On-the-fly audio source separation

Dalia El Badawy, Ngoc Q. K. Duong, Alexey Ozerov

2014 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 21–24, 2014, REIMS, FRANCE

ON-THE-FLY AUDIO SOURCE SEPARATION

Dalia El Badawy, Ngoc Q. K. Duong and Alexey Ozerov

Technicolor
975 avenue des Champs Blancs, CS 17616, 35576 Cesson Sévigné, France
{dalia.elbadawy, quang-khanh-ngoc.duong, alexey.ozerov}@technicolor.com

ABSTRACT

This paper addresses the challenging task of single channel audio source separation. We introduce a novel concept of *on-the-fly* audio source separation which greatly simplifies the user's interaction with the system compared to the state-of-the-art user-guided approaches. In the proposed framework, the user is only asked to listen to an audio mixture and type some keywords (*e.g.* "dog barking", "wind", etc.) describing the sound sources to be separated. These keywords are then used as text queries to search for audio examples from the internet to guide the separation process. In particular, we propose several approaches to efficiently exploit these retrieved examples, including an approach based on a generic spectral model with group sparsity-inducing constraints. Finally, we demonstrate the effectiveness of the proposed framework with mixtures containing various types of sounds.

Index Terms— On-the-fly source separation, user-guided, non-negative matrix factorization, group sparsity, universal spectral model.

1. INTRODUCTION

For a wide range of applications in audio enhancement and post-production, audio source separation still remains a very hot research topic. The problem becomes more challenging in the single-channel case where spatial information about the sources cannot be exploited. Thus most state-of-the-art approaches rather rely on the spectral diversity of individual sound sources, which is usually learned from relevant training data in order to separate them from the mixture [1, 2]. Such a class of supervised algorithms is often based on Non-negative Matrix Factorization (NMF) [3, 4, 5] or its probabilistic formulation known as Probabilistic Latent Component Analysis (PLCA) [2, 6]. However, relevant training data is not often available or representative enough, especially for non-popular sounds such as animal or environmental sounds.

Another type of so-called *user-guided* approaches rely on source-specific information provided by a user to guide the source separation process. For example, this information can be user-"hummed" sounds that mimic the sources in the mixture [6] or a speech transcription used to produce speech examples via a speech synthesizer [7]. Alternative user-guided approaches allow the end-user to manually annotate information about the activity of each source in the mixture [8, 9]. The annotated information is then used, instead of training data, to guide the separation process. In this line of annotation-based approaches, recent publications disclose an interactive strategy [10, 11] where the user can even perform annotation on the spectrogram of intermediate separation results so as to gradually correct the remaining errors. Despite the effectiveness of these user-guided approaches, they are usually very time consuming and require significant effort from the user. Additionally, the annotation process is only suitable for experienced people since they have to understand the spectrogram display in order to annotate it.

With the motivation of greatly simplifying the user interaction so as non-experienced people can easily do the job, we introduce in this paper a new concept of on-the-fly source separation for which the user guides the separation at a higher semantic level. More specifically, we propose a framework that only requires the user to listen to the mixture and to semantically describe the sources he/she would like to separate. For example, a user may wish to separate the "dog barking" (source 1 description) from the "bird song" (source 2 description). We then use these semantic descriptions as text queries to retrieve example audio files from the internet and use them to guide the source separation process. This strategy is akin to on-the-fly methods in visual search [12, 13] where an end-user searching for a certain person or a visual object is only required to type the person's name or the object's description. The corresponding representative example images are then retrieved via Google Image Search and used for training an appropriate classifier. Figure 1 depicts the workflow of the proposed system.

