INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Blind Speech Separation with GCC-NMF

Sean UN Wood, Jean Rouat

NECOTIS, GEGI, Université de Sherbrooke, Canada
sean.wood@usherbrooke.ca, jean.rouat@usherbrooke.ca

Abstract

We introduce a blind source separation algorithm named GCC-NMF that combines unsupervised dictionary learning via non-negative matrix factorization (NMF) with spatial localization via the generalized cross correlation (GCC) method. Dictionary learning is performed on the mixture signal, with separation subsequently achieved by grouping dictionary atoms, over time, according to their spatial origins. Separation quality is evaluated using publicly available data from the SiSEC signal separation evaluation campaign consisting of stereo recordings of 3 and 4 concurrent speakers in reverberant environments. Performance is quantified using perceptual and SNR-based measures with the PEASS and BSS Eval toolkits, respectively. We compare our approach with other NMF-based speech separation algorithms including unsupervised and semi-supervised approaches. GCC-NMF outperforms the unsupervised model-based approach that combines NMF with spatial covariance mixture models, and compares favourably to semi-supervised approaches that leverage prior knowledge and information, despite being purely unsupervised itself.
Index Terms: cocktail party problem, blind speech separation, interaural time difference, NMF, GCC, PHAT, CASA

1. NMF and Blind Speech Separation

1.1. The Cocktail Party Problem

The cocktail party problem [1, 2] is a classic blind source separation problem that involves separating mixtures of concurrent speech signals in real-world environments. Improving separation algorithms, resulting in greater suppression of interference and fewer artifacts, will impact the quality and robustness of assistive listening devices including hearing aids and cochlear implants, as well as the performance of speech recognition systems increasingly pervasive in recent years. Major challenges of the blind speech separation problem stem from underdetermined mixing systems, reverberant environments, the presence of noise, and the non-stationarity of speech. However, with the advent of powerful machine learning algorithms including non-negative matrix factorization (NMF) [3, 4], in addition to increasing computational power, significant progress is being made [5, 6].

1.2. NMF and Speech Separation

NMF is an unsupervised dictionary learning algorithm that lies at the heart of a wide variety of sound separation techniques ranging from blind to weakly and strongly guided approaches [7, 8]. Well-suited to the compositional nature of sound mixtures, NMF yields non-destructive, parts-based representations of mixture spectrograms. However, when applied to mixtures of complex sounds including speech, sources are encoded across multiple dictionary atoms, requiring subsequent grouping of atoms to achieve separation. While many solutions to this problem involve some form of supervision, unsupervised approaches, including that presented in this paper, have also been proposed.

Supervised solutions to the over-segmentation problem leverage prior knowledge or information. For simple sounds, dictionary elements may be grouped by hand via inspection [9], however this approach quickly becomes cumbersome as the source complexity or number of sources increases. A more common supervised approach is to use isolated source recordings to adapt source-specific dictionaries, and subsequently concatenate the dictionaries to encode mixture signals [10]. The encoding process then achieves separation, as each source is encoded by its corresponding dictionary. Another approach involves using prior knowledge of the kinds of sources present in the mixture signal to constrain parts of the NMF dictionary such that they correspond to the sources of interest [11].

Unsupervised solutions typically make use of spatially distributed microphones, combining NMF with spatial information to achieve separation. A common *model-based* approach is to learn a set of source-specific dictionaries, adapting a set of corresponding mixing models in parallel. Mixing models may take the form of spatial covariance matrices [12], while the dictionaries may be made more complex with a multi-layer structure [11]. However, the spatial covariance matrix approaches are sensitive to initialization, and require constrained dictionaries in practice for good results [11, 13]. Another unsupervised approach is to combine NMF with traditional beamforming algorithms [14, 15], however these approaches are developed for large microphone arrays as opposed to the two-channel case we consider here.

1.3. Proposed Approach: GCC-NMF

In this work, we propose a new approach to combining spatial information with NMF, providing a means to group dictionary atoms based on their spatial origin in an unsupervised fashion. By combining the Generalized Cross Correlation (GCC) source localization method with an NMF dictionary learned on a mixture signal, individual dictionary atoms are localized over time, and grouped according to their spatial origin. We begin with a presentation of NMF and GCC in Section 2, followed by a development of the GCC-NMF source separation algorithm in Section 3. Experimental analysis of the effects of NMF parameters on separation performance as well as a comparison with other unsupervised and semi-supervised approaches is then presented in Section 4, followed by a conclusion in Section 5.

