BLIND SEPARATION OF SPATIAL AUDIO USING PROJECTION-BASED METHOD

First Author
Narges Mohammadi
nmohamm4@ur.rochester.edu

ABSTRACT

Audio objects are characterized by spatial and spectro- temporal modeling which spatial modeling neither assume an object has only one underlying source point nor attempt to model complex room acoustics; hence, it is required to focus on the listener's perspective and take each object as the superposition of many contribution with various direction and inter-channel delays. The main goal of this project is to unmix multi-channel audio signals to different constitutive spatial objects which also requires estimating and exploiting inter-channel dependencies of an object. In this case, the geometry of microphone array is unknown and using anechoic mixing model we would be able to analyze the projection of mixture signal on many spatial direction. For separation evaluation performance, Signal to Distortion Ratio (SDR), Signal to Interference Ratio (SIR) and Image to Spatial Distortion Ratio (ISR) would be used.

Keywords: source separation ,multi-channel signal ,spatial modeling, anechoic model

1. INTRODUCTION

Blind separation of spatial multi-channel audio signals has attracted many attentions first by introducing NMF (Non-negative Matrix Factorization) as a solver for under- determined sound demixing problem where the number of audio objects in the mixture signal is greater than the number of received mixture signals. In fact, the feasibility of NMF is due to its ability to give a parts-based decomposition of audio spectrograms and assist to interpolate returned the returned frequency and time-based functions; consequently, this technique is more effective when repeating parts are available in the audio signals such as repeating notes, chords or drum hits. Following NMF method, it has been realized that grouping the time-basis functions to sound objects was a complicated task and some approaches has beeen proposed to tackle this problem. In [1] it has been tried to incorporate additional constraints such as shift invariance in time and frequency [2] or incorporation of harmonicity constraint [3],[4] as well as employing more structured statistical models for audio time-frequency representation. In [5], non-parametric fusion of NMF and hidden Markov models has been used to cluster basis functions associated with the same note together. To overcome the clustering problem, [6] has proposed to train separation models for specific instrument or sources. This method required original isolated objects for analyze along with the mixture and this would introduce the concepts of informed source separation.

Apart from aforementioned methods, there is of particular interest to use spatial information for employing source separation application. The initial efforts for this purpose has been applied in [7] which attempts to incorporate spatial cues utilized Non-negative Tensor Factorization (NTF) and then group the recovered basis function using their spatial positions. More improved NTF-based spatial methods include the weighting of time-frequency bins based on interaural intensity difference function [8]. In [9], the spatial covariance matrix has been used as utilizing spatial information into the separation process which has been proposed to deal with reverberant environment. Furthermore, [10] have a beam-space approach to use spatial cue; however, this method requires knowledge of distance between the microphone of arrays. Moreover, an approach that combines the direction of arrival estimation with spatial covariance matrices is presented in [11] which again requires knowledge of microphone array geometry.

In this work, we aim to implement a novel method to separate multi-channel audio signals into their different constitutive spatial objects while the microphone array geometry is unknown by using anechoic mixing model. Furthermore, instead of analyzing the received multi-channel mixture signal, we prefer to examine the projection of mixture signal on many spatial direction[12],[13]. This method benefits of combining the computational effectiveness of NTF methods with the adequacy of probabilistically methods for spatial audio and this combination could be result in demixing even diffuse spatial objects in a reasonable time.

2. METHOD

We suppose that an I-channel audio signal is received called mixture and its STFT represented as x is a $F \times T \times I$ tensor where F indicates the number of frequency bins and T stands for number of time frames. $x(f,t)$ indicates a $I \times 1$ vector which represent the value of complex spectrum of each channel of the mixture at TF bin (f,t). Now, we introduce the mixing process as:

$$\forall(f,t) \quad x(f,t) = \sum_j y_j(f,t) \tag{1}$$

which denotes that the mixture is sum of J multi-channel signal y_j called as $object\ images$. Generally the mixing

process is such that each image appear as mostly located along one particular direction and ambient sound objects are characterized by the fact that they are coming from many directions; therefore, we need to use a probabilistic model to represent such spatial audio signals. In this regard, the whole procedure could be summarized as in Figure (1).