However, several challenges arise when using the aforementioned retrieved audio examples. First, examples retrieved from the internet are not guaranteed to contain a sound with spectral characteristics similar to those of the source in the mixture. Second, these examples may also be mixtures of several sources. Thus it is desired to have a mechanism to allow selecting only the most representative examples to

978-1-4799-3694-6/14/$31.00

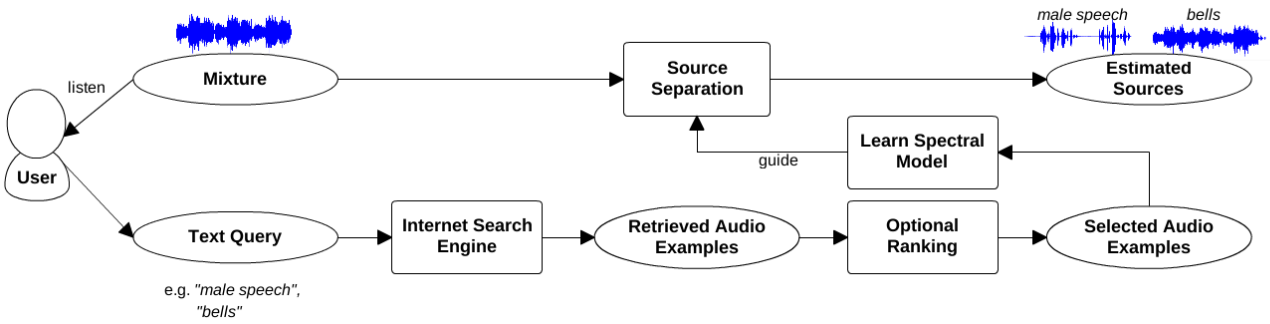


Fig. 1. General workflow of the proposed on-the-fly framework. A user listens to a mixture and types some keywords describing the sources. The keywords are used to retrieve examples to learn a spectral model for each source.

improve the separation result. We propose two alternative strategies to address this issue. The first one is based on the pre-selection of examples via a ranking scheme. Whereas the second exploits a universal spectral model learned from examples and handles the selection of the appropriate spectral patterns via some group sparsity-inducing constraints [4].

The rest of the paper is organized as follows. In Section 2 we summarize the supervised source separation approach based on the NMF model. We then present two classes of algorithms for the on-the-fly system in Section 3. In Section 4, we conduct experiments to validate the effectiveness of the proposed approach. Finally we conclude in Section 5.

2. NMF-BASED SUPERVISED SOURCE SEPARATION

This section discusses a standard supervised source separation approach for the single-channel case based on NMF, one of the most popular and widely used models in the state-of-the-art source separation. The general pipeline, which has been considered *e.g.* in [2, 3], consists in first learning corresponding source spectral models from some training data. Then these pre-learned models are used to guide the mixture decomposition.

Let $\mathbf{X}$ and $\mathbf{S}_j$ be the $F \times N$ matrices of the short-time Fourier transform (STFT) coefficients of the observed mixture signal and the j-th source signal, respectively, where F is the number of frequency bins and N the number of time frames. The mixing model writes

$$\mathbf{X} = \sum_{j=1}^{J} \mathbf{S}_j, \tag{1}$$

where J is the total number of sources. Let $\mathbf{V} = |\mathbf{X}|^{\cdot 2}$ be the power spectrogram of the mixture where $\mathbf{X}^{\cdot p}$ is the matrix with entries $[\mathbf{X}]_{il}^{p}$. NMF aims at decomposing the $F \times N$ non-negative matrix $\mathbf{V}$ as a product of two non-negative matrices $\mathbf{W}$ and $\mathbf{H}$ of dimensions $F \times K$ and $K \times N$, respectively, such that $\mathbf{V} \approx \hat{\mathbf{V}} = \mathbf{WH}$. This decomposition is done by optimizing the following criterion [5]

$$\underset{\mathbf{H}\geq 0, \mathbf{W}\geq 0}{min}\ D(\mathbf{V}\|\mathbf{WH}), \tag{2}$$

where $D(\mathbf{V}\|\hat{\mathbf{V}}) = \sum_{f,n=1}^{F,N} d_{IS}(\mathbf{V}_{fn}\|\hat{\mathbf{V}}_{fn})$ and $d_{IS}(x\|y) = \frac{x}{y} - \log(\frac{x}{y}) - 1$ is the Itakura-Saito divergence measure [4] which is a popular choice for audio applications. The parameters $\boldsymbol{\theta} = \{\mathbf{W}, \mathbf{H}\}$ are initialized with random non-negative values and are iteratively updated via multiplicative update (MU) rules [5].

