Unsupervised Techniques for Blind Audio Source Separation

Sushmita T

Speech Processing Laboratory, International Institute of Information Technology,
Hyderabad - 500032, India

sushmita.t@research.iiit.ac.in

1. Motivation For Research

A typical audio scene is generated by several concurrent audio sources. For example, the speech of a speaker is often obscured by concurrent background speakers or other indoor and outdoor sounds. A human listener's brain can easily track the audio source of interest [1]. The aim of source separation is to equip machine listeners with similar skills and to solve the cocktail party problem [2]. The problem dimensionality (relation between number of speakers and microphones), the type of mixing environment (instantaneous or convolutive), and noisy, non-stationary, reverberative environments influence the complexity of source separation.

The segregated target signal is either listened to or processed further, resulting in better performance in several applications like speech enhancement for hearing aids, automatic speech recognition, automatic speaker recognition and identification, real time speech translation, automatic indexing of large audio databases, audio information retrieval, automatic music transcription etc. Since the advantages offered by BSpS are many, different kinds of algorithms have been proposed for a few decades now.

Currently, the state-of-the-art in speech separation is DNN based monaural and array separation algorithms [3]. But DNNs fail in fully blind settings. The solution is to develop unsupervised techniques which need no training data. Though many unsupervised or fully blind methods have been proposed, there are many potential techniques that remain unexplored. The motivation for research has stemmed from these factors. Further, these unsupervised methods may be used for initialization of DNNs for faster convergence and better performance, thus combining the aspects of unsupervised techniques and deep learning in source separation.

2. Key Issues Identified/Addressed

Since 2005, probabilistic model methods [4–11] and unsupervised CASA approaches [12–14] are among the widely used unsupervised techniques for BSpS.

The key point identified is that probabilistic models have good potential for unsupervised BSpS and that they are not fully explored. The fact that these methods are not fully explored is stated in [6]. It is stated that among Gaussian model-based approaches for audio source separation, only a few combinations of spatial covariance models and spectral variance models were combined and many combinations were not investigated.The total number of configurations for J sources is 2×24^J but only 16 were investigated. The fact that probabilistic methods have good potential for unsupervised BSpS when deep learning fails is given in [4]. It is stated that for audio source separation, probabilistic models (such as multichannel NMF methods) are applicable in situations where deep learning methods are not applicable.

Among CASA based approaches there are some simple, powerful and flexible unsupervised BSpS algorithms. These algorithms may be improved and new algorithms along similar lines may be proposed to achieve source separation. In [12], a CASA based unsupervised algorithm in reverberant environments generates time-frequency masks to achieve BSpS. It combines generalized cross correlation (GCC) [15] and non-negative matrix factorization (NMF) [16] and is hence called GCC-NMF. This is a simple, powerful and flexible state-of-the-technique for unsupervised BSpS.The algorithm is so flexible that it can be applied to separation of concurrent speech in reverberant environments, to speech enhancement in real-world background noise and to segregation of noisy mixtures of moving speakers. To improve the performance of GCC-NMF, specific subtasks within the algorithm with potential for improvement are identified. Also, methods to improve these subtasks are proposed. In Section 3, results and discussions of the proposed methods are mentioned. The objectives and road map of the thesis are presented in Section 4.

3. Results And Discussions

CASA approach for BSpS: Following the discussion in Section 2, a brief summary of subtasks identified and the rationale behind the identification is presented here. In GCC-NMF, STFT time-frequency (T-F) representation is used. Time-Frequency masking generates artifacts due to time-frequency overlap of the sources [17, 18]. To reduce the artifacts, a high resolution T-F representation is needed. Hence, the first subtask identified is to use a better T-F representation. In GCC-NMF the localization technique used is generalized cross-correlation with phase transform (GCC-PHAT) [15]. GCC based technique has proved to be a good localization technique [12, 19]. However, in [20] it is stated that the adaptive eigenvalue decomposition (AED) algorithm demonstrated the best performance among various localization methods in both noise and reverberation conditions, showing its applicability for real applications. Therefore, the second subtask is either to use AED instead of GCC-PHAT or to develop new localization techniques robust to reverberation and noise. As mentioned in [20], drawbacks of AED were its computational complexity and its need for synchronization of signals at various channels. But today, computational complexity is not a major issue while synchronization still remains a major challenge as discussed in the *SiSEC2018* Challenge [21]. In addition, recently proposed approaches for localization using SFF [22] may be used.

The first subtask is addresed as mentioned here. It is proposed to use a high T-F resolution representation like Single Frequency Filtering (SFF) representation [23] in the GCC-NMF framework. It is mentioned in [23] that SFF provides better T-F resolution compared to STFT.

Database: The *dev1* development set of *UND 2016* of Signal Separation Evaluation Campaign *SiSEC2016* is used. It consists of stereo recordings of 3 and 4 concurrent speakers in reverberant conditions. Further details are available at [24].

Evaluation metrics: The toolbox used for the objective evaluation of a source separation system is PEASS toolbox [25]. It provides four perceptually motivated criteria: the overall perceptual score (OPS), the target-related perceptual score (TPS), the interference-related perceptual score (IPS), and the artifact-related perceptual score (APS).

