A Local Discontinuity Based Approach for Monaural Singing Voice Separation from Accompanying Music with Multi-stage Non-negative Matrix Factorization

Hatem Deif[1,2], Wenwu Wang[4], Lu Gan[1], and Saadat Alhashmi[3]

[1]College of Engineering, Design and Physical Sciences, Brunel University London, Uxbridge, Middlesex, UK
{[2]University College, [3]College of Engineering}, Abu Dhabi University, Abu Dhabi, UAE
[4]Center for Vision, Speech and Signal Processing, University of Surrey, Guildford, UK

Email: hatem.deif@brunel.ac.uk, w.wang@surrey.ac.uk, lu.gan@brunel.ac.uk, saadat.alhashmi@adu.ac.ae

***Abstract*— This paper presents a new method to improve a multi-stage non-negative matrix factorization (NMF) based algorithm for separating singing voice from accompanying music in single channel recordings. In the proposed method, local spectral and temporal discontinuity measures are used to refine the vocal and music components obtained by the baseline NMF algorithm. The effectiveness of the proposed algorithm is demonstrated using the MIR-1K data set.**

Keywords—non-negative matrix factorization; single channel voice/music separation; harmonic/percussive separation

I. INTRODUCTION

Non-negative matrix factorization (NMF) [1], [2], a popular technique for learning parts-based representations, has been used for monaural source separation of acoustic inputs [3], [4]. In particular, it has been used for singing voice separation based on the assumption that spectrogram of music can be expressed by a limited number of spectral templates [5].

A number of attempts were made to separate singing voice from monaural music using NMF. For example: Vembu and Baumann [6] applied NMF to decompose the mixture spectrogram into a set of components and used unsupervised learning algorithms to cluster their spectral bases into vocal and nonvocal ones. Chanrungutai and Ratanamahatana [7] performed separation based on the rhythmic structure of music components. In the hybrid system by Virtanen et al. [8], NMF is utilized to learn the accompaniment model once vocals are removed using a pitch detection algorithm. Then the learned accompaniments are subtracted from the original mixture to yield the separated vocals. The system proposed by Durrieu et al. [9] represents the leading voice using a source/filter model while an unconstrained NMF model is used to represent the background music.

A more recent approach for monaural voice/music separation is the use of the harmonic-percussive sound separation (HPSS) algorithm developed by Ono et al. in [10]. HPSS is designed as an optimization problem to minimize the temporal/spectral gradients of the separated spectrograms to enhance the horizontal/vertical ridges. Tachibana et al. [11] used HPSS in two stages with high and low frequency resolution spectrograms to separate the pitched and percussive instruments respectively from the mixture signal. Jeong and Lee extended this idea in [12] by including the vocal signal along with the harmonic and percussive instruments in a single optimization framework. FitzGerald used a median filtering approach in two stages to address the same problem in [13].

An interesting algorithm by Zhu et al. [14] replaces the HPSS in each stage by NMF. In their algorithm, each NMF component in the two stages is classified as either a vocal component or a musical one based on the thresholding of a discontinuity measure. The algorithm is fast and effective. However, we observed that many of the NMF components were actually a mixture of both voice and music. In this paper, we propose to decompose each one of these components into a vocal component and a musical one. The decomposition is achieved by measuring discontinuities at different parts of the component rather than the overall discontinuity of the whole component.

The rest of the paper is organized as follows: Section II briefly summarizes the baseline multi-stage NMF algorithm in [14]. Section III presents our method for improving this algorithm with the use of local discontinuity measures for further refining the NMF components before reconstructing sound sources. Section IV shows the results of applying the proposed method on the MIR-1K dataset as compared with the baseline method. Finally, section V gives the conclusion.

II. THE BASELINE ALGORITHM

A. Non-negative matrix factorization for sound source separation

Let $\mathbf{X}$ be the $K \times T$ non-negative matrix that represents the magnitude spectrogram of the mixture signal $\boldsymbol{x}$, where K is the number of frequency bins and T is the number of time frames. The approximate non-negative factorization of $\mathbf{X}$ is given by

$$\mathbf{X} \approx \mathbf{BG} \tag{1}$$

where $\mathbf{B}$ and $\mathbf{G}$ are the basis and gains matrices of dimensions $K\times J$ and $J\times T$ respectively, and J represents the number of components. Each component $\mathbf{X}^j$ is defined as the product of a spectral basis $\mathbf{b}^j$ (the j^{th} column in $\mathbf{B}$) and the corresponding temporal gain $\mathbf{g}^j$ (the j^{th} row in $\mathbf{G}$)

$$\mathbf{X}^j = \mathbf{b}^j\mathbf{g}^j \tag{2}$$

where $j = 1, \dots, J$ is the component index.