2. Foundations: NMF and GCC

In this section, we present the foundations of the GCC-NMF separation algorithm, namely the NMF dictionary learning algorithm and the GCC method of source localization.

http://dx.doi.org/10.21437/Interspeech.2016-1449

2.1. NMF

Input to the NMF algorithm consists of a magnitude time-frequency representation of the mixture signal, represented mathematically as a non-negative matrix V_{ft} with f and t indexing frequency and time respectively. NMF decomposes this spectrogram into two non-negative matrices: a dictionary matrix W_{fd} and a coefficient matrix H_{dt}, such that their product $\Lambda = \mathrm{WH}$ approximates V. The d columns of W are referred to as dictionary atoms: non-negative functions of frequency that are combined linearly with the corresponding coefficients at each point in time to reconstruct the corresponding column of the input spectrogram.

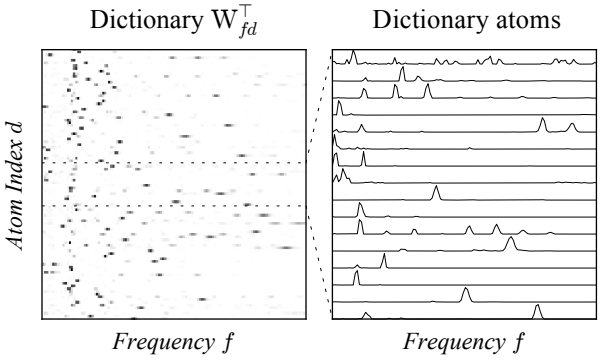

Figure 1: Dictionary learned by NMF on a mixture of speech signals: atoms are non-negative functions of frequency.

The NMF learning algorithm optimizes a cost function that includes a reconstruction error term and an optional coefficient sparsity inducing term. Various measures of reconstruction error have been used, several of which generalize to the β divergence $D_\beta(\mathrm{V}|\Lambda)$ including the Euclidian distance and generalized Kullback-Leibler divergence [16]. An l_1 norm is typically used for coefficient sparsity [17]. Multiplicative update rules are then defined such that by initializing W and H randomly and updating them iteratively, the algorithm converges to a local minimum of the cost function. The update rules for $D_\beta(\mathrm{V}|\Lambda)$ with l_0 sparsity are defined as,

$$\mathrm{H} \leftarrow \mathrm{H} \odot \frac{\mathrm{W}^\top \left(\mathrm{V} \odot \Lambda^{\beta-2}\right)}{\mathrm{W}^\top \Lambda^{\beta-1} + \alpha} \tag{1}$$

$$\mathrm{W} \leftarrow \mathrm{W} \odot \frac{\left(\Lambda^{\beta-2} \odot \mathrm{V}\right) \mathrm{H}^\top}{\Lambda^{\beta-1} \mathrm{H}^\top} \tag{2}$$

where $\odot$ is the Hadamard (element-wise) product, matrix exponentials are elementwise, and α weights coefficient sparsity against reconstruction error. To remove the scaling indeterminacy between W and H, the dictionary atoms are typically normalized after each update, and their coefficients adjusted accordingly.

In the case of stereo audio signals we study here, the left and right input spectrograms may be concatenated in time prior to learning, i.e. $\mathrm{V}_{ft} = [\mathrm{V}_{lft}|\mathrm{V}_{rft}]$, where the resulting coefficients are correspondingly $\mathrm{H}_{dt} = [\mathrm{H}_{ldt}|\mathrm{H}_{rdt}]$, and the dictionary remains as above.

2.2. GCC

Time differences of arrival (TDOA) of signals between pairs of spatially distributed sensors are used in a variety of sensor array applications for beamforming and localization. The Generalized Cross-Correlation (GCC) is a classic method for estimating TDOAs for an arbitrary set of frequencies [18, 19]. The GCC represents an *angular spectrogram* (see Figure 2a): a function of time-delay τ and time t, defined mathematically as,

$$G_{\tau t} = \sum_f \psi_{ft} \mathrm{V}_{lft} \mathrm{V}^*_{rft} e^{j2\pi f \tau} \tag{3}$$

where V_{lft} and V_{rft} are the left and right complex spectrograms, * is elementwise complex conjugation, and ψ_{ft} is a time-varying frequency-weighting function.