In the supervised setting, the factorization of $\mathbf{V}$ is guided by a pre-learned spectral model. In other words, the matrix $\mathbf{W}$ is obtained (and fixed) by

$$\mathbf{W} = [\mathbf{W}_{(1)}, \ldots, \mathbf{W}_{(J)}], \tag{3}$$

where $\mathbf{W}_{(j)}$ is spectral model for j-th source learned also in the NMF decomposition of the training examples. Correspondingly, the activation matrix is also partitioned into blocks as $\mathbf{H} = [\mathbf{H}_{(1)}^T, \ldots, \mathbf{H}_{(J)}^T]^T$, where $\mathbf{H}_{(j)}$ denotes a block characterizing the time activations for j-th source. Thus, first $\mathbf{W}$ is estimated from training data by optimizing (2). Then, $\mathbf{H}$ is estimated from the mixture by optimizing (2) but using the previously calculated $\mathbf{W}$ and keeping it fixed. Once the parameters $\boldsymbol{\theta} = \{\mathbf{W}, \mathbf{H}\}$ are obtained, the source STFT coefficients are computed by Wiener filtering as

$$\hat{\mathbf{S}}_j = \frac{\mathbf{W}_{(j)}\mathbf{H}_{(j)}}{\mathbf{WH}} \odot \mathbf{X}, \tag{4}$$

where $\odot$ denotes the element-wise Hadamard product and the division is also element-wise. And finally, the time domain source estimates are obtained via the inverse STFT.

3. PROPOSED ON-THE-FLY SOURCE SEPARATION

The state-of-the-art supervised approach described in Section 2 will work efficiently with "good" training examples, *i.e.* the

ones whose spectral characteristics are similar to that of the source in the mixture. However, in the considered on-the-fly framework there is no guarantee that the audio examples retrieved through the internet from an external database will sound similar to the source in the mixture. For instance, the retrieved audio data for a query "bird" may contain various bird songs from different bird species. Thus using all retrieved examples would be less efficient than using only those corresponding to the bird song in the mixture. In this section we therefore present two different approaches that allow to overcome this limitation and efficiently use the examples to guide the separation process.

3.1. Example pre-selection-based approach

In order to discard inappropriate retrieved examples, *i.e.* those containing spectral characteristics that are quite different from the source in the mixture, in the training step, we propose pre-ranking schemes to first roughly select the more likely "good" candidates among all the retrieved ones. These ranking schemes are based on the *similarity* between each example and the mixture computed in one of the following ways:

(i) Similarity based on temporal correlation: in this scheme, the normalized cross correlation between each example for each source and the mixture signal is computed. Examples with higher correlation values are selected.

(ii) Similarity based on audio feature correlation: in this scheme, the spectral magnitudes of the examples and the mixture are considered. Features such as the spectral centroid and the spectral spread are computed for each frame to form a sequence of 2D feature vectors for each signal. Then the 2D correlation between these feature vectors is computed. Examples with higher correlation values are selected.

After the ranking process, only a short list of the retrieved examples is retained. For each source in the mixture, the corresponding selected examples are concatenated and used to learn the spectral model $\mathbf{W}_{(j)}$ in the NMF framework by solving the minimization problem:

$$\underset{\tilde{\mathbf{H}}_{(j)}\geq 0, \mathbf{W}_{(j)}\geq 0}{min} D(\mathbf{V}_j \| \mathbf{W}_{(j)}\tilde{\mathbf{H}}_{(j)}), \tag{5}$$

where $\mathbf{V}_j$ is the power spectrogram of the concatenated training examples for the j-th source. Once $\mathbf{W}_{(j)}$ are learned for all sources, they are used to guide the mixture separation, as explained in Section 1.