The factorization in (1) is often achieved by minimizing a cost function defined on $\mathbf{X}$ and $\mathbf{BG}$ while applying the non-negativity contraints. The Kullback-Leibler (K-L) divergence is commonly used in source separation and it performed better in the current algorithm

$$D(\mathbf{X} \parallel \mathbf{BG}) = \sum_{k=1}^{K}\sum_{t=1}^{T} \mathbf{X}_{k,t} \log \frac{\mathbf{X}_{k,t}}{[\mathbf{BG}]_{k,t}} - \mathbf{X}_{k,t} + [\mathbf{BG}]_{k,t}. \tag{3}$$

This minimization problem was solved as in [2] with the following multiplicative update rules.

$$\mathbf{B} \leftarrow \mathbf{B} \otimes \frac{\frac{\mathbf{X}}{\mathbf{BG}}\mathbf{G}^{\mathrm{T}}}{\mathbf{1}\mathbf{G}^{\mathrm{T}}}, \qquad \mathbf{G} \leftarrow \mathbf{G} \otimes \frac{\mathbf{B}^{\mathrm{T}}\frac{\mathbf{X}}{\mathbf{BG}}}{\mathbf{B}^{\mathrm{T}}\mathbf{1}} \tag{4}$$

where $\otimes$ and / represent element-wise multiplication and division respectively, $\mathbf{1}$ denotes an all-one matrix of the same size as $\mathbf{X}$, and T is the matrix transpose.

In many of the existing algorithms [6], [7], each NMF component is ideally assumed to be coming from one sound source and thus classified as either vocal or instrumental. This is also the assumption of the baseline algorithm in [14] summarized in the following section.

B. Classifying each component using spectral and temporal discontinuity measures

The algorithm presented in [14] contains two stages, one for separating pitched instruments from the mixture, and the other for separating percussion instruments. The separation of pitched instruments is based on the observation that in a spectrogram with a long FFT window, pitched instruments have a stable pitch and thus appear continuous in the temporal direction and discontinuous in the spectral direction. To filter out these pitched instruments, the magnitude spectrogram is decomposed into a set of NMF components and those components that are spectrally discontinuous are removed.

Spectral discontinuity of each component is measured by summing and normalizing the squared differences between adjacent elements in its spectral basis. Specifically, for each component $\mathbf{X}^j$, the spectral discontinuity measure $d_{\mathrm{s}}(\mathbf{X}^j)$ is defined as

$$d_{\mathrm{s}}(\mathbf{X}^j) = \frac{\sum_{k=2}^{K}(\mathbf{B}_{k,j}-\mathbf{B}_{k-1,j})^2}{\sum_{k=1}^{K}\mathbf{B}_{k,j}^2} \tag{5}$$

and if it is larger than a threshold θ_{s}, the component is considered to be originating from a pitched instrument. The suitable value for θ_{s} was found empirically to be 0.4 as explained in [14].

A new magnitude spectrogram $\mathbf{X}'$ is formed by subtracting all pitched components from the input mixture spectrogram $\mathbf{X}$

$$\mathbf{X}' = \max\left(\mathbf{0},\ \mathbf{X} - \sum_{\substack{j=1,\dots,J \\ d_{\mathrm{s}}(\mathbf{X}^j)>\theta_{\mathrm{s}}}} \mathbf{X}^j\right) \tag{6}$$

where $\mathbf{0}$ is an all-zero matrix of the same size as $\mathbf{X}$, and $\max(\mathbf{Y},\mathbf{Z})$ takes the element-wise maximum of matrices $\mathbf{Y}$, $\mathbf{Z}$, which is used to ensure there are no negative elements in $\mathbf{X}'$. After that, $\mathbf{X}'$ is inverted back to time domain using the phase information of the original sound mixture, then it is used as an input to the second stage of the algorithm.