Among the most robust localization algorithms in the presence of interfering sounds and reverberation is the GCC Phase Transform (GCC-PHAT) [20], for which the frequency-weighting function is the inverse product of the left and right magnitude spectrograms,

$$\mathrm{G}^{\mathrm{PHAT}}_{\tau t} = \sum_f \frac{\mathrm{V}_{lft} \mathrm{V}^*_{rft}}{|\mathrm{V}_{lft}|\,|\mathrm{V}_{rft}|} e^{j2\pi f \tau} \tag{4}$$

The angular spectrogram is then pooled over time, yielding a summary angular spectrum, with the locations of the highest peaks then corresponding to the source TDOA estimates (see Figure 2b). The number of sources may be specified a priori, or estimated, for example by performing a k-means clustering of the local maxima amplitudes with $k = 2$. For small microphone separations, a nonlinearity must be applied to compensate the wide lobes of the resulting GCC,

$$\mathrm{G}^{\mathrm{PHAT\cdot NL}}_{\tau t} = 1 - \tanh\left(\gamma \sqrt{1 - \mathrm{G}^{\mathrm{PHAT}}_{\tau t}}\right) \tag{5}$$

where $\gamma = 2$ is shown to work well in practice [20, 21].

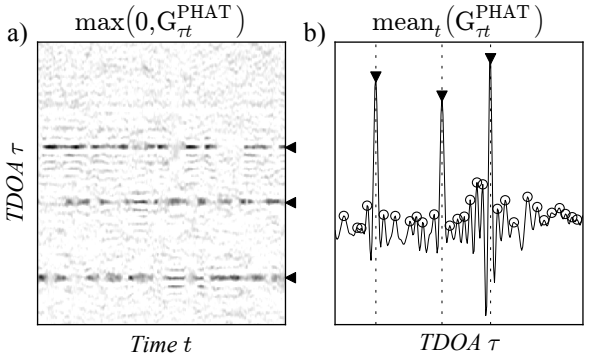

Figure 2: Source localization with GCC-PHAT for a 2 second mixture of 3 speakers. a) The GCC-PHAT angular spectrogram, rectified for clarity. The intermittent horizontal traces correspond to energy from the stationary speakers. b) The time-averaged GCC-PHAT angular spectrum. Source TDOA estimates τ_s are highlighted with dotted lines and triangle markers.

3. GCC-NMF Blind Speech Separation

In this section, we present the GCC-NMF separation algorithm. We first combine NMF and GCC to provide spatial information of individual dictionary atoms over time. Atoms are subsequently grouped into sources according to their spatial origin, with each group then reconstructed independently.

3.1. Combining GCC and NMF

We begin by defining a set of GCC frequency-weighting functions ψ_{dft}^{NMF} from the normalized NMF dictionary atoms,

$$\psi_{dft}^{\text{NMF}} = \frac{1}{|\text{V}_{lft}|\,|\text{V}_{rft}|}\frac{\text{W}_{fd}}{\sum_f \text{W}_{fd}} \tag{6}$$

constructed such that for a given atom d, frequencies are weighted according to their relative prominence. GCC-NMF is then the resulting set of atom-specific angular spectrograms,

$$\text{G}_{d\tau t}^{\text{NMF}} = \sum_f \psi_{dft}^{\text{NMF}} \text{V}_{lft} \text{V}_{rft}^{*} e^{j2\pi f\tau} \tag{7}$$

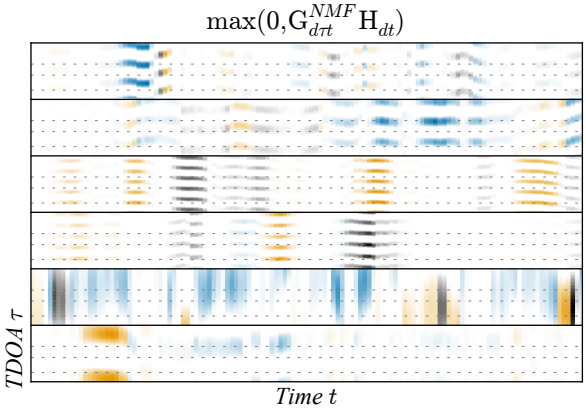

Figure 3: Example GCC-NMF angular spectrograms $\text{G}_{d\tau t}^{\text{NMF}}$. For clarity of presentation, angular spectrograms are multiplied by their corresponding coefficients, such that only active periods are shown, then rectified. Source-specific colors indicate to which of three sources atoms are associated over time, according to the procedure described in Section 3.2.

