Blind Speech Source Localization, Counting and Separation for 2-channel Convolutive Mixtures in a Reverberant Environment

Sayeh Mirzaei[1], Hugo Van hamme[1], Yaser Norouzi[2]

[1]Department of Electrical Engineering-ESAT, KU Leuven, Belgium
[2]Department of Electrical Engineering, Amirkabir University of Technology, Tehran, Iran
smirzaei@esat.kuleuven.be, hugo.vanhamme@esat.kuleuven.be, y.norouzi@aut.ac.ir

Abstract

In this paper, the tasks of speech source localization, source counting and source separation are addressed for an unknown number of sources in a stereo recording scenario. In the first stage, the angles of arrival of individual source signals are estimated through a peak finding scheme applied to the angular spectrum which has been derived using non-linear GCC-PHAT. Then, based on the known channel mixture coefficients, we propose an approach for separating the sources based on Maximum Likelihood (ML) estimation. The predominant source in each time-frequency bin is identified through ML assuming a diffuse noise model. The separation performance is improved over a binary time-frequency masking method. The performance is measured by obtaining the existing metrics for blind source separation evaluation. The experiments are performed on synthetic speech mixtures in both anechoic and reverberant environments.

IndexTerms— non-linear GCC-PHAT, angular spectrum, maximum likelihood (ML), binary masking, blind source separation (BSS)

1. Introduction

Extracting individual sources from a mixture of several audio signals is required in many applications including hearing aids, automatic meeting transcription systems, polyphonic music source separation and hands-free teleconference systems. When the impulse responses of the channel between sources and microphones are not known, we deal with a blind source separation (BSS) task.

In [1], a Non-negative Tensor Factorization (NTF) structure is proposed as the generative model of the mixture signal spectrogram. Then, the spectral components, time activations and channel mixture coefficients are estimated on the basis of instantaneous mixtures where the channel coefficients are assumed constant positive real values. In [2], the same NTF modeling is considered and the performance of the EM algorithm is measured and compared with the Multiplicative Update (MU) solution. Also, convolutive mixtures are accounted for by allowing frequency dependent channel coefficients.

There are several drawbacks associated with the above methods. Firstly, they disregard the phase information of the signal since they try to factorize the power or absolute value of the Short Time Fourier Transform (STFT) of the mixture signal. Secondly, the number of sources is taken known in advance to enable them to cluster the learned spectral components.

In a recent work [3], a Bayesian non-parametric approach is considered for estimating the number of sources, channel coefficients and individual source signals. However, the performance is very sensitive to the initial values chosen for the parameters and it is prone to finding a locally optimum solution. Moreover, it is computationally intensive because it involves variational inference for a large set of parameters.

Our proposed approach regards the source separation task as the second stage after source localization. Knowing the channel mixture weights, there will be three categories of methods we can apply for extracting individual sources: (1) Masking the time-frequency (TF) representation of the mixture signal [4,5]; (2) Beamforming-based source separation [6] and (3) Statistical methods where the individual source signals can be derived e.g. as a maximum likelihood solution.

In this paper, we consider speech source separation in a convolutive stereo mixture scenario. In contrast to most conventional methods, our approach to direction finding of the sources is not based on the Time Difference of Arrival (TDOA) estimation from some variant of Generalized Cross-Correlation with Phase Transform (GCC-PHAT) function, hence alleviating the spatial aliasing issue. Instead, we apply a peak finding algorithm to an angular spectrum calculated for individual angle of arrival values based on a non-linear version of GCC-PHAT. The number of sources is then estimated as the number of peaks found subject to some constraints. Even if it is overestimated, in the next step the irrelevant sources will be automatically eliminated (taking near zero values). In the source separation stage, we propose a technique to improve the performance over binary masking. It is based on ML estimation of the predominant source in each TF bin. The BSS evaluation metrics are finally obtained for performance comparison.

The signal model adopted here is based on the far-field assumption, so the received signal in two channels can be represented as

$$X_{ift} = \sum_{k=1}^{K} S_{kft} \exp\left(\frac{j2\pi df(i-1)\cos(\theta_k)}{C} \right) + n_{ift} \qquad i = 1,2 \quad (1)$$

where X_{ift} denotes the complex value of the mixture signal STFT in frequency bin f and time frame t for the i^{th} channel.

n_{ift} represents the ambient noise or room reverberations. d is the distance between two microphones, C is the sound propagation velocity, θ_k denotes the angle of arrival for source k with respect to the line connecting both microphones, S_{kft} is the complex contribution of each source in each TF bin and K is the total number of sources.