3.2. Universal model with a group sparsity constraint-based approach

Since the mixture contains several sources and a retrieved example may also contain several sources or additional noise, the similarity measure between them, *e.g.* as described in Section 3.1, may be very low so that even some "good" examples could be eventually discarded. In this section, we propose an alternative approach where the selection of "good" examples is done jointly in the model fitting step.

The proposed approach employs the so called universal model[1] with group sparsity constraints on the activation matrix $\mathbf{H}$ to enforce the selection of only few representative spectral patterns learned from all training examples. To begin, each retrieved example q corresponding to j-th source is used to learn the NMF spectral model denoted by $\mathbf{W}_{jq}$. Then the universal spectral model for j-th source is constructed as

$$\mathbf{W}_{(j)} = [\mathbf{W}_{j1}, \ldots, \mathbf{W}_{jQ_j}], \tag{6}$$

where Q_j is the number of retrieved examples for j-th source. In the NMF decomposition of the mixture, the spectral model $\mathbf{W}$ is constructed by (3). Then the activation matrix is estimated by solving the following optimization problem

$$\underset{\mathbf{H}\geq 0}{min}\ D(\mathbf{V}\|\mathbf{W}\mathbf{H}) + \lambda\Psi(\mathbf{H}), \tag{7}$$

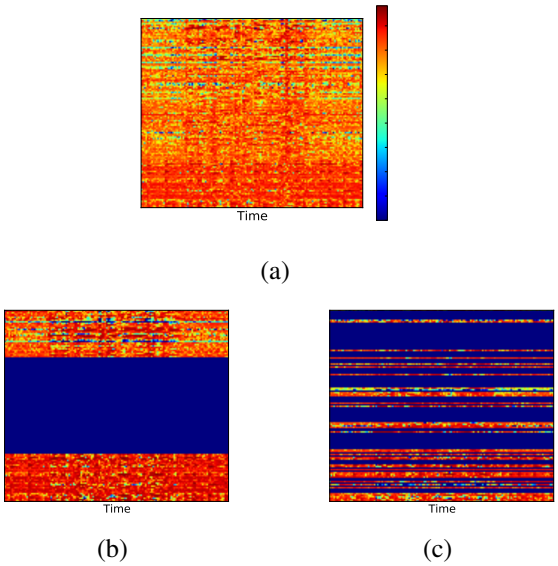
where $\Psi(\mathbf{H})$ denotes a penalty function imposing group sparsity on $\mathbf{H}$, and λ is a trade-off parameter determining the contribution of the penalty. When $\lambda = 0$, $\mathbf{H}$ is not sparse and the entire universal model is used as illustrated in Figure 2a. For $\lambda > 0$, different penalties can be chosen (*e.g.* as in [3, 4]); and in this paper we propose to use two alternative group sparsity-inducing penalties as follows.

(i) **Block sparsity-inducing penalty**

$$\Psi_1(\mathbf{H}) = \sum_{g=1}^{G} \log(\epsilon + \|\mathbf{H}_{(g)}\|_1), \tag{8}$$

where $\mathbf{H}_{(g)}$ is a subset of $\mathbf{H}$ representing the activation coefficients for g-th block, $\|.\|_1$ is the ℓ_1 norm, G is the total number of blocks, and ϵ is a small positive constant. In this case, a non-overlapping block represents one training example and G is the total number of examples used. This penalty is motivated by the fact that if some of the retrieved examples are more representative for the corresponding source in the mixture than the others, then it may be better to use only the former examples. It thus enforces the activation for "good" examples only while omitting the poorly fitting examples since their corresponding activation blocks will likely converge to zero, as visualized in Figure 2b. This block sparsity constraint was shown to be effective with the universal speech model in [3] in a denoising task; and in this paper we argue that it could also bring benefit in handling the selection of relevant training examples retrieved on-the-fly.

[1]The term "universal model" was introduced in [3] for the separation of speech and noise, which is also in analogy to the universal background models for speaker verification [14].



