A SOURCE/FILTER MODEL WITH ADAPTIVE CONSTRAINTS FOR NMF-BASED SPEECH SEPARATION

Damien Bouvier, *Nicolas Obin*, *Marco Liuni*, *Axel Roebel*

IRCAM - UMR STMS IRCAM-CNRS-UPMC
Paris, France

ABSTRACT

This paper introduces a constrained source/filter model for semi-supervised speech separation based on non-negative matrix factorization (NMF). The objective is to inform NMF with prior knowledge about speech, providing a physically meaningful speech separation. To do so, a source/filter model (indicated as Instantaneous Mixture Model or IMM) is integrated in the NMF. Furthermore, constraints are added to the IMM-NMF, in order to control the NMF behaviour during separation, and to enforce its physical meaning. In particular, a speech specific constraint - based on the source/filter coherence of speech - and a method for the automatic adaptation of constraints' weights during separation are presented. Also, the proposed source/filter model is semi-supervised: during training, one filter basis is estimated for each phoneme of a speaker; during separation, the estimated filter bases are then used in the constrained source/filter model. An experimental evaluation for speech separation was conducted on the TIMIT speakers database mixed with various environmental background noises from the QUT-NOISE database. This evaluation showed that the use of adaptive constraints increases the performance of the source/filter model for speaker-dependent speech separation, and compares favorably to fully-supervised speech separation.
Index Terms: speech separation, non-negative matrix factorization, source/filter model, constraints.

1. INTRODUCTION

Speech separation consists in the separation of a speech signal from a background environment, referred as *noise*, which is defined as everything but the speaker of interest (i.e., environmental sounds such as background non-speech sounds or background speech). Speech separation is essential for further speech processing in real speech technologies, such as speech recognition, speaker recognition, speaker localization, and audio multi-media technologies for speech extraction and remixing. Audio source separation methods have been recently introduced for speech separation, in which the audio signal is described as the sum of two sources: a speech signal and a background noise signal. In particular, the non-negative matrix factorization (NMF) of an audio signal is extremely popular for source separation, and is widely used in recent times for speech separation [1, 2, 3, 4, 5, 6]. In the original formulation, the NMF decomposition of an audio signal is strictly unsupervised [7]. In the last decade, audio and speech separation has massively converged to informed audio source separation, in order to provide prior knowledge about the audio sources to be separated [8]. In the context of speech separation, two main trends co-exist: semi-supervised speech separation uses prior knowledge about speech only [5, 6], and supervised speech separation adds prior knowledge about the background environment [3, 9]. This latter case remains extremely limited when the background environment is unknown, which is the case of most real-world applications.

In this context, semi-supervised speech separation is the most common approach. The main advantage of semi-supervised speech separations is that robust prior knowledge about speech can be exploited, while the integration of prior knowledge about the background environment is clearly not realistic, regarding the extreme variability of the background environment. For semi-supervised speech separation, the experience into speech processing and speech recognition can be exploited: a source/filter model can be used to inform NMF-based separation [1, 2]. Also, a universal speech model (USM) has been proposed for speaker-independent speech separation [3], in which a speaker can be represented by a combination of the most similar speakers bases. A real-time implementation of the USM has been recently proposed for on-line background noise estimation [4]. Furthermore, hidden Markov models (HMM) has been added to NMF speech separation, in order to construct a language model [5] and to use prior text information for speech separation [6]. Finally, deep neural networks (DNN) has been successfully introduced for speech separation [10], and deep-NMF [11] has been proposed in order to integrate the advantages of DNN within the NMF framework.

In this paper, we propose a constrained source/filter model for semi-supervised and physically-motivated NMF-based speech separation. To do so, a source/filter NMF model is described in Section 2. Then, constraints are added to this model in Section 3, in which a specific speech constraint and a method for the automatic adaptation of the constraints' weights during separation are proposed. An experiment is conducted in Section 4 in order to explore the use of constraints within the source/filter model, with comparison to state-of-the-art speech separation methods.

