USING BEAMFORMING IN THE AUDIO SOURCE SEPARATION PROBLEM

Nikolaos Mitianoudis and Michael E. Davies

DSP Group, Queen Mary
London E1 4NS, UK
nikolaos.mitianoudis@elec.qmul.ac.uk

ABSTRACT

The problem of separating audio sources observed in a real room environment is a very challenging task, also known as the *cocktail party problem*. Much work has been presented on audio separation, even in cases of high reverb. However, various problems remain unsolved in a real-world scenario. In this paper, the authors review proposed solutions employing *Independent Component analysis* (ICA), discussing possible solutions to various problems that arise during the analysis (i.e. the *permutation* problem). In particular, the use of *beamforming* techniques in parallel with the ICA framework is discussed. Finally, some of the open problems in audio source separation are considered.

1. INTRODUCTION

A real-world *Blind Source Separation* problem is described as follows. Assume there are N audio sources in a room $\underline{s}(n) = [s_1(n), \ldots, s_N(n)]^T$ and the auditory scene is captured by M sensors $\underline{x}(n) = [x_1(n), \ldots, x_M(n)]^T$. The common approach to model the mixing procedure is the following:

$$x_i(n) = \sum_{j=1}^{N} \alpha_{ij}(n) * s_j(n) \qquad i = 1, \ldots, M \qquad (1)$$

where α_{ij} is a FIR filter that models the *room transfer function* between the i^{th} sensor and the j^{th} source. This is also refered to as *convolutive mixtures* model. Assuming equal number of sources and sensors ($N = M$) and no additive noise, one could solve the source separation problem, by estimating the unmixing FIR filters w_{ij}, whenever that is possible.

$$u_i(n) = \sum_{j=1}^{N} w_{ij}(n) * x_j(n) \qquad i = 1, \ldots, N \qquad (2)$$

An efficient way to perform the unmixing is within a *subband* (eg. *frequency*) domain, where the convolution can be modelled approximately as multiplication and therefore any *Independent Component Analysis* (ICA) algorithm for *instantaneous mixtures* can be employed [5].

One approach is to model the sources and work solely in the frequency domain [9]. The benefits are that we are working with a sparser representation in the frequency domain that enables better separation [5]. On the other hand, we assume no statistical dependence between the frequency bins that introduces the *permutation problem* of ordering the separated sources along the frequency axis. However, unmixing in a subband domain does not necessarily imply that we should use a frequency domain source model. Lee et al [4] imposed the source model in the time-domain and preformed the unmixing in the frequency domain. The advantage is that the permutation problem does not exist in this case. Nonetheless, it is more computationally expensive due to the repetitive mappings to and from the frequency domain.

2. SOLUTIONS FOR THE PERMUTATION PROBLEM

Current solutions for the permutation problem in the Frequency Domain ICA (FD-ICA) framework can be categorized into two basic groups:

2.1. Source modeling Solutions

In *source modeling* solutions, the aim is to exploit the coherence and the information between frequency bands in order to impose frequency coupling between the subbands and therefore alignment of the sources after separation.

A possible approach is to impose time-frequency source models, as proposed by Ikeda [1], Mitianoudis and Davies [5]. Ikeda used signal envelopes in the time-frequency representation to impose coupling, whereas Mitianoudis and Davies proposed a generative time-frequency model along with a likelihood ratio jump solution in order to force subband coupling and align the sources after unmixing (*flipping solutions*).

2.2. Channel modeling Solutions

In channel modeling solutions, the aim is to model the transfer functions in order to couple the unmixing filters and therefore align the sources.

One approach is to assume *smooth* filters, as a contraint to the unmixing algorithm. Smaragdis [9] used a heuristic approach to achieve that. Parra and Spence [7] aligned permutations using a constrained filter model, which has been reported to get trapped in local minima [2]. Both approaches can be characterized as *gradient solutions*, as the model is incorporated in the gradient algorithm.

Another approach is to consider the BSS setup as a N-sensor beamformer and employ its directivity pattern to resolve the permutations, as investigated by Saruwatari et al [8], Ikram and Morgan [3], Parra and Alvino [6]. We will analyse the application of beamforming in the BSS concept in detail in the following section.