In the second stage of the algorithm, percussion instruments are separated from the sound mixture based on the observation that in a short window spectrogram, they appear continuous in the spectral direction and discontinuous in the temporal direction. Therefore, NMF components that are temporally discontinuous can be considered as originating from percussive sounds and thus removed using a similar temporal discontinuity thresholding method. Specifically, for each component $\mathbf{X}^j$, the temporal discontinuity measure $d_{\mathrm{t}}(\mathbf{X}^j)$ is defined as

$$d_{\mathrm{t}}(\mathbf{X}^j) = \frac{\sum_{t=2}^{T}(\mathbf{G}_{j,t}-\mathbf{G}_{j,t-1})^2}{\sum_{t=1}^{T}\mathbf{G}_{j,t}^2} \tag{7}$$

and if it is larger than a threshold θ_{t}, the component is considered to be originating from percussion instruments.

The separated voice spectrogram is obtained by subtracting all percussion instruments components from $\mathbf{X}'$ then it is inverted back to time domain using the phase information of $\mathbf{X}'$ to yield the separated singing voice $\boldsymbol{v}$. Music signal can be obtained by subtracting $\boldsymbol{v}$ from the mixture signal $\boldsymbol{x}$.

The algorithm summarized above classifies each NMF component as a vocal component or a musical one. However, it was noticed that many of the components contain a mixture of both voice and music, rendering an inaccurate classification in practice.

III. Using Local Discontinuity Measures to Refine the NMF Components

To address the problem discussed above, we propose a method for improving the separation quality through the use of local discontinuity measures of the NMF components. To explain the idea, we first consider the long window spectrogram factorization stage where d_{s} is used to classify NMF components into pitched and non-pitched ones. At this stage we noticed that many of the components that were classified as non-pitched components still contain sounds of pitched instruments.

An example is shown in Fig. 1. Specifically, Fig. 1(a) and Fig. 1(b) are the spectrograms of the original pitched music and voice respectively, while Fig. 1(c) shows the spectrogram of one of the non-pitched components. It can be

observed that the non-pitched component still contains traces of pitched music.

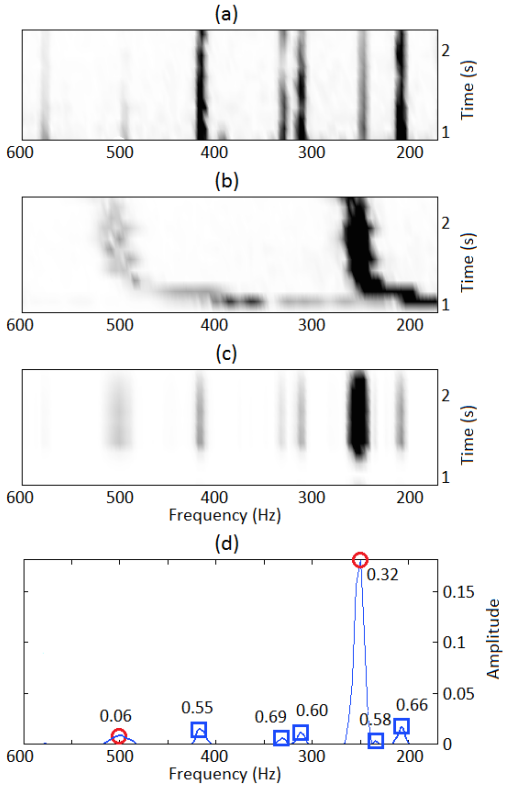

Fig. 1. Long-window spectrograms of (a) the original music, (b) the original voice, and (c) a component classified as non-pitched (vocals + percussions) component. The spectral basis of the component is shown in (d) where the local spectral discontinuity P_s of each peak is displayed.

To further refine this component, we first identify the I highest peaks in its spectral basis $\mathbf{b}^j$. Then, the local spectral discontinuity P_s around each peak is calculated as follows:

$$P_s(i,j) = \frac{\sum_{k=lo(i)}^{hi(i)} (\mathbf{B}_{k,j} - \mathbf{B}_{k-1,j})^2}{\sum_{k=lo(i)}^{hi(i)} \mathbf{B}_{k,j}^2} \quad (8)$$

where $i = 1, \ldots, I$ is the peak index, and the lower bound $lo(i)$ and the upper bound $hi(i)$ are given by

$$lo(i) = \max\left(0, f_i - \frac{l}{2}\right) \quad (9)$$

$$hi(i) = \min\left(f_i + \frac{l}{2}, K\right) \quad (10)$$

where f_i represents the frequency bin (index) of the peak and l is the peak width (in number of frequency bins) which is assumed to be constant for all peaks.