(a) (b) (c)

Fig. 2. Estimated activation matrix $\mathbf{H}$: (a) without a sparsity constraint, (b) with a block sparsity-inducing penalty (blocks corresponding to poorly fitting models are zero), and (c) with a component sparsity-inducing penalty (rows corresponding to poorly fitting spectral components from different models are zero).

(ii) **Component sparsity-inducing penalty**

$$\Psi_2(\mathbf{H}) = \sum_{k=1}^{K} \log(\epsilon + \|\mathbf{h}_k\|_1), \tag{9}$$

where $\mathbf{h}_k$ denotes k-th row of $\mathbf{H}$. This penalty is motivated by that fact that only a part of the spectral model learned from an example may fit well with the source in the mixture, while the remaining patterns (components) in the model do not (as in the case when an example is also a mixture of sounds). Thus instead of activating the whole block (all components in a spectral model $\mathbf{W}_{jp}$) as guided by $\Psi_1(\mathbf{H})$, the penalty $\Psi_2(\mathbf{H})$ allows to select only the more likely fitting spectral components from $\mathbf{W}_{jp}$. An example of $\mathbf{H}$ after convergence is shown in Figure 2c.

To derive algorithms optimizing (7) with different penalty functions (8) and (9), one can rely on MU rules and the majorization-minimization algorithm, as in [3] for the NMF with Kullback-Leibler divergence and in [4] for the NMF with Itakura-Saito divergence as considered in this paper. The resulting algorithm is summarized in Algorithm 1, where $\mathbf{P}_{(g)}$ is a matrix of the same size as $\mathbf{H}_{(g)}$ and $\mathbf{p}_k$ is a row vector of the same size as $\mathbf{h}_k$.

Algorithm 1 NMF with sparsity-inducing constraints

Input: $\mathbf{V}$, $\mathbf{W}$, λ
Output: $\mathbf{H}$
  Initialize $\mathbf{H}$ randomly
  $\hat{\mathbf{V}} = \mathbf{WH}$
  repeat
    if Block sparsity-inducing penalty **then**
      for $g = 1, ..., G$ **do**
        $\mathbf{P}_{(g)} \leftarrow \frac{1}{\epsilon + \|\mathbf{H}_{(g)}\|_1}$
      end for
      $\mathbf{P} = [\mathbf{P}_{(1)}^T, \ldots, \mathbf{P}_{(G)}^T]^T$
    end if
    if Component sparsity-inducing penalty **then**
      for $k = 1, ..., K$ **do**
        $\mathbf{p}_k \leftarrow \frac{1}{\epsilon + \|\mathbf{h}_k\|_1}$
      end for
      $\mathbf{P} = [\mathbf{p}_1^T, \ldots, \mathbf{p}_K^T]^T$
    end if
    $\mathbf{H} \leftarrow \mathbf{H} \odot \left(\frac{\mathbf{W}^T(\mathbf{V} \odot \hat{\mathbf{V}}^{\cdot -2})}{\mathbf{W}^T(\hat{\mathbf{V}}^{\cdot -1}) + \lambda \mathbf{P}} \right)^{\cdot \frac{1}{2}}$
    $\hat{\mathbf{V}} \leftarrow \mathbf{WH}$
  until convergence

4. EXPERIMENTS

4.1. Data and parameter settings

We evaluated the performance of the proposed on-the-fly approaches via a dataset containing 10 single-channel mixtures of two sources artificially mixed at 0 dB SNR. The mixtures were sampled at either 16000 Hz or 11025 Hz and vary in duration between 1 and 10 seconds. The sources in the mixtures represent different types of sound ranging from human speech to musical instruments and animal sounds. This variability of sound sources will demonstrate the power of the proposed on-the-fly strategy since *e.g.* appropriate training examples for non-popular sounds such as animal or environmental sounds are usually not available at the end-user's side. In our experiment, some example wave files were retrieved from `www.findsounds.com`, a search engine for audio where several parameters such as sample rate, number of channels, audio file format (wav, mp3), etc. can be specified and a list of URLs of audio files is accordingly retrieved. The keywords used included *guitar*, *bongos*, *drum*, *cat*, *dog*, *kitchen*, *river*, *chirps*, *rooster*, *bells*, and *car*. Additionally, speech examples were retrieved from the TIMIT database [15]. Note that the retrieved files were imposed to have sampling rates at least as high as that of the mixture; then the ones with higher sampling rates were downsampled to the mixture's sampling rate.