3. ICA AS A BEAMFORMER

One interpretation of the FD-BSS setup is a null-steering FD-Beamformer that uses a blind algorithm (ICA) to place nulls to the other sources present in the mixture, in order to separate one at a time. However, BSS does not utilize any information concerning the geometry of the auditory scene (position of sources and sensors, i.e Directions Of Arrival (DOA) of source signals to the array). One can use the sources' DOA to align the permutations along the frequency axis (flipping solution). We have to permute the sources along frequency, so that the *directivity pattern* of each beamformer is aligned. The directivity pattern is defined as follows:

$$F_i(f,\theta) = \sum_{k=1}^{N} W_{ik}^{phase}(f) e^{j2\pi f d_k \sin\theta_i / c} \qquad (3)$$

where $W_{ik}^{phase}(f) = W_{ik}(f)/|W_{ik}(f)|$ is the phase of the unmixing filter coefficient between the k^{th} sensor and the i^{th} source at frequency f, d_k is the distance of the k^{th} sensor from the origin, θ is the DOA of the i^{th} source and c is the velocity of sound. An ideal directivity pattern for a single delay system along frequency is depicted in figure 1, where we can clearly see a null at 25^o, which is the actual DOA of the source. Essentially, the single delay case approximates an anechoic room. However, the directivity patterns that describe real room transfer functions are not so smooth.

Applying beamforming in the FD-ICA framework is clearly a *channel modeling* solution. In other words, it is an attempt to exploit the *phase information* of the unmixing system.

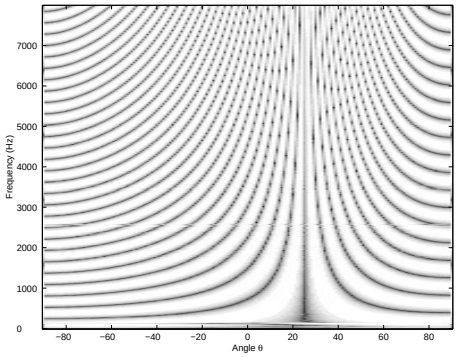

Figure 1: Frequency Dependent directivity pattern of a single-delay transfer function.

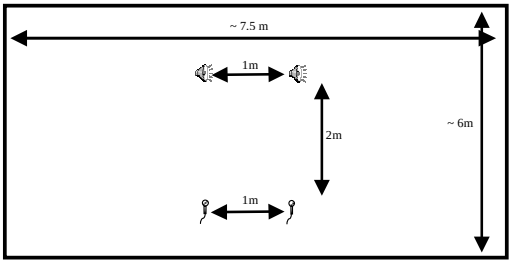

Figure 2: Experimental setup in a real room.

4. A REAL WORLD EXPERIMENT

In this section, we discuss the use of beamforming for permutation alignment through a real world experiment. We used a university lecture room to record a 2 sources - 2 sensors experiment. We used two speakers (source 1 and source 2) and two cardiod microphones (mic 1 and mic 2), arranged as in figure 2. We used the approach described in [5] to separate the sources, using the source modeling plus *Likelihood Ratio Jump* to tackle the permutation problem. Then, we tested the beamforming performance of the estimated filters along frequency, making some interesting observations.

4.1. General Observations

In figure 3, we can see the beamforming pattern of the estimated unmixing filter between source 1 and microphone 1. The first observation is that the beamforming pattern is more smeared compared to the single delay beamforming plot. This is due to the room's complex transfer function, that slightly shift the sources' DOAs along frequency. However, we can still spot a main direction of arrival that can help align the permutations. This is an attempt to approximate the transfer function with a single delay.

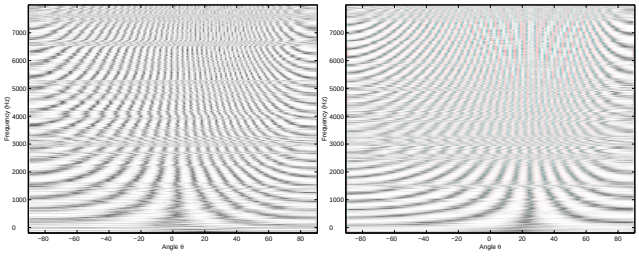

Figure 3: Beamforming pattern of the estimated unmixing filters.