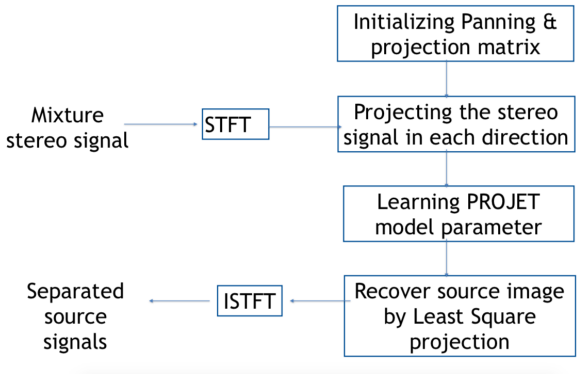


Figure 1. Projection-based source separation workflow

2.1 Punctual anechoic model

We assume that for each object image y_j there is one single underlying monophonic signal called $object\ source$ where its STFT is written as s_j. In Punctual anechoic model, each channel of object model is obtained by a simple delay and gain applied to this unique source as below:

$$\forall(f,t) \quad y_j(f,t) = h(\phi,\tau \mid f) * s_j(f,t) \tag{2}$$

with:

$$h(\phi,\tau \mid f)_i = \phi_i.\exp(-j\frac{2\pi f}{2F-1}\tau_i) \tag{3}$$

where $\phi = [cos\theta, sin\theta]$ indicates the $panning\ vector$ and $\theta \in [0, \pi/2]$ denotes the $panning\ angle$, τ is the $delay\ vector$, τ_i is a delay in sample and f is an index for the frequency band considered (from 0 to $F-1$). Moreover, h as mixing filter is modeling the acoustic transportation of each source to the microphones. This model benfits from the fact that the energy of y_j is the same as s_j. When listening to the resulting image y_j, we have the clear sensation that the object comes from a particular point in space, hence named punctual for this model.

2.2 Diffuse object model

A typical audio object does not come from only one point in space and for real audio signals, it is modeled as superposition of many point sources; therefore, the Punctual anechoic model should be modified as:

$$y_j(f,t) = \int_{(\phi,\tau)\in L} h(\phi,\tau \mid f) q_j(\phi,\tau) s_j(f,t \mid (\phi,\tau)) d(\phi,\tau) \tag{4}$$

where $s_j(f,t \mid (\phi,\tau))$ indicates the object sources each one originates from location(ϕ,τ) and $q_j(\phi,\tau)$ denotes the $Pan-delaygain$ which shows the strength of location (ϕ,τ) in the object. As this model still requires prior knowledge of microphone array to efficiently initialize the mixing filters, some modification would be applied to drop the physical acoustic model and rather implement a receiver point of view. In this regard, each object image y_j is now modeled as the superposition of contributions with various panning directions ϕ and delays patterns τ, from the point of view of the listener, weighted by a gain $q_j(\phi,\tau)$

$$y_j(f,t) = \sum_{l=1}^{L} h_l(f) q_j(l) s_j(f,t \mid l) \tag{5}$$

where $h_l(f)$ and $s_j(f,t \mid l)$ are short-hand notations for $h(\phi_l,\tau_l \mid f)$ and $s_j(f,t \mid (\phi_l,\tau_l))$ and

$$\mathfrak{L} = (\phi_1,\tau_1),\cdots,(\phi_L,\tau_L) \in \mathbb{L}^L \tag{6}$$

is the $source\ location\ set$. This model assumes all object sources $s_j(f,t \mid l)_l$ are independent, although sharing the same underlying energy. This relaxation permits us to model diffuse audio objects, for which the different punctual sources are not necessarily coherent. Furthermore, we need the probabilistic model precisely satisfies assumption that they all share the same "energy", while having their own random phase. It has been assumed that all TF bins (f,t) are independent and choose all $s_j(f,t \mid l)_l$ as independent and distributed from isotropic $\alpha-stable$ distribution denoted by $S_\alpha S_c$[14]:

$$s_j(f,t \mid l) \sim S_\alpha S_c(P_j^\alpha(f,t)) \tag{7}$$

which notably handles large dynamics range and would be simplified to Gaussian distribution for $\alpha = 2$.
$P_j^\alpha(f,t) \geq 0$ is Fractional Power Spectral Density (α-PSD) as energy of object j at TF bin (f,t) shared for this object by the sources at all panning-delays. This model is equivalent to assuming that the underlying object source waveform is both locally stationary and α-stable, which better fits the large dynamic ranges found in audio signals.