The rest of this paper is organized as follows: In section 2, the approach for estimating the source directions of arrival (DOA) is described. In section 3, the proposed source separation method based on the known DOAs is explained. The experiments are outlined in section 4. Section 5 presents our conclusions.

2. Source localization and counting

Most of the source localization methods are based on TDOA estimation in every single TF bin followed by clustering of the estimated values for finding the location of the individual sources [7]. The fundamental deficiency of most of these schemes is the spatial aliasing problem which limits the distance between microphones in the array. Furthermore, the number of the clusters should be known in advance. There are also some drawbacks associated with clustering methods such as k-means, which are very sensitive to the initial choice of centroids.

2.1. Angular spectrum derivation

Due to the above mentioned disadvantages, we did not opt for clustering based methods. Instead, we compute a metric against a range of angles of arrival values distributed uniformly over the interval $[0, \pi]$. The source localization task is carried out by finding the peak locations of this metric. First, we evaluate the following function based on GCC-PHAT [8] in each TF bin against θ values:

$$\mathrm{R}(f,t,\theta) = real\left(\frac{X_{1ft} X_{2ft}^{*}}{\left| X_{1ft} X_{2ft}^{*} \right|} \exp\left(\frac{j2\pi df \cos(\theta)}{C} \right) \right) \qquad (2)$$

where * denotes complex conjugation. For increasing the spatial resolution, a monotonically decreasing non-linear function in the range $[0, 1]$ is applied to the GCC-PHAT metric based on [9]:

$$\mathrm{M}(f,t,\theta) = 1 - \tanh(\alpha \sqrt{1 - R(f,t,\theta)}) \qquad (3)$$

where α is the non-linearity parameter. This non-linear function can make the algorithm more effective in a reverberant environment because it results in sharper peaks corresponding to true source DOAs. To obtain the final angular spectrum, $\mathrm{F}(\theta)$, a summation over all frequency bins and a maximization over all time frames is performed:

$$\mathrm{F}(\theta) = \max_{t} \sum_{f} M(f,t,\theta) \qquad (4)$$

2.2. Peak finding algorithm

In this stage, the tasks of source localization and counting are accomplished using a peak finding algorithm. First the minimum value of the angular spectrum $\mathrm{F}(\theta)$ is subtracted and then it is normalized, i.e. the vector is divided by its maximum value. Afterwards, two constraints on the minimum distance between the peaks and minimum peak height can eliminate the irrelevant peak locations. Here, we set the threshold for minimum peak height to 0.55 and for the minimum peak distances to 4 degrees, implying that the angular distance between the speech sources is assumed to be more than 4 degrees. The decision criterion was optimized on a range of simulated test cases, including reverberated conditions.

3. Extracting individual sources

The approach in this stage relies on the sources found in the previous stage. If the number of sources is less than or equal to two, the task will be carried on more easily. Since we are considering two channel models, in the case where there are two sources, they can be found easily by linearly inverting the channel mixture matrix which can be computed from the previous stage. Most multichannel BSS methods based on ICA also require that the number of microphones be at least as large as the number of sources [10]. However in the underdetermined case, where there are more sources than sensors, they are not applicable even if the channel mixture matrix is known. What we propose here is more general and can be applied to the underdetermined case.

Most conventional speech source separation methods assume sparsity in the TF representation of the speech mixture. This means that just one source is assumed active in each TF bin. Therefore, the dominant source in each bin is identified and then individual binary masks are applied to segregate the source signals. A common issue with binary TF masking methods is the musical noise artifacts which degrade the separation performance. This noise consists of short tones randomly distributed in the separated signal over the time and frequency. Here, we propose a new approach to overcome the shortcomings caused by binary masking.

For binary masking, we discriminate the dominant active source in each TF bin by evaluating the M function (non-linear GCC-PHAT metric) for the estimated DOA's found in the previous step and choosing the source that maximizes this criteria:

$$\begin{aligned} & i_{BM}(f,t) = \arg\max_{k} \mathrm{M}(f,t,\theta_k) \quad k = 1...K \\ & S_{i_{BM}(f,t)ft} = X_{1ft} \\ & S_{kft} = 0 \qquad \forall k \neq i_{BM}(f,t) \end{aligned} \qquad (5)$$

Hence, in the binary masking method, the index i_{BM} specifies the dominant source corresponding to the bin (f,t).