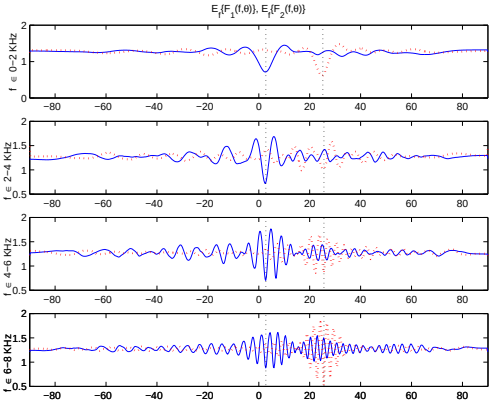

Figure 4: Average Beampatterns along certain frequency bands for both sources.

Looking at these beamforming plots, it seems that aligning the permutations around the DOAs, using only beamforming information, is difficult. The main reason being that after $f_0 \approx c/d$, we start getting multiple nulls, due to wavelength restrictions. Saruwatari et al [8] used the overall statistics of the nulls to get DOA estimates. However, Ikram and Morgan [3] proposed to use only lower frequencies to get DOA estimates (see figure 4) and then align permutations, by looking for nulls in the neighbourhood of the estimated DOA. Parra and Alvino [6] used more sensors than sources to get more accurate estimates for the DOAs with a subspace technique and add this information as a geometric constraint to their unmixing algorithm.

However, it seems to be rather difficult to perform permutation alignment in higher frequencies, even with very accurate estimates for the DOA. In figure 4, we can see that it is difficult to define the "neighbourhood" around the DOA, as the nulls corresponding to the both sources are really close and the probability of error is high. Therefore, one solution can be to rely mainly on amplitude only criteria for mid-higher frequency band in order to sort the permutations. On the other hand, phase information (beamforming) can be useful for sorting out the lower frequency band permutations.

Equally important is the choice of the sensor spacing d_k in eq 3. It is obvious that choosing smaller spacing will reduce the multiple nulls at mid-high frequencies. In contrast, the *Signal-to-Noise Ratio* will decrease, as the signals captured by the microphones will be similar, due to the far-field effect. Therefore, the choice of sensor spacing is a tradeoff between *separation quality* and *beamforming pattern clarity*.

4.2. Sensitivity Analysis

In this section, we discuss the sensitivity of our system to movement. We repeated two new recordings with source 2 moved $20cm$ and $50cm$ to the left. We unmixed the sources in either case and observed the following:

4.2.1. Beamformer's sensitivity to movement

Comparing the beampatterns estimated for either case, we observed that the beamformer's sensitivity to movement is a function of frequency. We can clearly see that in lower frequencies, beamformer's null has been slightly shifted, however in higher frequencies the shift is a function of frequency. This is to show that if we have a moving source in our source separation problem, a very small change will not greatly affect our beamformer in lower frequencies. However, our beamformer can be rendered useless in higher frequencies in cases of small movements (figure 5). This can be supported by the way humans perform *source localization*. Many psychologists observed that the human ear tends to localize lower frequency sounds by phase difference and higher frequency sounds by amplitude difference.

4.2.2. Distortion introduced due to movement

Using the filters estimated for the original speakers position to unmix the two new recordings, we explore how tolerant the system is to movement. We observed that source 2 was separated from the mixture, however it sounded more "echoic". On the other hand, source 1 contained more crosstalk, but no added distortion. As BSS is a null-steering procedure, source 2 will have no contamination from the other source, as the sensors will place a correct null to source 1. However, because we are mapping back to the microphones to remove the *scale ambiguity* [5], the source 2 will be mapped incorrectly to the microphone space. In contrast, source 1 will have contamination from source 2, due to misaligned beamforming, but will have correct mapping back to the mics. This observation is more evident in the case of $50cm$ movement.

