Separation of Vocals From Monaural Music Recordings Using Diagonal Median Filters and Practical Time-Frequency Parameters

Hatem Deif[1,2], Derry Fitzgerald[3], Wenwu Wang[4], and Lu Gan[1]

[1]College of Engineering, Design and Physical Sciences, Brunel University London, Uxbridge, Middlesex, UK
[2]University College, Abu Dhabi University, Abu Dhabi, UAE
[3]NIMBUS Centre, Cork Institute of Technology, Bishopstown, Cork, Ireland
[4]Center for Vision, Speech and Signal Processing, University of Surrey, Guildford, UK

Email: hatem.deif@brunel.ac.uk, derry.fitzgerald@cit.ie, w.wang@surrey.ac.uk, lu.gan@brunel.ac.uk

***Abstract*— In this paper we present a novel method for using median filters to separate vocals from single channel music recordings that contain harmonic and percussive sounds. The method incorporates diagonal median filters in addition to the horizontal and vertical ones to better match the characteristics of the vocal track. Further improvements are obtained through proper choices of the filters' lengths using time and frequency units. The effectiveness of the proposed method is demonstrated using a test set of songs by the Beach Boys.**

Keywords—single channel voice/music separation, harmonic-percussive separation

I. Introduction

Separating singing voice (or voices) from single channel music recordings is a challenging problem in audio signal processing. Addressing this problem would have impacts in several applications. For example, the separated voice can be used for lyrics alignment [1] and singer identification [2] while the separated instrumental track could be used for melody extraction and transcription [3].

There are a wide variety of approaches to this problem, such as pitch based approaches [4], probabilistic approaches [5], non-negative matrix factorization [6], sparse and low rank matrix decomposition [7], repetition based techniques [8], and harmonic-percussive separation techniques [9]. The later ones are of interest here since no training is required and they produce good results.

The main idea in harmonic-percussive separation approaches for separating vocals from music signals is that voice appears like percussive sounds at high frequency resolution (long FFT window) spectrograms while it appears somewhat closer to pitched instruments at low frequency resolution spectrograms. Thus two stages operating at two different frequency resolutions are required to separate vocals from both harmonic and percussive instruments.

Inspired by this observation, Tachibana et al. built a two-stage diffusion based system for enhancing singing voice utilizing the anisotropic smoothness of harmonic and percussive instruments [9]. Jeong and Lee extended this idea by including the vocal signal along with harmonic and percussive instruments in a single optimization framework [10]. Zhu et al. used non-negative matrix factorization at two different resolutions to separate the instrumental track [11].

A parallel approach presented in [12] uses median filtering along the time and frequency axes of the spectrogram to separate harmonic and percussive components. Applying median filters along the horizontal and vertical directions of the spectrogram is suitable for separating harmonic instruments from percussive ones since they appear as horizontal and vertical lines respectively. However, it is interesting to note that vocals spectrograms contain more modulation in many cases and may be distorted by filters designed for pitched and percussion instruments. For that reason, we propose to use diagonal median filters to further enhance the separation of vocals.

The rest of the paper is organized as follows: Section II explains briefly the use of median filters for separating the vocal track from monaural songs as in [12]. Section III introduces the novel use of diagonal median filters for better separation of vocals. Section IV introduces the use of seconds and hertz for filters' lengths. Finally, section V presents the experiments for finding practical filters' lengths as well as the results when using diagonal filters, while section VI gives the conclusion.

II. Review of The Baseline Algorithm

The median of a list of values is the value at the center of the sorted list. If the number of values is even, it is the mean of the two values at the center. When a median filter of length l is applied on an input vector $\boldsymbol{x}$, the result is the output vector $\boldsymbol{y}$ defined in (1) if l is odd and in (2) if l is even,

$$y(n) = median\{x(n-\frac{l-1}{2} : n+\frac{l-1}{2})\} \quad (1)$$

$$y(n) = median\{x(n-\frac{l}{2} : n+\frac{l}{2}-1)\} \quad (2)$$

where n is the index of the processed element of the output vector $\boldsymbol{y}$.

Since percussion instruments form vertical ridges in the magnitude spectrogram as in Fig. 1(a), applying a median filter MD_h with length l_h for each frequency slice in the spectrogram will remove these ridges if the filter length is large enough compared to the percussion instrument duration as they will be treated like outliers. We call this the horizontal filter since it is applied along the horizontal (time) axis.

On the other side, harmonics of pitched instruments form horizontal ridges in the magnitude spectrogram as in Fig. 1(b), so applying a median filter MD_p with length l_p for each time frame in the spectrogram will remove these ridges as they will be treated like outliers if the filter length is large enough compared to the ridges frequency span. We call this the vertical filter since it is applied along the vertical (frequency) axis.