2. NMF AND SOURCE/FILTER MODEL

2.1. NMF principle

Let $\boldsymbol{V}$ denotes our observation matrix, with only non-negative coefficients (for audio, usually the STFT magnitude of the observed mixture signal); the NMF consists of finding the best approximation given a chosen cost $\mathcal{C}$:

$$\boldsymbol{V} \simeq \boldsymbol{W}\boldsymbol{H} \tag{1}$$

where $\boldsymbol{W}$ and $\boldsymbol{H}$ also contains only non-negative coefficients. $\boldsymbol{W}$ represents a dictionary matrix and $\boldsymbol{H}$ is the activation matrix (it can be seen as the gains of the projection of $\boldsymbol{V}$ onto the space defined by $\boldsymbol{W}$). Afterwards, source separation can be made using Wiener filters [12].

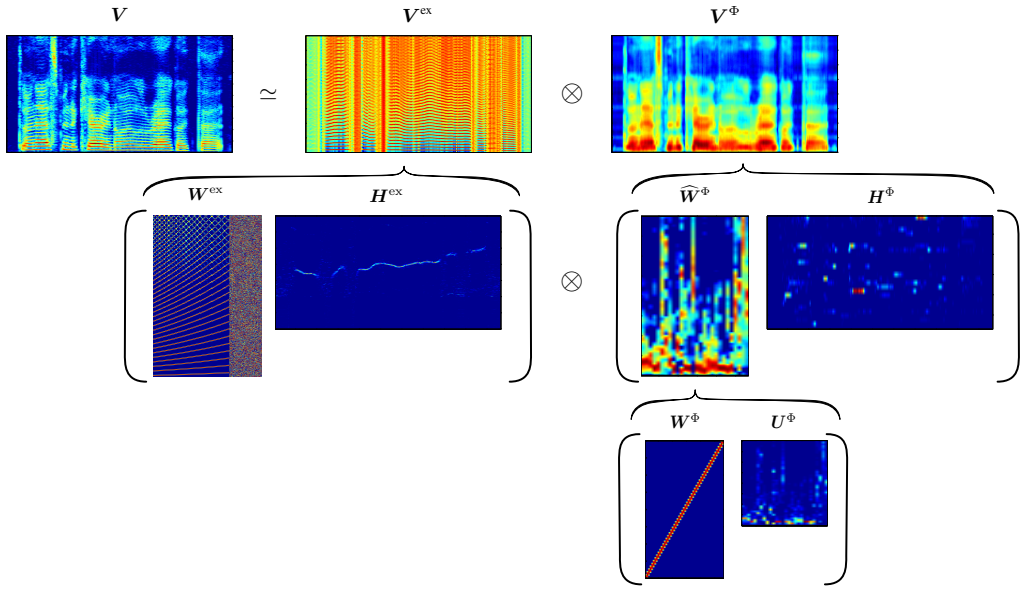


Fig. 1. Illustration of the source/filter decomposition for the IMM-NMF described in (4).

In audio, usual costs are Kullback-Leiber (KL) and Itakura-Saito (IS) divergence [13], which are both limit cases of the β-divergence (respectively for $\beta = 1$ and $\beta = 0$). In this paper, we use the IS divergence for its scale-invariance, which is an interesting property for audio signals:

$$\mathcal{C} = D_{IS}\left(\boldsymbol{V}|\boldsymbol{W}\boldsymbol{H}\right) = \sum_{f}^{F}\sum_{n}^{N} d_{IS}\left(\boldsymbol{V}_{fn}|\left(\boldsymbol{W}\boldsymbol{H}\right)_{fn}\right) \quad (2)$$

with $d_{IS}(x|y) = \frac{x}{y} - \log\frac{x}{y} - 1$. Moreover, we use the artificial noise floor introduced in [14] in order to control the noise robustness of the IS-NMF.

The solution of the NMF problem, using β-divergence, can be efficiently obtained by applying an iterative algorithm, derived from a gradient step descent technique [13]. The i-th iteration is based on the application, on both $\boldsymbol{W}$ and $\boldsymbol{H}$, of the following multiplicative rule:

$$\boldsymbol{\Theta}^{(i+1)} \longleftarrow \boldsymbol{\Theta}^{(i)} \otimes \frac{\nabla^{-}_{\boldsymbol{\Theta}^{(i)}}\mathcal{C}}{\nabla^{+}_{\boldsymbol{\Theta}^{(i)}}\mathcal{C}} \quad (3)$$

where $\boldsymbol{\Theta}$ represents either $\boldsymbol{W}$ or $\boldsymbol{H}$, $\nabla^{+}_{\boldsymbol{\Theta}^{(i)}}$ and$\nabla^{-}_{\boldsymbol{\Theta}^{(i)}}$ are the positive and negative parts of the gradient of the cost $\mathcal{C}$ with respect to $\boldsymbol{\Theta}^{(i)}$, $\otimes$ denotes the Hadamard product, and the division is pointwise.