2.3 Spatial Projection

In this part, the main concentration would be on how to estimate the parameters of this multi-channel α-model. In this regard, it is proposed to not handle the original multivariate observed mixture $x(f,t) \in \mathbb{C}^I$, but rather many projections of it onto the complex plane (pooling the parameters obtained with these scalar projections is much easier). The first step will then be to select a projection set P, consisting of M elements from the pan-delay set L:

$$P = (\phi_1,\tau_1),\cdots,(\phi_M,\tau_M) \in \mathbb{L}^M \tag{8}$$

For now, consider one of $(\phi_m,\tau_m) \in P$ and the corresponding complex vector $h(\phi_m,\tau_m \mid f)$ short-handed as $h_m(f)$. In this study, we are interested in scalar $\langle h_m(f), y(f,t)\rangle$ which still has $S_\alpha S_c$ distribution:

$$\langle h_m(f), y(f,t)\rangle \sim S_\alpha S_c(P_j^\alpha(f,t) k_m(f)^T Q_j) \tag{9}$$

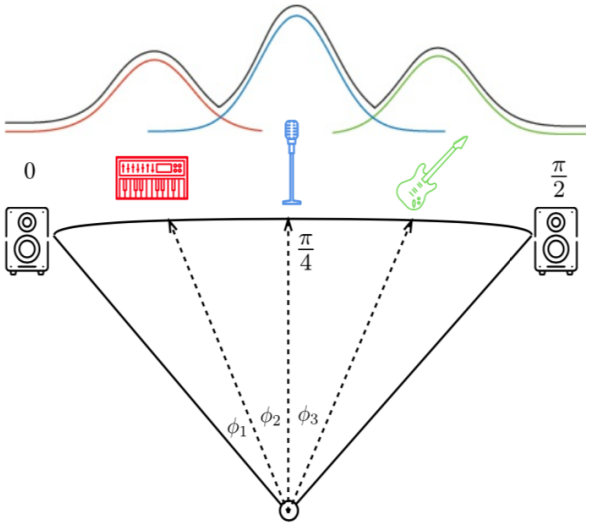


Figure 2. Illustration of standard Pan Law

where $k_m(f)$ is introduced as :

$$k_m(f) \triangleq \left[|\langle h_m(f, h_1(f))\rangle|^\alpha \quad , ..., \quad |\langle h_m(f, h_L(f))\rangle|^\alpha\right]^T \tag{10}$$

and $L \times 1$ vector Q_j as:

$$Q_j \triangleq [q_j^\alpha(1), ..., q_j^\alpha(L)]^T \tag{11}$$

This result can be used for the dot product between $h_m(f)$ and mixture signal as below:

$$\begin{aligned} \langle h_m(f), x(f,t)\rangle &= \sum_j \langle h_m(f), y(f,t)\rangle \\ &= S_\alpha S_c \sum_j (P_j^\alpha(f,t) k_m(f)^T Q_j) \end{aligned} \tag{12}$$

For simplification, all $h_m(f)^T$ would be gathered in a matrix as $Projection\ Matrix$ for frequency f written as N_f:

$$N_f = \begin{bmatrix} h_1(f)^T \\ ... \\ h_M(f)^T \end{bmatrix} \tag{13}$$

The resulting signals $\langle h_m(f), x(f,t)\rangle$ are also gathered in projection matrices c_m as follows:

$$c_m(f,t) \triangleq \langle h_m(f), x(f,t)\rangle \tag{14}$$

For convenience, $c(f,t)$ will indicate the $M \times 1$ vector gathering all $c_m(f,t)$ as shown below:

$$c(f,t) \triangleq [c_1(f,t), ..., c_M(f,t)]^T \tag{15}$$

therefore, c could be called $Projection\ Tensor$ which is equivalent to:

$$c(f,t) = N_f x(f,t) \tag{16}$$

Following equation (12), the marginal distribution of each entry of the projection tensor could be represented as:

$$\begin{aligned} c_m(f,t) &\sim \langle h_m(f), x(f,t)\rangle \\ &= S_\alpha S_c \sum_j (P_j^\alpha(f,t) k_m(f)^T Q_j) \end{aligned} \tag{17}$$

where $k_m(f)$ would be computed only once. The free parameters of this model, consist of the $L \times J$ pan-delays gains Q_j, as well as the $F \times T$ objects α-PSD ($P_j(\alpha)$) which could be written as $\Theta = \left\{P_j^\alpha, Q_j\right\}$

2.4 Separation

For now, assume that all parameters have been estimated and the remaining problem is how to perform demixing and separate the original multi-channel mixture x given the parameters estimated using the projection tensor c? For this purpose, first, the $M \times 1$ projection entries $c(f,t)$ are decomposed into J contributions $y_j^c(f,t)$, which also are of dimension $M \times 1$ and called the *projected images of the objects* as:

$$c(f,t) = \sum_j y_j^c(f,t) \tag{18}$$

For source separation purpose, one solution is using Wiener Filtering to the case of α-harmonisable processes by following equation:

$$\hat{y}_{m,j}^c(f,t) = \frac{P_j^\alpha k_m(f)^T Q_j}{\sum_{j'} P_{j'}^\alpha k_m(f)^T Q_{j'}} c_m(f,t) \tag{19}$$