3.2. Coefficient Masking

The GCC-NMF angular spectrograms are used to associate each dictionary atom at each time with a single source s. Source TDOAs τ_s are first estimated using GCC-PHAT as described in Section 2.2. For each time t, dictionary atoms are then attributed to the source for which $\text{G}_{d\tau_s t}^{\text{NMF}}$ is highest. This defines a set of source-specific binary coefficient masks,

$$\text{M}_{sdt} = \begin{cases} 1 & \text{if } s = \text{argmax}_s \text{G}_{d\tau_s t}^{\text{NMF}} \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

that are multiplied with the mixture coefficients element-wise to create masked coefficients for each source.

3.3. Source Reconstruction

Source reconstruction is achieved by performing the inverse NMF and time-frequency functions using the source-specific masked coefficients,

$$\hat{\text{V}}_{scft} = \text{W}_{fd}\,(\text{M}_{sdt} \odot \text{H}_{cdt}) \tag{9}$$

$$\hat{\text{X}}_{scn} = \text{STFT}^{-1}\left(\hat{\text{V}}_{scft} \angle \text{V}_{cft}\right) \tag{10}$$

where $\hat{\text{V}}_{scft}$ are the source spectrograms estimates, $\hat{\text{X}}_{scn}$ are the time-domain source estimates, c indexes the stereo channels, and t and n index time in the frequency and time domains. Note that $\hat{\text{V}}_{scft}$ are combined with the mixture spectrogram phase prior to inverting the time-frequency transform.

3.4. GCC-NMF Separation System

We present a block diagram for the separation system in Figure 4, followed by a description of the system variables in Table 1.

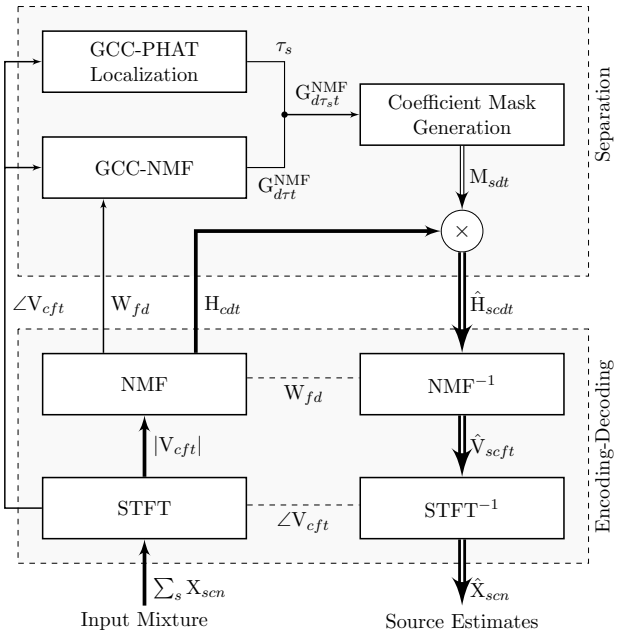

Figure 4: GCC-NMF source separation system, see Table 1 for variable descriptions. The separation system starts with an encoding-decoding block, consisting of STFT and NMF. A coefficient-masking block then interrupts the encoding-decoding process, resulting in an encoding-separation-decoding architecture. Bold arrows emphasize the encoding-decoding process, while double arrows highlight source-specific signals.

s	Source index	$\sum_s \text{X}_{scn}$	Input mixture
c	Channel index	V_{cft}	STFT mixture
n	Time index (input)	$\lvert\text{V}_{cft}\rvert$	STFT magnitude
t	Time index (STFT)	$\angle\text{V}_{cft}$	STFT phase
f	Frequency index	W_{fd}	Dictionary atoms
d	Atom index	H_{cdt}	Atom coefficients
τ_s	Source TDOAs	M_{sdt}	Coefficient masks
X_{scn}	Source signals	$\hat{\text{H}}_{scdt}$	Masked coefficients
$\hat{\text{X}}_{scn}$	Source estimates	$\text{G}_{d\tau t}^{\text{NMF}}$	GCC-NMF

Table 1: Variable descriptions. Subscripts index dimensions of multidimensional variables, lowercase symbols used as indexes.