2.2. A source/filter-NMF for speech

In order to provide an explicit representation of speech, we use the source/filter model for NMF proposed in [2], in which a training step is added to estimate the vocal filters. The NMF source/filter decomposition of the STFT magnitude $\boldsymbol{V}^{\text{S}}$ of a speech signal can be expressed as:

$$\begin{aligned}\boldsymbol{V}^{\text{S}} &= \boldsymbol{V}^{\text{ex}} \otimes \boldsymbol{V}^{\Phi} \\ &\simeq \underbrace{\left(\boldsymbol{W}^{\text{ex}}\boldsymbol{H}^{\text{ex}}\right)}_{\text{excitation}} \otimes \underbrace{\left(\widehat{\boldsymbol{W}}^{\Phi}\boldsymbol{H}^{\Phi}\right)}_{\text{filter}}\end{aligned} \quad (4)$$

where $\boldsymbol{V}^{\text{ex}}$ and $\boldsymbol{V}^{\Phi}$ are respectively the magnitude STFT of the excitation part and the filter part, $\boldsymbol{W}^{\text{ex}}$ and $\boldsymbol{H}^{\text{ex}}$ are the standard NMF decomposition for the speech excitations $\boldsymbol{V}^{\text{ex}}$ (where $\boldsymbol{W}^{\text{ex}}$ is a fixed dictionary, including periodic and noisy basis), and $\widehat{\boldsymbol{W}}^{\Phi}$ and $\boldsymbol{H}^{\Phi}$ are the standard NMF decomposition for the speech filters $\boldsymbol{V}^{\Phi}$. In order to ensure the smoothness of the speech filters $\widehat{\boldsymbol{W}}^{\Phi}$, we further decompose $\widehat{\boldsymbol{W}}^{\Phi}$ as the product $\boldsymbol{W}^{\Phi}\boldsymbol{U}^{\Phi}$, where $\boldsymbol{W}^{\Phi}$ is a fixed dictionary of smooth "atomic" filters (here, Hann windows) and $\boldsymbol{U}^{\Phi}$ is the coefficient matrix linearly combining those elementary filters to form a speech filter. Figure 1 illustrates the architecture of the source/filter model for NMF.

For speech separation, the observed signal $\boldsymbol{V}$ is assumed to be a mixture of a speech signal $\boldsymbol{V}^{S}$ and a background noise signal $\boldsymbol{V}^{N}$. $\boldsymbol{V}$ can be approximated by $\widetilde{\boldsymbol{V}}$ as follows:

$$\boldsymbol{V} \simeq \widetilde{\boldsymbol{V}} = \left(\boldsymbol{W}^{\text{ex}}\boldsymbol{H}^{\text{ex}}\right) \otimes \left(\boldsymbol{W}^{\Phi}\widehat{\boldsymbol{U}}^{\Phi}\boldsymbol{H}^{\Phi}\right) + \boldsymbol{W}^{N}\boldsymbol{H}^{N} \quad (5)$$

in which the background noise signal $\boldsymbol{V}^{N}$ is expressed using a standard NMF decomposition. Following the denomination used in [2], we will refer to this mixture decomposition as the "Instantaneous Mixture Model for NMF" (IMM-NMF).

2.3. Semi-supervision of the IMM-NMF

In the IMM-NMF, the speech filters $\widehat{\boldsymbol{W}}^{\Phi}$ of a speaker are explicitly represented by the coefficients matrix $\boldsymbol{U}^{\Phi}$. In [2], this matrix was directly estimated from the observed signal $\boldsymbol{V}$, thus fully unsupervised. Here, we propose to estimate the speech filters from clean speech signals of a speaker. To do so, the speech filter matrice $\boldsymbol{V}^{\Phi}$ is first estimated by a spectral envelope estimation algorithm [15], which is then approximated by the NMF filter decomposition:

$$\boldsymbol{V}^{\Phi} \simeq \boldsymbol{W}^{\Phi}\boldsymbol{U}^{\Phi}\boldsymbol{H}^{\Phi} \quad (6)$$

Furthermore, we used phonetic information in order to train phonemes separately and to have one basis for each speech filter (i.e. each phoneme).