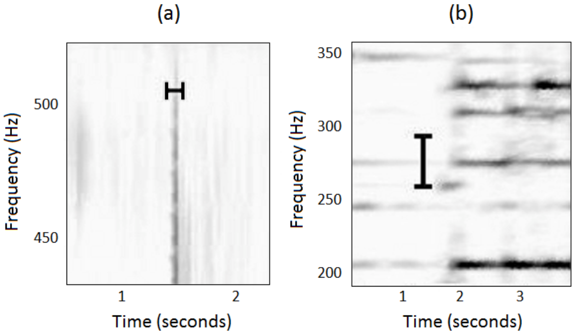

Fig. 1. (a) Horizontal median filter for removing vertical ridges of percussive instruments, (b) vertical median filter for removing horizontal ridges of pitched instruments.

Applying the previous two filters (one at a time) on every sample in the magnitude spectrogram S will produce the harmonic-enhanced spectrogram H and the percussion-enhanced spectrogram P.

$$H = MD_h\{S, l_h\} \tag{3}$$

$$P = MD_p\{S, l_p\} \tag{4}$$

To reduce median filter artifacts and to improve separation at overlap regions, Wiener filter masks M_H and M_P are generated from H and P as in (5) and (6) then multiplied (element-wise) by the original complex spectrogram to produce the harmonic instruments and percussive instruments spectrograms respectively. These spectrograms are transformed back to time domain to yield the separated harmonic and percussive signals.

$$M_H = \frac{H^2}{H^2 + P^2} \tag{5}$$

$$M_P = \frac{P^2}{H^2 + P^2} \tag{6}$$

In order to separate singing voice, the above procedure is implemented twice, once at high frequency resolution to separate pitched instruments from the vocals and percussions (remember that voice appears like percussive sounds at high frequency resolution) and once again at low frequency resolution to separate the voice from percussive sounds (remember that voice looks more like pitched instruments at low frequency resolution).

The algorithm summarized above uses vertical and horizontal median filters to separate vertical ridges of percussive instruments and horizontal ridges of pitched instruments from the mixture signal. However, when carefully examining the fluctuations in vocals, one can see that they usually contain a combination of diagonal and horizontal ridges. Hence, we suggest the use of diagonal median filters to capture more details of the vocal components as explained in the following section.

III. Using Diagonal Median Filters to Improve Vocal Separation

In order to improve the separation of vocals, we propose to use diagonal median filters during the low frequency resolution stage of the algorithm. This is because the diagonal characteristics of vocals are more evident at low frequency resolution spectrograms.

As an example, Fig. 2(a) shows an original voice component that has a clear diagonal ridge. On one hand, Fig. 2(b) shows the separated voice when using the horizontal median filter alone in the low frequency resolution stage. The circle indicates that the diagonal ridge almost disappeared after the separation. On the other hand, the diagonal ridge was preserved when using the diagonal median filters as shown in Fig. 2(c).

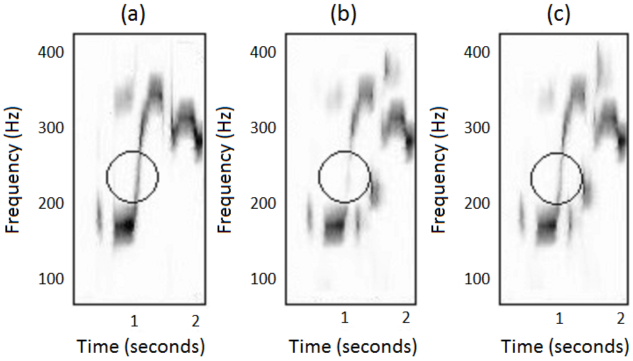

Fig. 2. Low frequency resolution spectrogram of vocals: (a) original voice, (b) separated voice using horizontal median filter, (c) separated voice using diagonal median filters.

To accommodate a wide variety of singing voices, six diagonal median filters MD_{d1} through MD_{d6} are applied along the diagonals of the magnitude spectrogram matrix in six different directions as shown in Fig. 3. The diagonal median filters' lengths are set to be the same as the length of the horizontal median filter. The results are the diagonally enhanced spectrograms D_1 to D_6, defined as:

$$D_i = MD_{di}\{S, l_h\} \tag{7}$$

where i is the index of the diagonal filter used, $i = 1, \ldots, 6$.