Then using these projected images y_j^c the corresponding object images would be estimated by:

$$\hat{y_j}(f,t) = N_f^\dagger y_j^c(f,t) \tag{20}$$

where .† indicates pseudo-inversion and waveforms of the separated objects images in the time domain are then easily recovered through inverse STFT transforms of $\hat{y_j}$. The whole procedure is called project (Projection Estimation Technique) which requires mixture signal as input as well as different initialized parameters to construct the projection tensor c. Then, the parameters are iteratively estimated and finally used for separation through (19),(20)

3. PERFORMANCE EVALUATION

To measure the efficiency of the proposed method, two test scenario has been implemented in which different aspects are the methods are analyzed in each of these two cases. In the first scenario, The method is applied on a single recording and in the second scenario it is implemented on an open-sourse testset to measure the generalizibility of proposed approach.

3.1 Performance Evaluation on single recording

To analyze the results of this method in details, the projection-based method is applied on a stereo mixture signal with spectrogram presented in Figure (3) and tried to separate 6 sources (we are not sure about the number of sources, so we just set it as an initial guess) by estimating their corresponding panning angle for each time-frequency point. (The No. of panning direction is set to 41 and the No. of panning projection is set to 15). The estimated panning angle Q_j in Figure (5), indicates that there are just 3 distinct object placed at separated spatial position at 0,39 ,64

degrees . Furthermore, the estimated normalized spectrogram (α-PSD) of separated sources in Figure (4) shown as ($P_j(\alpha)$) illustrates that again 3 distinct spectrogram of sources are present and all the other spectrograms are combination of the separated the ones.

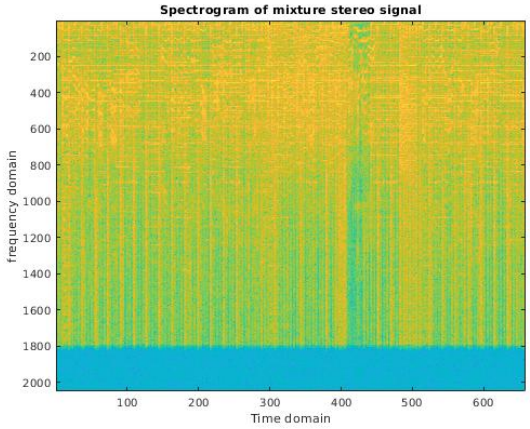


Figure 3. Spectrogram of mixture signal

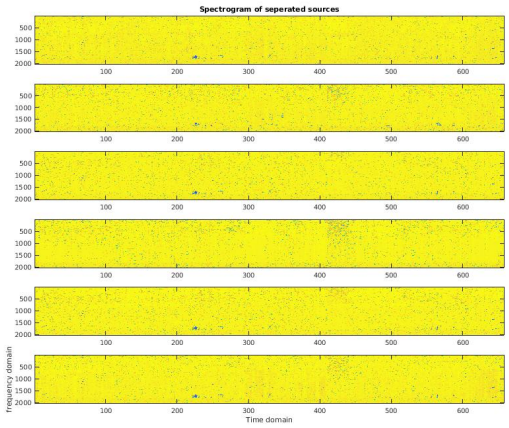


Figure 4. Spectrogram of separated sources

3.2 Performance Evaluation on MSD100 Dataset

To analyze the generalizibility of this method, it has been applied on a testset.The data set is created using open-source MSD100 development set which is consist of 50 full length songs created from mixtures of 4 objects. The testset is created by extracting 30s of each song to create stereo image using pan-delay model and the mixture signal x is created by summing stereo images. In this testset, objects were mixed with an equal angle between them, with no delay between the channels (angle between objects was varied from 10 degrees to 30 degrees in steps of 10) which result in total of 150 test mixtures, with a total of 600 separated sources. The separation evaluation criterias which have been used to compare the results are Signal to Distortion Ratio (SDR), Signal to Interference Ratio (SIR) and