We propose the ML metric for recognizing the dominant source. The source contribution in each TF bin is also derived from the ML solution. Supposing one predominant source in each TF bin, we have:

$$\begin{aligned} & \underline{X}_{ft} = \underline{a}(\theta_k) S_{kft} + \underline{n}_{ft} \\ & \underline{a}(\theta_k) = \left[1 \quad \exp\left(\frac{j2\pi df \cos(\theta_k)}{C} \right) \right]^{T} \end{aligned} \qquad (6)$$

We assume a diffuse noise model, thus the covariance matrix of the zero-mean complex Gaussian noise $\underline{n}_{ft} \sim N(\mathbf{0}, \mathbf{\Sigma_n})$ is given with the following form [11,12]:

$$\mathbf{\Sigma_n} = \sigma^2 \begin{bmatrix} 1 & \text{sinc}\left(\frac{2fd}{C}\right) \\ \text{sinc}\left(\frac{2fd}{C}\right) & 1 \end{bmatrix}, \quad \text{sinc}(x) = \frac{\sin(\pi x)}{\pi x} \tag{7}$$

In this paper, we choose σ^2 equal to an arbitrarily small constant 10^{-7}, independent of the data. The mixture signal log-likelihood conditioned on the k^{th} source being active in bin (f,t), is given by:

$$\begin{aligned} \log P(\underline{X}_{ft} \mid S_{kft}) = &-2\log\pi - \log\left|\mathbf{\Sigma_n}\right| \\ &-(\underline{X}_{ft} - \underline{a}(\theta_k)S_{kft})^h \mathbf{\Sigma}_n^{-1}(\underline{X}_{ft} - \underline{a}(\theta_k)S_{kft}) \end{aligned} \tag{8}$$

where superscript h denotes the conjugate transpose operation. The ML solution for source k, $\hat{S}_{kft}$ is obtained by derivation of the above expression w.r.t S_{kft}:

$$\hat{S}_{kft} = \frac{\underline{a}^h(\theta_k)\mathbf{\Sigma_n^{-1}}\underline{X}_{ft}}{\underline{a}^h(\theta_k)\mathbf{\Sigma_n^{-1}}\underline{a}(\theta_k)} \tag{9}$$

which has a similar structure as the MVDR solution [13]. Substituting (9) in (8) will result in the maximum log-likelihood of the data corresponding to the k^{th} source, $\text{ML}(f,t,\theta_k)$. We take the source with maximum corresponding log-likelihood as the dominant active source in each TF bin. Then, the contribution of the active source is derived from (9):

$$\begin{aligned} &i_{ML}(f,t) = \arg\max_k \text{ML}(f,t,\theta_k) \quad k = 1...K \\ &S_{i_{ML}(f,t)ft} = \hat{S}_{i_{ML}(f,t)ft} \quad (\textit{given by } (9)) \\ &S_{kft} = 0 \qquad \forall k \neq i_{ML}(f,t) \end{aligned} \tag{10}$$

Obtaining the ML solution for the dominant source will lead to better BSS evaluation metrics as will be demonstrated in the experiments. The metrics are compared with those obtained for the binary masking case.

The major advantage of the proposed method is that it is easy to implement and it is not computationally expensive. Perhaps better performance can be achieved by calculating the output power of a proper beamformer for deciding about the predominant sources in each TF bin but this would need the derivation of an empirical covariance matrix based on some adjacent TF bins which makes the algorithm more time consuming and also more sensitive to the accuracy of the estimated covariance matrix.

4. Experiments

The experiments are performed on synthetic stereo mixtures of speech signals. The male speech signals are taken from dev2 dataset of the SiSEC'08 "under-determined speech and music mixtures" task [14]. The sampling frequency is 16 kHz. The time duration of all individual sources is 10s. The STFT was computed with cosine windows of length 1024 and 50% overlap. Three male speech signals are synthetically mixed based on the far-field model given in (1) without considering noise. The source directions w.r.t the microphone array axis are taken [30° 80° 130°]. Microphone spacing is equal to 10 cm. The angular spectrum is calculated for 180 uniformly spaced angles in the range $[0, \pi]$. The non-linearity parameter α is taken equal to 2. First, we investigate the performance of the DOA estimation stage in the presence of an uncorrelated additive zero-mean Gaussian noise in the anechoic case. Figure 1 depicts the RMS error of the estimated source directions against different SNR values and compares the performance of nonlinear GCC-PHAT with the GCC-PHAT metric. The RMSE is measured considering 100 different noise realizations corresponding to each SNR value. The three largest peak locations give the estimated directions in this case; however the constraint on the minimum peak distance (section 2.2) is applied to avoid very close estimated directions. It can be seen that the nonlinear version of GCC-PHAT leads to lower estimation errors. The performance of the source counting against SNR is plotted in figure 2. The probability of finding the true number of sources is drawn versus overestimation or underestimation probabilities obtained for 100 runs. In this case, both of the constraints are applied to the peak finding algorithm.