3. SOURCE/FILTER MODEL UNDER CONSTRAINTS

3.1. Constrained-NMF

The main objective of this work is to inform NMF speech separation with a physical model of speech. For this purpose, we use constraints into the NMF to penalize solutions not respecting the speech model. Accordingly, the cost $\mathcal{C}$ is modified with the addition of the constraint penalty cost $\mathcal{P}$:

$$\mathcal{C} = D_{IS}\left(\boldsymbol{V}|\widetilde{\boldsymbol{V}}\right) + \mu\mathcal{P} \tag{7}$$

with μ a positive value determining the weight of the constraint. The new multiplicative update for the i-th iteration will be of the form :

$$\boldsymbol{\Theta}^{(i+1)} \longleftarrow \boldsymbol{\Theta}^{(i)} \otimes \frac{\nabla^{-}_{\boldsymbol{\Theta}^{(i)}} D_{IS} + \mu\nabla^{-}_{\boldsymbol{\Theta}^{(i)}}\mathcal{P}}{\nabla^{+}_{\boldsymbol{\Theta}^{(i)}} D_{IS} + \mu\nabla^{+}_{\boldsymbol{\Theta}^{(i)}}\mathcal{P}} \tag{8}$$

for any non-fixed matrix $\boldsymbol{\Theta}$ in the model. In our method, we use several constraints ans sum their values in order to obtain the total penalty cost $\mathcal{P}$.

3.2. State-of-the-art constraints

We used three constraints from the literature :

- the sparsity constraint described in [16] as the Column-Normalized $\ell 1$-norm, in order to promote the activation, at any given time, of a single filter basis and a single excitation basis.
- the normalized decorrelation constraint, based on the correlation measure proposed in [17], in order to penalize simultaneous activation between bases.
- a smoothness constraint proposed in [18] to prevent filter activation to jump from one phoneme to another between frames.

3.3. Source/filter coherence constraint

The first contribution of this paper consists in the elaboration of a speech-specific source/filter coherence constraint for the IMM-NMF. This constraint is based on the fact that all phonemes corresponds to a match between one excitation one corresponding filter:

- vocal filters corresponding to voiced phoneme will always be used simultaneously with a periodic excitation;
- vocal filters corresponding to unvoiced phoneme will always be used with a noisy excitation;

Because the IMM-NMF allows unreal combination leading to audible artefacts, we propose a source/filter coherence constraint which aims to avoid unrealistic combinations between excitation and filter (see Figure 2). This constraint is inspired by the normalized decorrelation constraint, and is expressed as follow:

$$\mathcal{P}_{\phi} = \sum_{\substack{k\in\text{periodic}\\ l\in\text{unvoiced}}} \frac{\left[\boldsymbol{H}^{\text{ex}}\boldsymbol{H}^{\Phi\,T}\right]_{kl}}{\|\boldsymbol{H}^{\text{ex}}_{k}\|_{\ell_2}\|\boldsymbol{H}^{\Phi}_{l}\|_{\ell_2}} + \sum_{\substack{k\in\text{noise}\\ l\in\text{voiced}}} \frac{\left[\boldsymbol{H}^{\text{ex}}\boldsymbol{H}^{\Phi\,T}\right]_{kl}}{\|\boldsymbol{H}^{\text{ex}}_{k}\|_{\ell_2}\|\boldsymbol{H}^{\Phi}_{l}\|_{\ell_2}} \tag{9}$$

The left term of the sum is a measure of the correlation between periodic excitation basis and filter basis corresponding to unvoiced phoneme, normalized by their power; the right term is the same measure, but between noisy excitation basis and filter basis corresponding to voiced phoneme.