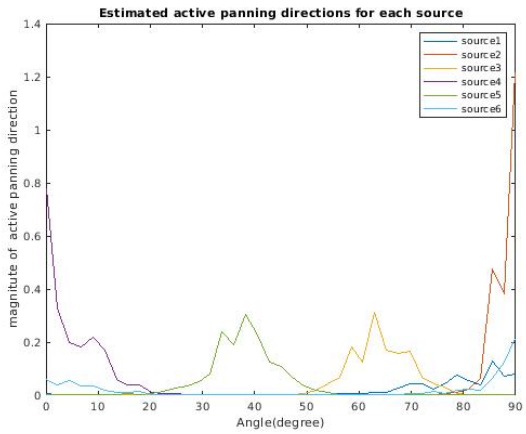


Figure 5. Estimated active panning direction of each source

Angle(degree)	10	20	30
SDR(dB)	5.1	6.1	6.2
SIR(dB)	10.2	9.4	9.1
SAR(dB)	10.1	9.5	9.6
ISR(dB)	11.9	12.2	13.1

Table 1. Separation metrics vs. Angle

Image to Spatial Distortion Ratio (ISR) using BSS Eval Toolbox.

The achieved results in Table (1) indicated that increasing the angle between the spatial position of the objects would results in improved performance metrics.

4. CONCLUSION

In this work, a spatial projection-based method was presented for application of multi-channel audio separation using a mixing model which benefit from the fact that the spatial image of each object is a weighted sum of independent contributions originating from all pan-delay directions. It is shown how to create a set of observations in which some objects are attenuated or enhanced by projecting the multi-channel mixture signal onto a range of spatial-delay direction. The simulation results are expected to outperform other conventional methods using spatial and delay diversity of the objects with no more constraint on spectro-temporal characteristics of audio signals.

5. REFERENCES

[1] R. Jaiswal, D. FitzGerald, D. Barry, E. Coyle and S. Rickard, "Clustering NMF basis functions using Shifted NMF for monaural sound source separation," 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, 2011, pp. 245-248.

[2] D. FitzGerald, M. Cranitch and E. Coyle, "Shifted 2D Non-negative Tensor Factorisation," 2006 IET Irish

Signals and Systems Conference, Dublin, 2006, pp. 509-513.

[3] D. FitzGerald, M. Cranich, and E. Coyle. "Extended non-negative tensor factorisation models for musical sound source separation". Computational Intelligence and Neuroscience, vol. 2008, Article ID 872425, 2008

[4] B. Fuentes, R. Badeau, and G. Richard. "Blind Harmonic Adaptive Decomposition Applied to Supervised Source Separation". In Proceedings European Signal Processing Conference (EUSIPCO 2012), Bucharest, Romania, 2012.

[5] M. Nakano, J. Le Roux, H. Kameoka, T. Nakamura, N. Ono, and S. Sagayama. "Bayesian nonparametric spectrogram modeling based on infinite factorial infinite hidden markov model". In Proceedings of the WASPAA 2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, October 2011

[6] J. Carabias-Orti, M. Cobos, P. Vera-Candeas, and F. Rodríguez-Serrano. "Nonnegative signal factorization with learnt instrument models for sound source separation in close-microphone recordings". EURASIP Journal on Advances in Signal Processing, 2013(1):1–16, 2013

[7] D. FitzGerald, M. Cranitch, and E. Coyle. "Non-negative tensor factorisation for sound source separation". In Proceedings of the Irish Signals and Systems Conference, 2005

[8] Y. Mitsufuji and A. Roebel. "On the use of a spatial cue as prior information for stereo sound source separation based on spatially weighted non-negative tensor factorization". EURASIP Journal on Advances in Signal Processing, 2014(1), 2014.

[9] N.Q.K Duong, E. Vincent, and R. Gribonval. "Underdetermined convolutive blind source separation using spatial covariance models". In Acoustics, Speech and Signal Processing, IEEE Conference on (ICASSP'10), pages 9–12, Dallas, United States, March 2010.

[10] S.N. Lee, S.H. Park, and K. Sung. "Beamspace-domain multi-channel non-negative matrix factorization for audio source separation". Signal Processing Letters, IEEE, 19(1):43–46, Jan 2012.

[11] J. Nikunen and T. Virtanen. "Direction of arrival based spatial covariance model for blind sound source separation". IEEE/ACM Transactions on Audio, Speech Language Processing, 22(3):727–739, 2014.

[12] D. FitzGerald, A. Liutkus and R. Badeau, "Projection-Based Demixing of Spatial Audio," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 9, pp. 1560-1572, Sept. 2016.

[13] E. Cano, D. FitzGerald, A. Liutkus, M. D. Plumbley and F. Stöter, "Musical Source Separation: An Introduction," in IEEE Signal Processing Magazine, vol. 36, no. 1, pp. 31-40, Jan. 2019.

[14] D. FitzGerald, A. Liutkus and R. Badeau, "Projection-Based Demixing of Spatial Audio," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 9, pp. 1560-1572, Sept. 2016.