Fig. 1(d) shows the spectral basis $\mathbf{b}^j$ as well as the values of $P_s(i,j)$ for each peak. In our experiments, we observed that peaks with $P_s > \theta_s$ ($\theta_s = 0.4$) mostly belong to pitched instruments (denoted by blue squares); otherwise, they are from the voice (denoted by red circles).

Following this observation, we propose to remove the pitched peaks (with $P_s > \theta_s$) from the basis $\mathbf{b}^j$ of this component (as well as all non-pitched components) in order to obtain a 'cleaner' non-pitched component. The removed pitched peaks are added together to form a new pitched component. Algorithms 1 and 2 depict the new long window spectrogram factorization stage in detail.

Algorithm 1 Separating pitched instruments from the sound mixture $\boldsymbol{x}$

Input: Mixture signal $\boldsymbol{x}$
Output: Pitched-instruments-removed signal $\boldsymbol{x}'$
Initialization: J
Calculate $\mathbf{B}, \mathbf{G}$ from (3), (4)
for $j = 1{:}J$
  if $(d_s(\mathbf{X}^j) > \theta_s)$
    $\mathbf{X}^j{}_{pitched} \leftarrow \mathbf{X}^j$
  else
    Run Algorithm 2 to extract $\mathbf{X}^j{}_{pitched}$ from $\mathbf{X}^j$
  end if
end for
$\mathbf{X}' \leftarrow$ Calculate from (6) using all $\mathbf{X}^j{}_{pitched}$ above
$\boldsymbol{x}' \leftarrow$ Inverse STFT of $\mathbf{X}'$

Algorithm 2 Split a component $\mathbf{X}^j$ into a pitched and a non-pitched one based on P_s

Input: Component $\mathbf{X}^j$ with $d_s \le \theta_s$ $(\mathbf{X}^j = \mathbf{b}^j\mathbf{g}^j)$
Output: Extracted pitched component $\mathbf{X}^j{}_{pitched}$
Initialization: θ_s, l
$\mathbf{v}^j \leftarrow \mathbf{b}^j$
$\mathbf{f} \leftarrow$ Locations of the I highest peaks in $\mathbf{v}^j$
for $i = 1{:}I$
  Calculate $P_s(i,j)$, $lo(i)$ and $hi(i)$ from (8)-(10)
  if $(P_s(i,\mathrm{j}) > \theta_s)$
    $ind = lo(i){:}hi(i)$
    $\mathbf{v}^j{}_{ind} \leftarrow \mathbf{0}$
  end if
end for
$\mathbf{m}^j \leftarrow \mathbf{b}^j - \mathbf{v}^j$
$\mathbf{X}^j{}_{pitched} \leftarrow \mathbf{m}^j\mathbf{g}^j$

Similarly, at the second stage where percussion instruments are separated from vocals using short window spectrogram factorization, it was noticed that many of the NMF components that were classified as originating from percussion instruments ($d_t > \theta_t$), still contain vocal sounds. Again, we searched for the I highest peaks in the temporal gain $\mathbf{g}^j$ of each of these components and we calculated the local temporal discontinuity P_t around each peak defined as:

$$P_t(i,j) = \frac{\sum_{t=lo(i)}^{hi(i)} (\mathbf{G}_{j,t} - \mathbf{G}_{j,t-1})^2}{\sum_{t=lo(i)}^{hi(i)} \mathbf{G}_{j,t}^2} \quad (11)$$

with

$$lo(i) = \max\left(0, c_i - \frac{w}{2}\right) \quad (12)$$

$$hi(i) = \min\left(c_i + \frac{w}{2}, K\right) \quad (13)$$

where c_i represents the time frame (index) of the i^{th} peak and w is the peak width measured in terms of the number of time frames and assumed to be constant for all peaks.

Peaks are assumed to belong to vocals if $P_t \leq \theta_t$ and thus removed from the percussion component gain $\mathbf{g}^j$ to obtain a refined one. The removed peaks are added together to form a new vocal gain. In this way, the percussion component is split into a new vocal component and a refined percussion one. All refined percussion components are used to re-synthesize the singing voice as explained at the end of section II.

