjoint orthogonal signals: Demixing n sources from 2 mixtures," In Proc. ICASSP, June 5-9, Istanbul, Turkey, June 2000.

[4] D. L. Wang N. Roman and G. J. Brown, "Speech segregation based on sound localization," *J. Acoust. Soc. Am.*, vol. 114, 2003.

[5] Y. Deville, "Temporal and time-frequency correlation based blind source signal separation methods," In Proc. 4th Int. Workshop Indep. Compon. Anal. and Blind Sig. Sep., April, Nara, Japan, 2003.