The separation quality is measured by calculating the evaluation metrics including Signal to Distortion Ratio (SDR), Signal to Interference Ratio (SIR) and Signal to Artifact Ratio (SAR) [15]. Initially, we performed experiments in a noiseless anechoic case without considering reverberation. The estimated source directions and the resulted evaluation metrics for the noiseless anechoic case are reported in table 1. It can be seen that the estimated directions perfectly match with the true ones in this case. The proposed approach (ML) for source separation has improved the SDR and SAR measures corresponding to the first and third separated sources. The SIR of the second source is also increased compared to the binary masking (BM) method.

For the reverberant case, the mixture signal is synthetically generated using the Roomsimove Toolbox [16] for a rectangular room of dimensions 4m × 3m × 2.5m and omnidirectional microphones. The reflection coefficient of the walls is taken 0.2. Uncorrelated additive noise is not considered here. The microphone spacing and the source directions have the same values as in the anechoic case. The source distance from the array is 50cm. This configuration leads to the reverberation time (RT60) of about 300ms. The results of the estimated DOAs and separation quality are listed in table 2. Reverberation leads to errors in the estimated directions, but the proposed approach for source separation outperforms the binary masking method. All performance metrics are superior for the ML solution relative to the binary mask, a conclusion which is in line with [17].

The diffuse noise assumption appears to be a sufficiently good model for reverberation noise. A proper choice of the non-linear parameter α along with applying the mentioned peak finding constraints can lead to more accurate estimation of the source directions in the reverberant case. Here, we have opted for α=2 empirically. When there is a stronger reverberation effect, α should be increased. An alternative for a highly reverberant environment could be to consider a preprocessing stage, such as the first scheme proposed in [18], to dereverberate the mixture. Then, our proposed approach can be applied.

The results indicate that the proposed ML approach outperforms the binary masking scheme based on non-linear GCC-PHAT metric. This performance difference can be attributed to the fact that the ML approach provides a better means for recognizing the dominant source as well as obtaining the contribution of that source in each TF cell.

5. Conclusion

In this paper, a novel approach with low computational cost is proposed for the tasks of source localization, counting and separation. The localization algorithm outperforms some existing clustering based algorithms due to the points mentioned in section 2. The nonlinear GCC-PHAT metric can improve the direction estimation accuracy in both noisy and reverberant environments compared to GCC-PHAT. The applied constraints for peak finding can give us a proper estimation of the number of sources. The separation is done knowing the channel mixture coefficients given by the previous step. The dominant active source in each TF bin is recognized by an ML approach considering diffuse noise model. The source contribution is also given by the ML solution. The experiments show that the proposed separation algorithm improves the performance compared to the binary masking approach in both anechoic and reverberant cases.

6. Acknowledgements

This research was funded by the KU Leuven research grant GOA/14/005 (CAMETRON).

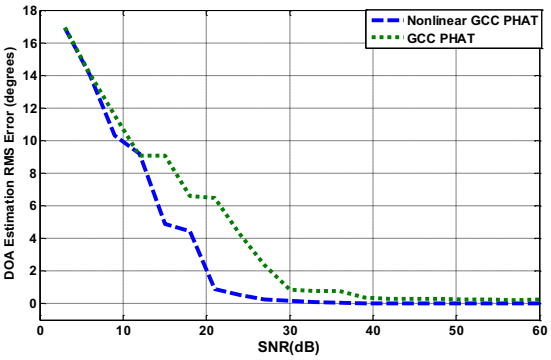

Figure 1. *DOA Estimation RMS Error*

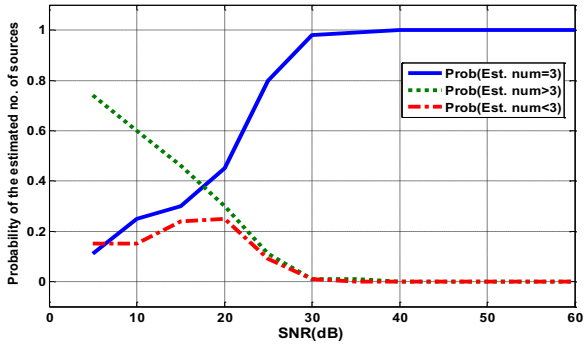

Figure 2. *Probability of finding the true number of sources against SNR*

Table 1. *Evaluation metrics for the noiseless anechoic case. DOA is expressed in degrees, SDR, SIR and SAR in dB.*