IV. Experimental Results

The MIR-1K dataset [15] is used to evaluate the effectiveness of the proposed algorithm in comparison to the baseline algorithm in [14]. The dataset consists of 1000 song clips with duration ranging from 4 to 13 seconds, extracted from 110 karaoke Chinese pop songs performed mostly by amateurs. The sampling rate of each song is 16 kHz, and the singing voice and music accompaniment were recorded in the right and left channels respectively. The voice and music signals were linearly mixed with equal energy to generate the mixture signal.

The separation performance was measured using the source-to-distortion ratio (SDR), source-to-interferences ratio (SIR), and sources-to-artifacts ratio (SAR) defined in [16]. These were calculated using the BSS_Eval toolbox [17] where higher values indicate better separation quality.

The first experiment was run using the original algorithm with all its parameters as in [14]. In the first stage, pitched instruments were separated using a spectrogram with a long FFT window of 4096 samples and an overlap of 50%. The spectral discontinuity threshold θ_s was set to 0.4. Percussion instruments were separated in the second stage where the FFT length was set to 256 samples with also 50% window overlap, and the temporal discontinuity threshold θ_t was set to 0.2. The number of components J was fixed to 15 in the two stages.

In the second experiment, the long window spectrogram factorization stage was implemented using the original algorithm as in [14] without any modification while the short window stage (i.e. the second stage) was implemented using our proposed algorithm of removing the vocal peaks from percussion components gains. A fixed width w of 250 time frames (which corresponds to 2 seconds) was chosen empirically, and I was set to 20.

In the third experiment, we used our proposed algorithm only during the long window stage where pitched peaks are removed from non-pitched components basis. All peaks were assumed to have a width l of 6 frequency bins (~24 Hz). Finally, in the fourth experiment our proposed algorithm was used in both stages.

Fig. 2 shows the results based on the three metrics, namely, SDR, SIR and SAR for the separated voice for the four experiments. We noticed that the best separation performance was achieved when using our proposed algorithm during the long window stage only, where the median SDR improved by 1 dB, and the median SIR improved by 1.2 dB, while the median SAR decreased by only 0.2 dB. It was also noted that using the proposed algorithm in the two stages leads to similar results.

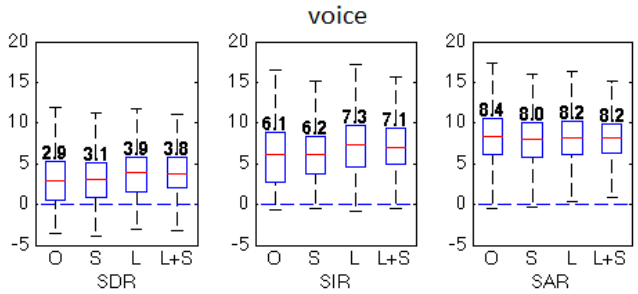

Fig. 2. Separation performance for singing voice using SDR (left), SIR (middle), and SAR (right) metrics. Four boxplots are shown for each metric; the leftmost one is for Zhu's original algorithm in the two stages (O), followed by the new modified algorithm during the short window stage (S), then during the long window stage (L), and finally combining both modifications (L+S). Outliers are not shown. Median values are displayed.

On the other hand, Fig. 3 shows the same three metrics for the separated music in all the four experiments. In this case, the best separation performance was obtained when using our proposed algorithm during the two stages, where the median SDR improved by 1.2 dB, and the median SIR improved by 1.8 dB, while the median SAR decreased by 2 dB. The reader can also check sound samples for the four experiments in [18].

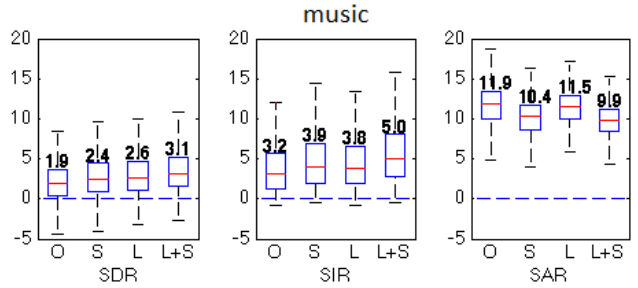

Fig. 3. Separation performance for music instruments using the same metrics as in Fig. 2.