4. Speech Separation Experiments

Experiments are performed using the SiSEC *dev1* live speech recordings dataset, constructed as "static sources played through loudspeakers in a meeting room, recorded one at a time" [6], consisting of sixteen 10-second mixtures of 3 and 4 female and male speakers, with 5 cm and 1 m microphone separations, and 180 ms and 250 ms reverberation times. Complex spectrograms are generated from the 16 kHz mixture signals with a short-time Fourier transform (STFT) using a 1024-sample Hann window (64 ms), and 16-sample hop size (1 ms). Default NMF parameters are set to 1024 dictionary atoms, 100 iterations, sparsity α = 0, cost function $\beta = 1$. The GCC nonlinearity is used for 5 cm microphone separations with $\gamma = 3$.

Separation performance is quantified in terms of overall quality, target fidelity, interference suppression, and lack of artifacts using two open-source toolkits: PEASS [23] and BSS Eval [24]. While the latter measures traditional signal-to-noise ratio (SNR), the former is a perceptually-motivated approach whose scores better correlate with human assessments.

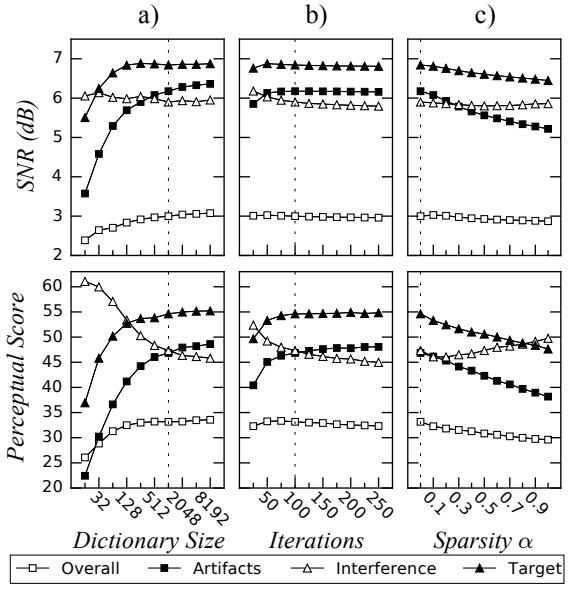

Figure 5: Effect of NMF parameters on separation performance as measured with BSS Eval (top) and PEASS (bottom) scores. Subplots depict average scores over the 56 sources of the *dev1* SiSEC dataset for varying a) NMF dictionary size b) number of NMF iterations c) sparsity coefficient α. Default values are shown with dashed-lines.

4.1. Effects of NMF Parameters

In Figure 5, we present the effects on separation performance of NMF dictionary size, number of iterations, and the sparsity coefficient α. For both SNR and perceptual measures, increasing dictionary size results in increased target fidelity, lack of artifacts and overall score, saturating for larger dictionaries. While SNR measures suggest interference suppression is independent of dictionary size, the perceptual score shows a clear decrease with increasing dictionary size. Dictionary size therefore offers control of the tradeoff between interference suppression and overall, target, and artifact scores. The number of iterations has a similar, though less drastic effect, while increasing coefficient sparsity has the opposite effect: target, artifact, and overall scores decrease with increasing sparsity, while interference suppression increases. We hypothesize that increasing sparsity pushes NMF to learn atoms that are less well-suited for separation with GCC-NMF, but leave a proper study to future work.

4.2. Comparison with Model-based Approaches

In Table 2, we compare GCC-NMF with other NMF-based speech separation algorithms, in addition to an oracle baseline. FASST is a flexible, open-source, model-based approach combining NMF with a spatial covariance mixing model [22]. In the purely unsupervised setting, it is too sensitive to initialization and lacks robustness. For FASST-init, we therefore use an oracle mixture initialization procedure, resulting in significantly better performance, however requiring prior mixing model information. Note that while this semi-supervised approach outperforms GCC-NMF according to the BSS Eval metrics, GCC-NMF results in significantly better overall, target and interference based PEASS scores, with a cost of increased artifacts. We also present results of other semi-supervised and constrained dictionary algorithms from the SiSEC campaign as presented in [5, 6, 25], with GCC-NMF performing favourably despite being a purely unsupervised approach.