Results and Discussions: Angular spectrogram obtained after GCC-PHAT on SFF of dual speech mixtures of 3 speakers and the estimated time difference of arrivals (TDOAs) are shown in Figure 1.

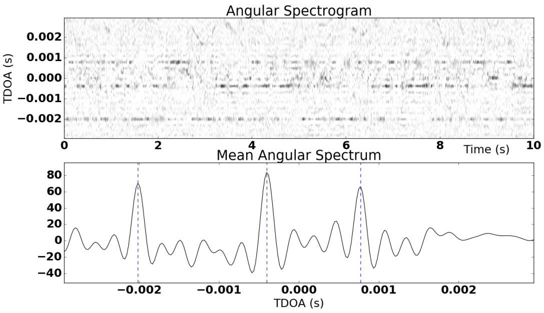

Figure 1: *Angular spectrogram and estimated TDOAs by GCC-PHAT on SFF of dual speech mixtures comprising three concurrent speakers.*

The implementation of modified GCC-NMF with high resolution T-F representaion gave the following results. Tables (1) and (2) show the mean PEASS scores and standard deviations obtained using the data of 3 speakers and 4 speakers respectively. In Table 1, improved APS and OPS, but reduced TPS and IPS are observed.

Table 1: *Mean PEASS scores ± standard deviation, obtained with the SiSEC dev1 live speech recording dataset of 3 speakers.*

	OPS	TPS	IPS	APS
Baseline	19 ± 6	31 ± 15	76 ± 4	3 ± 2
Proposed	27 ± 5	15 ± 6	43 ± 16	23 ± 7

Table 2: *Mean PEASS scores ± standard deviation, obtained with the SiSEC dev1 live speech recording dataset of 4 speakers.*

	OPS	TPS	IPS	APS
Baseline	22 ± 8	40 ± 18	78 ± 4	1 ± 1
Proposed	23 ± 7	21 ± 8	41 ± 17	31 ± 12

In Table 2, improved APS and reduced TPS and IPS are observed. OPS remains the same. Increase in APS, in both the cases, may be attributed to increase in spectral and temporal resolutions in SFF output [17, 18]. Because of the increase in the resolutions in T-F domain, the overlap of sources is reduced resulting in less artifacts in the target signals. Decrease in TPS and IPS may be because of improper masks generated in the proposed method. To summarize, by using a high resolution T-F representation, artifacts such as musical noise have substantially reduced as proved by the increased APS. However, the reasons for decrease in TPS and IPS are yet to be analyzed.

Probabilistic models for BSpS: Probabilistic model methods encode some prior information of sources and mixing environments by assuming that each T-F point is an independently distributed random variable (Gaussian, Laplacian) in the T-F domain. An estimation criterion such as maximum likelihood (ML) or maximum-a-posteori (MAP) of the parameters is formed and an estimation algorithm like expectation maximization (EM) is implemented. The sources are reconstructed by Wiener filtering. The source and mixing models may be trained initially. These models perform well in reverberant conditions if the sources are modelled as full-rank spatial covariance matrix.

Database: The database is the same as mentioned above for *CASA approach for BSpS.*

Evaluation metrics: Apart from *PEASS* toolbox, *BSS-Eval* toolbox which features the signal to distortion ratio (SDR), the source image to spatial distortion ratio (ISR), the signal to interference ratio (SIR), and signal to artifacts ratio (SAR) metrics is used [26].

Current Status: The study and implementation of some of the established techniques [8, 10] in a general flexible framework for audio source separation [6] is completed . The model in [10], models the contribution of each source to all mixture channels in T-F domain as zero-mean Gaussian random variable with covariance capturing the spatial characteristics of the sources. It does not model spectral power of the sources. In [8] the spectral power of sources is modelled as non-negative matrix factorization (NMF) with the Itakura-Saito divergence in a T-F domain along with spatial modelling resulting in improved performance.

4. Objectives And Road Map Of Thesis

Objectives Of The Thesis

To develop unsupervised techniques for BSpS using either the probabilistic modelling approach or CASA approach that are at par with current state-of-the-art. To exhaustively test the developed algorithms for speech mixtures collected in natural environments such as day-to-day home and office environments consisting of many challenging issues such as moving sources in noisy and reverberant environments.

Road Map To The Thesis

In CASA based approaches, the first task is to study , analyze and improve existing localization techniques or develop a new one that is robust in practical environment because a good localization technique leads to a good source separation system. This task is partially accomplished. The next goal is to decide a suitable T-F domain in which the sources would be sparsely represented. Currently, SFF representation is being researched.The final task is to develop an algorithm for BSpS in the chosen T-F domain using the proposed localization technique.

In probabilistic methods the first task is to analyze the implemented methods and identify the subtasks, within the methods, which when modified will lead to performance improvement. Subtasks could be modifying either spatial covariance model of the mixing environment or spectral variance model of the sources in the Local Gaussain Framework (LGM) as cited in [6].The subsequent goal is to propose a new model for the identified subtask.The final goal is to integrate the new model into the existing framework and test for the performance improvement.