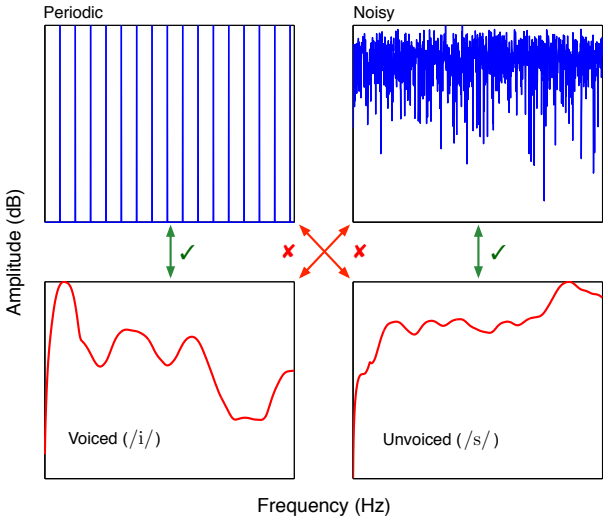


Fig. 2. Schematic example of realistic (green check marks) and unrealistic (red crosses) combinations of speech excitations (top) and speech filters (bottom).

3.4. Adaptation of constraints weights

One of the main limitations of constrained-NMF is the difficulty for finding a good constraint weight for speech separation. A small weight would conduct to a small or null effect of the constraint, while a strong weight will over-consider the constraint over the reconstruction cost, thus lead to a wrong solution depending on the initialization (generally random). The second contribution of this paper consists in adapting the constraint's weight at each iteration during speech separation: from small to strong depending on the evolution of the reconstruction (i.e. the evolution rate of the β-divergence value). At the i-th iteration, the constraint's weight is updated as:

$$\mu^{(i)} = \mu_{max}\frac{D_{IS}\left(\boldsymbol{V}|\widetilde{\boldsymbol{V}}^{(i-2)}\right) - D_{IS}\left(\boldsymbol{V}|\widetilde{\boldsymbol{V}}^{(i-1)}\right)}{D_{IS}\left(\boldsymbol{V}|\widetilde{\boldsymbol{V}}^{(i-2)}\right)} \tag{10}$$

where μ is initialized at 0 for the first two iterations, and after varies in the interval $[0\ \mu_{max}]$, where μ_{max} is a chosen value. The stronger the β-divergence diminishes, the smaller the constraint; the smaller it diminishes, the higher the constraint. Figure 3 shows an example of the effect of this adaptive method on costs evolution.

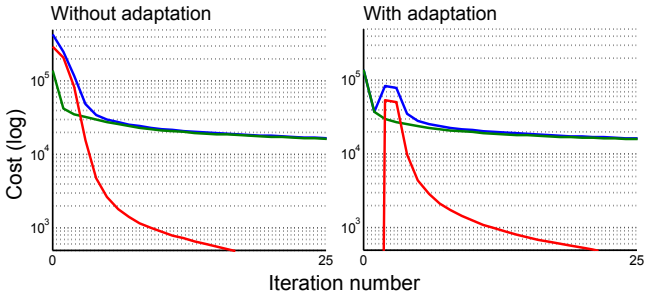


Fig. 3. Evolution of the reconstruction cost (in green), constraints cost (in red) and total cost (in blue) with (right graphic) and without (left graphic) the weight adaptation method, in function of the iteration number.

		Algorithms								
		References		Proposed						
		ASNA [19]	V-IMM [2]	# 1	# 2	# 3	# 4	# 5	# 6	# 7
Training	Speech	✓		✓	✓	✓	✓	✓	✓	✓
	Noise	✓								
IMM-NMF Constraints			✓	✓	✓ SoA	✓ coherence	✓ all	✓ SoA	✓ coherence	✓ all
Adaptation					Without			With		
-6dB	SDR	5.8	4.4	4.0	4.1	5.0	5.2	4.1	5.2	**5.4**
	PESQ	2.00	1.22	1.91	1.91	1.94	1.92	1.91	**2.01**	**2.01**
$+0$dB	SDR	10.7	7.8	9.1	9.2	9.0	8.9	9.2	**9.8**	**9.8**
	PESQ	2.44	1.54	2.30	2.30	2.24	2.23	2.30	2.34	**2.35**
$+6$dB	SDR	15.0	9.7	**13.0**	12.8	11.1	10.9	**13.0**	12.8	12.9
	PESQ	2.85	1.82	**2.62**	2.61	2.46	2.44	**2.62**	2.59	**2.62**
Mean	SDR	10.5	7.3	8.7	8.7	8.4	8.3	8.7	9.3	**9.4**
	PESQ	2.43	1.52	2.28	2.27	2.21	2.20	2.28	2.31	**2.33**

Table 1. Results from the experimental evaluation. *SoA* refers to the state-of-the-art constraints (decorrelation, sparsity and smoothness), *coherence* to the proposed source/filter coherence constraint and *adaptation* to the adaptive weight method.