5. Conclusion

We have introduced a new approach to combining spatial information with NMF for unsupervised speech separation. The GCC method of source localization is used to localize individual dictionary atoms over time, such that they may be grouped into sources based on their spatial origin. The resulting GCC-NMF blind speech separation approach outperforms the unsupervised model-based spatial covariance approach, and compares favourably to semi-supervised and constrained NMF-based approaches that leverage prior knowledge and information. While the simple combination of GCC and NMF performs well, more complex NMF models and other feature learning approaches are being studied. Finally, since coefficient masking is performed on a frame-by-frame basis, real-time speech separation is also theoretically possible, provided dictionary learning is performed offline.

Acknowledgements: ACELP/CEGI, NSERC discovery grant.

	PEASS				BSS Eval			
	OPS	TPS	IPS	APS	SDR	ISR	SIR	SAR
GCC-NMF	33.16±5.34	54.62±10.16	47.33±12.05	46.85±7.68	3.00±1.16	6.84±2.65	5.90±4.33	6.18±1.30
FASST [22]	16.98±3.57	32.34±7.21	19.03±6.93	47.66±4.77	-1.70±0.79	2.15±1.01	-3.85±2.05	3.39±1.09
FASST-Init [22] [1]	26.28±6.85	43.19±13.64	38.02±11.37	52.11±16.80	3.61±1.31	7.77±2.36	6.33±1.94	7.89±1.93
Ozerov [11] [2,3]	34.87±6.73	57.16±6.75	47.40±8.60	56.55±8.84	4.07±2.55	8.99±3.96	7.80±4.73	7.37±2.63
Adiloglu [13] [2]	34.59±6.05	58.11±5.13	45.24±8.75	58.80±7.75	3.61±1.96	8.35±3.32	7.03±4.18	7.49±2.38
IBM [6] [4]	38.40±7.94	66.48±4.58	73.26±1.91	34.37±11.62	8.99±1.26	17.51±1.82	19.33±1.94	9.31±1.39

Table 2: Mean PEASS and BSSEval separation scores ± standard deviation, taken over the SiSEC dev1 live speech recordings dataset. OPS, TPS, IPS, APS: Oracle, target, interference, artifact-related perceptual scores. SDR, ISR, SIR, SAR: signal to distortion, source image to spatial distortion, source to interference, source to artifacts ratios. [1] Oracle mixture initialization. [2] Constrained multilayer NMF. [3] Condition-specific settings. [4] Oracle baseline: ideal binary mask.

6. References

[1] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the acoustical society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[2] S. Haykin and Z. Chen, "The cocktail party problem," *Neural computation*, vol. 17, no. 9, pp. 1875–1902, 2005.

[3] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[4] ——, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2001, pp. 556–562.

[5] N. Ono, Z. Koldovsky, S. Miyabe, and N. Ito, "The 2013 signal separation evaluation campaign," *Proc. International Workshop on Machine Learning for Signal Processing*, pp. 1–6, 2013.

[6] N. Ono, Z. Rafii, D. Kitamura, N. Ito, and A. Liutkus, "The 2015 signal separation evaluation campaign," in *Latent Variable Analysis and Signal Separation*. Springer, 2015, pp. 387–395.

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

[8] E. Vincent, N. Bertin, R. Gribonval, and F. Bimbot, "From blind to guided audio source separation: How models and side information can improve the separation of sound," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 107–115, 2014.

[9] B. Wang and M. D. Plumbley, "Musical audio stream separation by non-negative matrix factorization," in *Proc. DMRN summer conf*, 2005, pp. 23–24.

[10] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *Spoken Language Proceesing, ISCA International Conference on (INTERSPEECH)*, 2006.

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

[12] S. Arberet, A. Ozerov, N. Q. Duong, E. Vincent, R. Gribonval, F. Bimbot, and P. Vandergheynst, "Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation," in *Information Sciences Signal Processing and their Applications (ISSPA), 2010 10th International Conference on*. IEEE, 2010, pp. 1–4.