		Binary Masking			Proposed ML method		
	DOA	**SDR**	**SIR**	**SAR**	**SDR**	**SIR**	**SAR**
Source 1	30	7.7	21.6	7.9	10.0	21.7	9.2
Source 2	80	7.2	14.9	8.1	7.4	18.8	8.2
Source 3	130	12.4	22.5	12.9	15.5	22.6	16.4

Table 2. *Evaluation metrics for the reverberant case. DOA is expressed in degrees, SDR, SIR and SAR in dB.*

		Binary Masking			Proposed ML method		
	DOA	**SDR**	**SIR**	**SAR**	**SDR**	**SIR**	**SAR**
Source 1	31	7.5	19.7	7.8	8.6	20.0	9.0
Source 2	78	6.6	13.3	7.8	7.3	17.7	7.9
Source 3	130	12.2	21.8	12.8	15.1	21.9	15.9

7. References

[1] C. Fevotte and A. Ozerov, "Notes on nonnegative tensor factorization of the spectrogram for audio source separation: statistical insights and towards self-clustering of the spatial cues," *Proc. 7th International Symposium on Computer Music Modeling and Retrieval (CMMR'2010)*, 2010.

[2] Ozerov, A.; Févotte, C., "Multichannel Nonnegative Matrix Factorization in Convolutive Mixtures for Audio Source Separation," *Audio, Speech, and Language Processing, IEEE Transactions on* , vol.18, no.3, pp.550,563, March 2010.

[3] H.Kameoka, M.Sato, "Bayesian nonparametric approach to blind separation of infinitely many sparse sources," *IEICE Trans. Fundamentals,* vol.E96-A, no.10, October 2013.

[4] Yilmaz, O.; Rickard, Scott, "Blind separation of speech mixtures via time-frequency masking," *Signal Processing, IEEE Transactions on* , vol.52, no.7, pp.1830,1847, July 2004.

[5] Reju, V.G.; Soo Ngee Koh; Soon, I.Y., "Underdetermined Convolutive Blind Source Separation via Time–Frequency Masking," *Audio, Speech, and Language Processing, IEEE Transactions on* , vol.18, no.1, pp.101,116, Jan. 2010.

[6] M. Dmour and M. Davies, "Under-determined speech separation using GMM-based non-linear beamforming," *European Signal Processing Conference (EUSIPCO'08)*, (Lausanne, Switzerland), Aug. 2008.

[7] S. Araki, H. Sawada, R. Mukai, and S. Makino, "A novel blind source separation method with observation vector clustering," *Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC '05)*, pp. 117–120, Eindhoven, The Netherlands, September 2005.

[8] C. Knapp, G. Carter, "The generalized cross-correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech and Signal Processing* 24 (4) (1976) 320-327.

[9] B. Loesch, B. Yang, "Blind source separation based on time-frequency sparseness in the presence of spatial aliasing," *9th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, 2010.

[10] Sawada, H.; Mukai, Ryo; Araki, S.; Makino, S., "A robust and precise method for solving the permutation problem of

frequency-domain blind source separation," *IEEE Transactions* on *Speech and Audio Processing*, vol.12, no.5, pp.530-538, Sept. 2004.

[11] I. A. McCowan, H. Bourlard, "Microphone Array Post-Filter Based on Noise Field Coherence," *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 6, pp. 709-716, 2003.

[12] C. Blandin, E. Vincent, A. Ozerov, "Multi-source TDOA estimation using SNR-based angular spectra, " *IEEE International Conference on Acoustics,Speech, and Signal Processing (ICASSP),* 2011.

[13] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969.

[14] E. Vincent, S. Araki, P. Bo_lld, "The 2008 signal separation evaluation campaign: A community-based approach to large-scale evaluation," *Proc .Int. Conf. on Independent Component Analysis and Signal Separation (ICA)*, 2009, pp. 734_741.

[15] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. on Audio, Speech and Language Processing*, 14(4):1462-1469, 2006.

[16] E. Vincent, D. Campbell, Roomsimove, a Matlab toolbox for the computation of simulated room impulse reponses for moving sources, http://www.irisa.fr/metiss/members/evincent/software.

[17] Jensen, J.; Hendriks, R.C., "Spectral Magnitude Minimum Mean-Square Error Estimation Using Binary and Continuous Gain Functions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol.20, no.1, pp.92,102, Jan. 2012.

[18] Jan, T., Wenwu Wang, "Joint blind dereverberation and separation of speech mixtures," Proceedings of the 20th European Signal Processing Conference (EUSIPCO), Aug. 2012.