Mathematically speaking, if we assume that

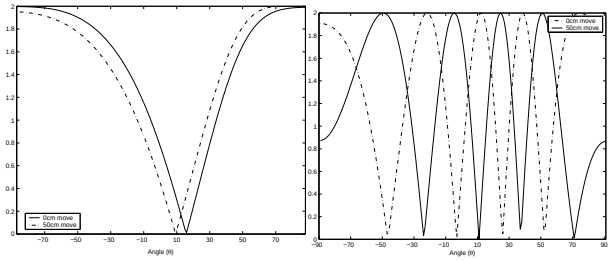

Figure 5: Comparing beamforming patterns at (a) 160Hz and (b) 750Hz.

$$\underline{X} = \begin{bmatrix} A_{11}A_{12} \\ A_{21}A_{22} \end{bmatrix} \begin{bmatrix} S_1 \\ S_2 \end{bmatrix} \quad (4)$$

represents the mixed signals. After separation, we map back to the microphone space, so we try to estimate the following signals:

$$\underline{X}_{s1} = \begin{bmatrix} A_{11} \\ A_{21} \end{bmatrix} S_1, \underline{X}_{s2} = \begin{bmatrix} A_{12} \\ A_{22} \end{bmatrix} S_2 \quad (5)$$

where

$$\underline{X} = \underline{X}_{s1} + \underline{X}_{s2} \quad (6)$$

However, due to misaligned beamforming, one source will get contamination from the other source. Therefore,

$$\underline{X}_{s1} = \begin{bmatrix} A_{11} \\ A_{21} \end{bmatrix} S_1 + \epsilon \begin{bmatrix} G_1 \\ G_2 \end{bmatrix} S_2 \quad (7)$$

where ϵ and G_1, G_2 model the error due to misaligned beamforming. Because equation 6 is a constraint to our reconstruction, this implies that the second source will get no contamination from source 1, but instead will be distorted, due to wrong mapping.

$$\underline{X}_{s2} = \left(\begin{bmatrix} A_{12} \\ A_{22} \end{bmatrix} - \epsilon \begin{bmatrix} G_1 \\ G_2 \end{bmatrix} \right) S_2 \quad (8)$$

5. CONCLUSIONS - OPEN PROBLEMS

In this study, we discussed some benefits and problems encountered using beamforming to solve the permutation problem in FD-ICA. More specifically, we observed that it is very difficult to align the permutations in higher frequencies due to multiple nulls present there. Therefore, a combined approach of *amplitude only* criteria in mid-higher frequencies and *phase information* in lower frequencies is proposed. In addition, using extra sensors, one can achieve better beamforming with subspace techniques (i.e. MuSIC).

We also explored the sensivity of the BSS setup to source movement. Source movement seems to affect more higher than lower frequencies. We also observed that a moving source can be separated without any contamination from the non-moving source, but distorted. On the other hand, the non-moving source remains contaminated from the moving source.

Dealing with moving sources is still an open problem for BSS systems. Most of the convolved mixtures solutions work in *batch mode*, and they are far from working online. This implies that the sources should remain stationary at least for the time that the BSS system needs to adapt, otherwise we will have distortion and contamination as described in the paper.

6. REFERENCES

[1] S. Ikeda and N. Murata. A method of ICA in time-frequency domain. In *Proc. Int. Workshop on ICA and Signal Separation*, pages 365–370, Aussois, France, 1999.

[2] M.Z. Ikram and D.R. Morgan. Exploring permutation inconsistency in blind separation of signals in a reverberant environment. In *ICASSP*, 2000.

[3] M.Z. Ikram and D.R. Morgan. A beamforming approach to permutation alignment for multichannel frequency-domain blind speech separation. In *ICASSP*, 2002.

[4] T.-W. Lee, A. J. Bell, and R. Lambert. Blind separation of delayed and convolved sources. In *Advances in Neural Information Processing Systems*, volume 9, pages 758–764. MIT Press, 1997.

[5] N. Mitianoudis and M. Davies. Audio source separation of convolutive mixtures. *to appear in IEEE Trans. Audio and Speech Processing*, 2003.

[6] L. Parra and C. Alvino. Geometric source separation: Merging convolutive source separation with geometric beamforming. *IEEE Trans. on Speech and Audio Processing*, 10(6):352–362, 2002.

[7] L. Parra and C. Spence. Convolutive blind source separation based on multiple decorrelation. In *Proc. IEEE Workshop on Neural Networks for Signal Processing (NNSP'97)*, Cambridge, UK, 1998.

[8] H. Saruwatari, T. Kawamura, and K. Shikano. Fast-convergence algorithm for ica-based blind source separation using array signal processing. In *Proc. Int. IEEE WASPAA*, pages 91–94, New Paltz, New York, 2001.

[9] P. Smaragdis. Blind separation of convolved mixtures in the frequency domain. *Neurocomputing*, 22:21–34, 1998.