V. Conclusion

We have presented a method to improve the performance of a multi-stage NMF algorithm for the separation of singing voice from monaural music recordings by applying the local spectral and temporal discontinuity measures on the peaks of basis and gains of the NMF components that were identified as containing both voice and music. These components were then split into vocal and music components instead of being classified as a whole to be either from singing voice or music instruments. Experiments indicated that the new algorithm improves the separation quality for both voice and music.

Acknowledgment

The authors of this paper would like to thank Dr. Wei Li and Dr. Bilei Zhu for providing the code of their algorithm.

REFERENCES

[1] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization.," Nature, vol. 401, no. 6755, pp. 788–91, Oct. 1999.

[2] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in Proc. Neural Inf. Process. Syst., Denver, CO, 2001, pp. 556–562.

[3] P. Smaragdis, "Discovering auditory objects through non-negativity constraints," in Proc. ISCA Tutorial and Research Workshop Statistical Perceptual Audio Process., Jeju Island, Korea, 2004.

[4] P. Smaragdis, "Non-negative matrix factor deconvolution ; Extracation of multiple sound sources from monophonic Inputs," in Proc. Int. Symp. Independent Component Analysis and Blind Signal Separation, Sep. 2004, pp. 494–499.

[5] J. C. Brown and P. Smaragdis, "Non-negative matrix factorization for polyphonic music transcription," in Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), Oct. 2003, pp. 177–180.

[6] S. Vembu and S. Baumann, "Separation of vocals from polyphonic audio recordings," in Proc. Int. Conf. Music Inf. Retrieval, London, U.K., 2005, pp. 337–344.

[7] A. Chanrungutai and C. A. Ratanamahatana, "Singing voice separation for mono-channel music using non-negative matrix factorization," in Proc. IEEE Int. Conf. Advanced Technologies for Communications, Bangkok, Thailand, Oct. 2008, pp. 243–246.

[8] T. Virtanen, A. Mesaros, and M. Ryynänen, "Combining pitch-based inference and non-negative spectrogram factorization in separating vo- cals from polyphonic music," in Proc. ISCA Tutorial and Research Workshop on Statistical and Perceptual Audition (SAPA'08), Brisbane, Australia, Sep. 2008, pp. 17–20.

[9] J.-L. Durrieu, B. David, and G. Richard, "A musically motivated mid-level representation for pitch estimation and musical audio source separation," IEEE J. Sel. Topics Signal Process., vol. 5, no. 6, pp. 1180–1191, Oct. 2011.

[10] N. Ono, K. Miyamoto, J. L. Roux, H. Kameoka, and S. Sagayama, "Separation of a monaural audio signal into harmonic/percussive com- ponents by complementary diffusion on spectrogram," in Proc. Eur. Signal Processing Conf., Lausanne, Switzerland, Aug. 2008.

[11] H. Tachibana, T. Ono, N. Ono, and S. Sagayama, "Melody line estimation in homophonic music audio signals based on temporal-variability of melodic source," in Proc. ICASSP, 2010, pp. 425–428.

[12] I.-Y. Jeong and K. Lee, "Vocal separation from monaural music using temporal/spectral continuity and sparsity constraints," IEEE Signal Processing Lett., vol. 21, no. 10, pp. 1197-1200, 2014.

[13] D. FitzGerald and M. Gainza, "Single channel vocal separation using median filtering and factorisation techniques," ISAST Trans. Electron. Signal Process., vol. 4, no. 1, pp. 62–73, Jan. 2010.

[14] B. Zhu, W. Li, R. Li, and X. Xue, "Multi-stage non-negative matrix factorization for monaural singing voice separation," IEEE Trans. Audio, Speech, Lang. Process., vol. 21, no. 10, pp. 2096–2107, Oct. 2013.

[15] C.-L. Hsu and J.-S. R. Jang, "On the improvement of singing voice sepa- ration for monaural recordings using the MIR-1K dataset," IEEE Trans. Audio, Speech, Lang. Process., vol. 18, no. 2, pp. 310–319, Feb. 2010.

[16] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," IEEE Trans. Audio, Speech, Lang. Process., vol. 14, no. 4, pp. 1462– 1469, Jul. 2006.

[17] BSS_Eval toolbox available at http://bass-db.gforge.inria.fr/bss_eval/

[18] Sound samples available at:
https://sites.google.com/site/voicemusicseparation/