For parameter settings, a frame length of 47 ms with 50% overlapping was used for the STFT. The number of iterations for MU updates in all algorithms was 200 for training and 100 for testing. The number of NMF components for each

Method	NSDR	NSIR
Baseline on-the-fly	2.0	6.6
Temp. corr. -based ranking	2.4	7.1
Feature-based ranking	3.2	7.8
Universal non-constraint	3.1	7.5
Universal block sparsity ($\lambda = 128$)	3.3	**7.9**
Universal component sparsity ($\lambda = 64$)	**3.7**	**7.9**

Table 1. Average source separation performance.

source in the example pre-selection-based approach and the number of NMF components for each spectral model learned from one example in the universal model-based approach was set to 32. Several values were tested for the trade-off parameter λ which weights the contribution of the sparsity-inducing penalty (7); it was finally set to 128 and 64 for block and component sparsity, respectively.

4.2. Results and discussion

We compare the separation performance obtained by the baseline on-the-fly algorithm (named *Baseline on-the-fly*)[2] described in Section 2, where all retrieved examples were used to train one spectral model for each source, with that achieved by the example pre-selection-based approach described in Section 3.1, where only the 3 top-ranked examples were used to train the corresponding source spectral model. These examples were chosen either via the temporal correlation scheme (named *Temp. corr. -based ranking*) or the audio feature correlation-based scheme (named *Feature-based ranking*). We also evaluated the performance of the universal model-based approaches described in Section 3.2 with either no sparsity constraints i.e. $\lambda = 0$ (named *Universal non-constraint*), or a block sparsity-inducing penalty (8) (named *Universal block sparsity*), or a component sparsity-inducing penalty (9) (named *Universal component sparsity*).

Separation results were evaluated using the normalized signal-to-distortion ratio (NSDR) measuring overall distortion as well as the normalized signal-to-interference ratio (NSIR) [16, 17], measured in dB and averaged over all sources. Note that the normalized values were simply computed by subtracting the SDR of the original mixture signal from the SDR of the separated source. In other words, these normalized values show the improvement compared to the case where the user does not have access to a source separation system.

The results obtained by different algorithms are shown in Table 1, and sound files for subjective listening are available online[3]. As can be seen, the proposed on-the-fly strategy for retrieving examples via a search engine to guide the source separation brings significant benefit where the average performance over all methods was of 3 dB NSDR and 7.5 dB NSIR. As expected, pre-selecting retrieved examples even by simple temporal correlation or feature correlation improves the result over the baseline, *e.g.* by 0.4 dB and 1.2 dB NSDR, respectively, since it allows to discard inappropriate examples in the training phase. Also as expected, feature-based correlation was slightly better than temporal correlation since it is unaffected by dynamic variations; indeed these variations may result in low temporal correlation values between otherwise similar sounds causing their unnecessary elimination. Moreover, better results were achieved by the universal model with group sparsity constraint-based approaches with an improvement of 0.2 and 0.6 dB NSDR over the non-constraint case. This shows that these proposed methods efficiently handle the use of representative spectral models learned from training examples in the parameter estimation process. Finally, it should be noted that the component sparsity-inducing penalty produces the best result with 3.7 dB NSDR and 7.9 dB NSIR. We think that this is thanks to the fact that this penalty allows exploiting the most representative spectral patterns from different spectral models.