4. EXPERIMENT

4.1. Experimental setups

An experiment was conducted to evaluate the performance of the semi-supervised and constrained source/filter model for speech separation. The benchmark includes: the semi-supervised source/filter model, with variants on the use of the constraints (with/without constraints, state-of-the-art constraints vs. source/filter coherence constraint, and with/without the constraint adaptation), with comparison to state-of-the-art unsupervised V-IMM source/filter model [2] (originally developed for singing voice / music separation), and the supervised ASNA algorithm [19] (see Table 1 for details).

The database used for the experiment is a mix of the TIMIT speech database for clean speech [20] and the QUT-NOISE database for environmental background noises [21]. We used 20 TIMIT speakers (10 women and 10 men), with each 10 sentences: 2 sentences, shared among all the speakers, were used for training, and the 8 remaining sentences, different for all speakers, were used as the test set for speech separation. We mixed those 160 test sentences with 4 different background noises from the QUT-NOISE database (city street, home kitchen, car window, cafe) and white noise, using 3 signal-to-noise-ratio (SNR) (-6 dB, 0 dB, $+6$ dB), resulting in $2,400$ mixture signals. For the training, the 2 shared sentences were used for each speaker for the semi-supervised and the supervised algorithms, and one 5 s. extract of the background noise (different of the one used for mixing) was used for the supervised algorithm. The performance of the speech separation was measured based on the signal-to-distortion ratio (SDR) [22] (in dB), and the perceptive evaluation of speech quality (PESQ) [23].

All benchmark speech separation algorithms were based on the STFT magnitude of the audio signal, using a Hamming window of 64 ms and a hop size of 32 ms. For training, the filterbank dictionary $\boldsymbol{W}^{\Phi}$ was created with 50 Hann windows linearly spaced from 0 to 8000 Hz, and the excitation dictionary $\boldsymbol{W}^{\text{ex}}$ was created with 250 periodic bases (spanning every twentieth of tone between 80 and 350 Hz) and 100 white noise bases. For testing, the maximum number of iteration of the NMF was set to 100, and we used the IS divergence with a noise floor of -60dB (see [14] for details). Various constraint weights (from 10^{-2} to 10^{3}) and number of background noise bases (from 5 to 100, used for all algorithms) were tested.

4.2. Results and Discussion

Table 1 summarizes the scores obtained for the benchmark algorithms, optimized for the state-of-the-art algorithms, and sharing the same optimal setup for all of the proposed algorithms. Firstly, the semi-supervised source/filter algorithm (# 1) improves the performance over a standard unsupervised source/filter, which naturally confirms the importance of training the filter dictionary for speech separation. Secondly, the use of constraints in the source/filter model without adaptation does not improves speech separation (# 1 vs. # 2, # 3, and # 4) on the one side. On the other side, the use of constraints with adaptation (# 5, # 6, and # 7) substantially improves speech separation. This shows the importance of the adaptation during speech separation, by gradually increasing the importance of the constraints depending on the convergence of the speech separation. This is especially true when using all constraints together (# 7). A comparison of the constraints reveals that the state-of-the-art constraints (decorrelation, sparsity and smoothness) have a small effect (# 5) whereas the source/filter coherence constraint provides a strong effect, and is more efficient for a high SNR (# 6). Finally, the semi-supervised constrained source/filter algorithm shows encouraging performance, compared to state-of-the-art algorithms. The semi-supervised algorithm stands in between unsupervised algorithm (V-IMM) and the supervised algorithm (ASNA). In particular, the semi-supervised algorithm is close to the supervised algorithm, without any prior knowledge about the nature the noise environment. This proves the importance of using prior knowledge on speech for informed speech separation.