[13] K. Adiloglu and E. Vincent, "Variational bayesian inference for source separation and robust feature extraction," Ph.D. dissertation, INRIA, 2012.

[14] J. Traa, P. Smaragdis, N. D. Stein, and D. Wingate, "Directional NMF for joint source localization and separation," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2015 IEEE Workshop on*. IEEE, 2015, pp. 1–5.

[15] T. T. Vu, B. Bigot, and E. S. Chng, "Speech enhancement using beamforming and non negative matrix factorization for robust speech recognition in the CHiME-3 challenge," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 423–429.

[16] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the β-divergence," *Neural Computation*, vol. 23, no. 9, pp. 2421–2456, 2011.

[17] J. Roux, F. Weninger, and J. Hershey, "Sparse NMF–half-baked or well done?" *Mitsubishi Electric Research Laboratories Technical Report*, 2015.

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

[19] X. Anguera, "Robust speaker diarization for meetings," Ph.D. dissertation, Universitat Politècnica de Catalunya, 2006.

[20] C. Blandin, A. Ozerov, and E. Vincent, "Multi-source TDOA estimation in reverberant audio using angular spectra and clustering," *Signal Processing*, vol. 92, no. 8, pp. 1950–1960, 2012.

[21] N. Q. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 7, pp. 1830–1840, 2010.

[22] Y. Salaün, E. Vincent, N. Bertin, N. Souviraa-Labastie, X. Jaureguiberry, D. T. Tran, and F. Bimbot, "The flexible audio source separation toolbox version 2.0," in *ICASSP*, 2014.

[23] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 7, pp. 2046–2057, 2011.

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

[25] S. Araki, F. Nesta, E. Vincent, Z. Koldovskỳ, G. Nolte, A. Ziehe, and A. Benichoux, "The 2011 signal separation evaluation campaign (SiSEC2011):-audio source separation," in *Latent Variable Analysis and Signal Separation*. Springer, 2012, pp. 414–422.

[26] S. Mirzaei, Y. Norouzi *et al.*, "Blind audio source separation of stereo mixtures using bayesian non-negative matrix factorization," in *Signal Processing Conference (EUSIPCO), 2014 Proceedings of the 22nd European*. IEEE, 2014, pp. 621–625.

[27] B. Cauchi, T. Gerkmann, S. Doclo, P. A. Naylor, and S. Goetze, "Spectrally and spatially informed noise suppression using beamforming and convolutive NMF," in *Audio Engineering Society Conference: 60th International Conference: DREAMS (Dereverberation and Reverberation of Audio, Music, and Speech)*. Audio Engineering Society, 2016.

[28] J. Thiemann and E. Vincent, "An experimental comparison of source separation and beamforming techniques for microphone array signal enhancement," in *Machine Learning for Signal Processing (MLSP), 2013 IEEE International Workshop on*. IEEE, 2013, pp. 1–5.

[29] G. Dekkers, T. van Waterschoot, B. Vanrumste, B. Van Den Broeck, J. F. Gemmeke, P. Karsmakers *et al.*, "A multi-channel speech enhancement framework for robust NMF-based speech recognition for speech-impaired users," in *INTERSPEECH 2015: proceedings*, no. accepted. ISCA, 2015.

[30] S. Lee, S. H. Park, and K.-M. Sung, "Geometric source separation method of audio signals based on beamforming and NMF," in *Audio Engineering Society Conference: 42nd International Conference: Semantic Audio*. Audio Engineering Society, 2011.

[31] C. Joder, F. Weninger, F. Eyben, D. Virette, and B. Schuller, "Real-time speech separation by semi-supervised nonnegative matrix factorization," in *Latent Variable Analysis and Signal Separation*. Springer, 2012, pp. 322–329.

[32] N. D. Stein, "Nonnegative tensor factorization for directional blind audio source separation," *arXiv preprint arXiv:1411.5010*, 2014. [Online]. Available: http://arxiv.org/abs/1411.5010

[33] D. Fitzgerald, A. Liutkus, and R. Badeau, "PROJET-spatial audio separation using projections," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 36–40.

[34] 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 *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2011, pp. 325–328.

[35] M. I. Mandel, "Binaural model-based source separation and localization," Ph.D. dissertation, Columbia University, 2010.