[2]Note that as on-the-fly source separation is a new approach, there is currently no state-of-the-art methods with which to compare; and thus we consider as a baseline the method in Section 2

[3]`http://audiosourceseparation.wordpress.com/`

5. CONCLUSION

In this paper, we introduced the novel concept of on-the-fly audio source separation and proposed several algorithms implementing it. In contrast with other state-of-the-art user-guided approaches, the considered framework allows to greatly simplify the user interaction with the system such that everyone can do source separation just by typing keywords describing audio sources in the mixture. In particular, we proposed to use a universal spectral model with group sparsity-inducing constraints so as to efficiently handle the selection of representative spectral patterns learned from retrieved examples. Experiments with mixtures containing various sound types confirm the potential of the proposed on-the-fly source separation concept as well as the corresponding algorithms. Future work includes addressing the case where the user does not completely specify all the sources in the mixture (e.g. describing one out of two sources). Additionally, a compressed sensing approach for overlapping blocks [18], and a mixed block and component sparsity-inducing penalty [19] would also be investigated within the considered universal spectral model framework.

6. REFERENCES

[1] E. Vincent, S. Araki, F. Theis, G. Nolte, P. Bofill, H. Sawada, A. Ozerov, V. Gowreesunker, D. Lutter, and N. Q. K. Duong, "The Signal Separation Campaign (2007-2010): Achievements and remaining challenges," *Signal Processing*, vol. 92, pp. 1928–1936, 2012.

[2] P. Smaragdis, B. Raj, and M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," in *Int. Conf. on Independent Component Analysis and Signal Separation (ICA)*, 2007, pp. 414–421.

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

[4] A. Lefèvre, F. Bach, and C. Févotte, "Itakura-Saito nonnegative matrix factorization with group sparsity," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011, pp. 21–24.

[5] C. Févotte, N. Bertin, and J. Durrieu, "Non-negative matrix factorization with the itakura-saito divergence. with application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830.

[6] P. Smaragdis and G. J. Mysore, "Separation by humming: User-guided sound extraction from monophonic mixtures," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2009, pp. 69–72.

[7] L. Le Magoarou, A. Ozerov, and N. Q. K. Duong, "Text-informed audio source separation using nonnegative matrix partial co-factorization," in *IEEE Int. Workshop on Machine Learning for Signal Processing (MLSP)*, 2013, pp. 1–6.

[8] A. Lefèvre, F. Bach, and C. Févotte, "Semi-supervised NMF with time-frequency annotations for single-channel source separation," in *Int. Conf. on Music Information Retrieval (ISMIR)*, 2012, pp. 115–120.

[9] N. Q. K. Duong, A. Ozerov, and L. Chevallier, "Temporal annotation-based audio source separation using weighted nonnegative matrix factorization," in *IEEE Int. Conf. on Consumer Electronics - Berlin (ICCE-Berlin)*, 2014.

[10] N. J. Bryan and G. J. Mysore, "Interactive refinement of supervised and semi-supervised sound source separation estimates," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013, pp. 883–887.

[11] N. Q. K. Duong, A. Ozerov, L. Chevallier, and J. Sirot, "An interactive audio source separation framework based on nonnegative matrix factorization," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014, pp. 1586–1590.

[12] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "On-the-fly specific person retrieval," in *13th Int. Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, 2012, pp. 1–4.

[13] K. Chatfield and A. Zisserman, "Visor: Towards on-the-fly large-scale object category retrieval," in *Asian Conference on Computer Vision*. 2012, Lecture Notes in Computer Science, pp. 432–446, Springer.

[14] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1, pp. 19–41, 2000.

[15] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, and N. Dahlgren, "DARPA TIMIT: Acoustic-phonetic continuous speech corpus," Tech. Rep., NIST, 1993, distributed with the TIMIT CD-ROM.

[16] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, 2006.

[17] A. Ozerov, P. Philippe, R. Gribonval, and F. Bimbot, "One microphone singing voice separation using source-adapted models," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2005, pp. 90–93.

[18] S. Gishkori and G. Leus, "Compressed sensing for block-sparse smooth signals," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014.

[19] A. Hurmalainen, R. Saeidi, and T. Virtanen, "Group sparsity for speaker identity discrimination in factorisation-based speech recognition," in *Interspeech*, pp. 17–20.