5. CONCLUSION

In this paper, we presented a semi-supervised method for speech separation, based on a constrained source/filter model for NMF-based speech separation, with the add of a speech specific constraint, and the adaptive weighting of constraints during separation. An experimental validation proved the efficiency of the constrains for speech separation, and beyond indicates the importance of prior knowledge about speech and physically-motivated speech separation. Further research will focus on the integration of a source/filter model for text-informed speech separation [5, 6], and speaker-independent speech separation (Universal Speech Model [3]), and on the unsupervised estimation of the background noise [4, 14].

6. REFERENCES

[1] T. Virtanen and A. Klapuri, "Analysis of polyphonic audio using source-filter model and non-negative matrix factorization," in *Advances in models for acoustic processing, neural information processing systems workshop*, 2006.

[2] J.-L. Durrieu, A. Ozerov, C. Févotte, G. Richard, and B. David, "Main instrument separation from stereophonic audio signals using a source/filter model," in *European Signal Processing Conference (EUSIPCO)*, 2009, pp. 15–19. [Online]. Available: http://www.durrieu.ch/phd/eusipco09/

[3] D. L. Sun and G. J. Mysore, "Universal speech models for speaker independent single channel source separation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013, pp. 141–145.

[4] F. G. Germain and G. J. Mysore, "Speaker and noise independent online single-channel speech enhancement," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015, pp. 71–75.

[5] G. J. Mysore and P. Smaragdis, "A non-negative approach to language informed speech separation," in *Latent Variable Analysis and Signal Separation*. Springer, 2012, pp. 356–363.

[6] L. Le Magoarou, A. Ozerov, and N. Q. Duong, "Text-informed audio source separation. Example-based approach using non-negative matrix partial co-factorization," *Journal of Signal Processing Systems*, pp. 1–15, 2014.

[7] 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.

[8] A. Liutkus, J.-L. Durrieu, L. Daudet, and G. Richard, "An overview of informed audio source separation," in *International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, 2013, pp. 1–4.

[9] T. Virtanen, B. Raj, J. F. Gemmeke *et al.*, "Active-set newton algorithm for non-negative sparse coding of audio," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014, pp. 3092–3096. [Online]. Available: http://www.esat.kuleuven.be/psi/spraak/downloads/

[10] Y. Wang and D. Wang, "A Neural Network For Time-Domain Signal Reconstruction: Towards Improving The Perceptual Quality Of Supervised Speech Separation," Department of Computer Science and Engineering, The Ohio State University, Tech. Rep., 2014.

[11] J. Le Roux, J. R. Hershey, and F. Weninger, "Deep NMF for speech separation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015, pp. 66–70.

[12] L. Benaroya, F. Bimbot, and R. Gribonval, "Audio source separation with a single sensor," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 191–199, 2006.

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

[14] A. Roebel, J. Pons, M. Liuni, and M. Lagrange, "On automatic drum transcription using non-negative matrix deconvolution and Itakura-Saito divergence," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015, pp. 414–418.

[15] F. Villavicencio, A. Röbel, and X. Rodet, "Improving LPC spectral envelope extraction of voiced speech by true-envelope estimation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, 2006, pp. 869–872.

[16] C. Joder, F. Weninger, D. Virette, and B. Schuller, "A comparative study on sparsity penalties for NMF-based speech separation: Beyond Lp-norms," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013, pp. 858–862.

[17] S. Z. Li, X. W. Hou, H. Zhang, and Q. Cheng, "Learning spatially localized, parts-based representation," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2001, pp. I–207.

[18] 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, 2007.

[19] T. Virtanen, J. F. Gemmeke, and B. Raj, "Active-set Newton algorithm for overcomplete non-negative representations of audio," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 11, pp. 2277–2289, 2013.

[20] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: TIMIT and beyond," *Speech Communication*, vol. 9, no. 4, pp. 351–356, 1990.

[21] D. B. Dean, S. Sridharan, R. J. Vogt, and M. W. Mason, "The QUT-NOISE-TIMIT corpus for the evaluation of voice activity detection algorithms," *Proceedings of Interspeech 2010*, pp. 3110–3113, 2010.

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

[23] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, 2001, pp. 749–752.