5. Target Beneficiaries

The developed unsupervised BSpS algorithm will be beneficial to many speech applications as mentioned in Section 1, when the training data is not available. Additionally, the algorithm may be used to initialize parameters in a supervised setting leading to faster convergence of supervised techniques.

6. References

[1] DeLiang Wang and Guy J Brown, *Computational Auditory Scene Analysis: Principles, algorithms, and applications*, 2006.

[2] Simon Haykin and Zhe Chen, “The cocktail party problem,” *Neural Computation*, vol. 17, no. 9, pp. 1875–1902, 2005.

[3] D. Wang and J. Chen, “Supervised speech separation based on deep learning: An overview,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing,*, vol. 26, no. 10, pp. 1702–1726, October 2018.

[4] Ozerov A., Fvotte C., and Vincent E., *Audio Source Separation*. Springer, 2018.

[5] T. Virtanen, J. F. Gemmeke, B. Raj, and P. Smaragdis, “Compositional models for audio processing: Uncovering the structure of sound mixtures,” *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 125–144, March 2015.

[6] A. Ozerov, E. Vincent, and F. Bimbot, “A general flexible framework for the handling of prior information in audio source separation,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1118–1133, May 2012.

[7] E.Vincent, M.Jafari, S.A.Abdallah, M.D.Plumbley, and M.E.Davies, *Machine Audition: Principles, Algorithms and Systems*, 2010.

[8] A. Ozerov and C. Fevotte, “Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, March 2010.

[9] N. Q. K. Duong, E. Vincent, and R. Gribonval, “Spatial covariance models for under-determined reverberant audio source separation,” in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, October 2009, pp. 129–132.

[10] N. Q. K. Duong and E. Vincent and R. Gribonval, “Under-determined reverberant audio source separation using a full-rank spatial covariance model,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, September 2010.

[11] C. Fevotte and J. F. Cardoso, “Maximum likelihood approach for blind audio source separation using time-frequency Gaussian source models,” in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, October 2005, pp. 78–81.

[12] S. U. N. Wood, J. Rouat, S. Dupont, G. Pironkov, S. U. N. Wood, J. Rouat, S. Dupont, and G. Pironkov, “Blind speech separation and enhancement with gcc-nmf,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing,*, vol. 25, no. 4, April 2017.

[13] T. Virtanen, “Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066–1074, March 2007.

[14] M. D. P. Beiming Wang, “Musical audio stream separation by non-negative matrix factorization,” in *Proc. DMRN Summer Conference*, 2005.

[15] C. Knapp and G. Carter, “The generalized correlation method for estimation of time delay,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, August 1976.

[16] Daniel D Lee and H Sebastian Seung, “Learning the parts of objects by non-negative matrix factorization,” *Nature*, vol. 401, no. 6755, p. 788, 1999.

[17] Emmanuel Vincent, *Blind Audio Source Separation: A review of state-of-the-art techniques*. [Online]. Available: https://www.irisa.fr/metiss/members/evincent/keynoteICArn05.pdf

[18] D. Wang, “Time-frequency masking for speech separation and its potential for hearing aid design,” *Trends in Amplification*, vol. 12, no. 4, pp. 332–353, 2008.

[19] Sean U. N. Wood and Jean Rouat, “Blind speech separation with GCC-NMF,” in *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2016, pp. 3329–3333.

[20] J. Chen, J. Benesty, and Y. Huang, “Time delay estimation in room acoustic environments: An overview,” *EURASIP Journal on Advances in Signal Processing*, May 2006.

[21] F.-R. Stöter, A. Liutkus, and N. Ito, “The 2018 signal separation evaluation campaign,” in *Proc. Int. Conf. Latent Variable Analysis and Signal Separation*, 2018, pp. 293–305. [Online]. Available: http://sisec.inria.fr/2018-asynchronous-recordings-of-speech-mixtures/

[22] N. Chennupati, B. H. V. S. N. Murthy, and B. Yegnanarayana, “A Signal Processing Approach for Speaker Separation Using SFF Analysis,” in *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2017.

[23] S. R. Kadiri and B. Yegnanarayana, “Epoch extraction from emotional speech using single frequency filtering approach,” *Speech Communication*, vol. 86, pp. 52 – 63, 2017.

[24] A. Liutkus, F.-R. Stöter, Z. Rafii, D. Kitamura, B. Rivet, N. Ito, N. Ono, and J. Fontecave, “The 2016 signal separation evaluation campaign,” in *Proc. Int. Conf. Latent Variable Analysis and Signal Separation*, 2017, pp. 323–332. [Online]. Available: https://sisec.inria.fr/sisec-2016/2016-underdetermined-speech-and-music-mixtures/

[25] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, “Subjective and objective quality assessment of audio source separation,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2046–2057, September 2011. [Online]. Available: http://bass-db.gforge.inria.fr/peass/

[26] E. Vincent, R. Gribonval, and C. Fevotte, “Performance measurement in blind audio source separation,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, July 2006.