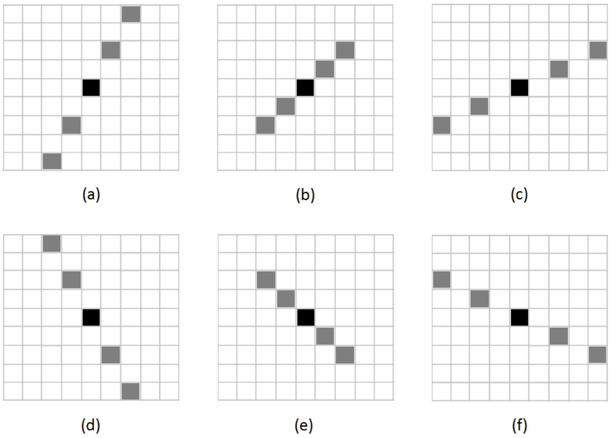

Fig. 3. (a-f) Samples used when applying MD_{d1} to MD_{d6} respectively with a length of 5 samples each on the center point.

Best results were obtained when combining the horizontal median filter originally used in [12] with the diagonal median filters. The harmonic-enhanced spectrogram H, and the diagonally enhanced spectrograms D_1 to D_6 are combined using an operator taking the maximum of all the matrices element-wise.

$$H' = \max(H, D_1, ..., D_6) \qquad (8)$$

H' then replaces H in (5), (6) to generate the Wiener filter masks which are used to estimate the sources as explained earlier in section II.

Although the diagonal filters described here could be considered as specific examples of kernels described in [13], here we use multiple diagonal filters (kernels) to identify a single source: the voice.

IV. Practical Time-Frequency Filters' Lengths

In order to further improve vocal separation, we also propose to use a practical set of filters' lengths that are independent of other parameters like fast Fourier transform (FFT) size, step size of the short-time Fourier transform (STFT) and the sampling frequency of the song. For that reason, we use seconds for measuring the lengths of the horizontal median filters and hertz for measuring the lengths of the vertical median filters. This is in contrast to using the number of time frames and frequency bins (columns and rows of the time-frequency matrix of the spectrogram) in [12] for measuring the lengths of horizontal and vertical median filters respectively.

Let f_v denotes the vertical filter length in Hz, then its length in frequency bins l_v can be calculated as:

$$l_v = \frac{f_v}{f_s} \times L \qquad (9)$$

where f_s is the sampling frequency and L is the window length of the STFT. Similarly, the horizontal filter length in seconds is denoted by f_h and the corresponding length in time frames l_h is calculated as:

$$l_h = \frac{f_h}{R} \times f_s \qquad (10)$$

where R is the step size of the STFT. In the following section we examine the effect of changing these lengths in an attempt to find practical values.

V. Experimental Results

In this section we first search for practical median filters parameters using one set of song clips, then test these parameters on another set of songs, followed by presenting the effect of using diagonal median filters.

A. *Estimating practical median filters' lengths from the MIR-1K dataset*

In our search for practical median filters' lengths, we used the MIR-1K dataset [4], which contains 110 karaoke Chinese pop songs performed mostly by amateurs. The sampling rate of each song is 16 kHz, and music accompaniment and singing voice were recorded in the left and right channels, respectively. We noticed some songs had traces of vocals in the music channel, which may affect accuracy of results, so we selected 50 clips from 50 different songs that do not contain vocal traces in its music channel. These clips length range from 5 to 12 seconds. We mixed the voice and music signals of these songs linearly 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 by Vincent et al [15]. These were calculated using the BSS_Eval toolbox [16] where higher values indicate better separation quality.

We set parameters of the experiment like those in [12]. Specifically, the median filters' lengths were all equal to 17 bins or frames. The FFT size for the high frequency resolution stage was 16384 samples with STFT step size 2048 samples. And the low frequency resolution stage FFT size was 1024 samples and the STFT step size was 256 samples.

Then we examined the effect of changing the vertical filter length at the high frequency resolution stage on the separation quality represented by the mean SDR, SIR, and SAR in dB. The next figure shows the results.

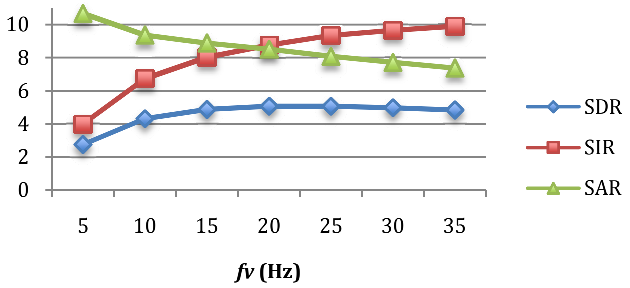

Fig. 4. Vocal separation metrics when changing the vertical median filter length in Hz at the high frequency resolution stage.

We found 20 Hz to achieve the highest SDR and it brings also a good compromise between SIR and SAR. So, we fixed the vertical filter length at this value and started to change the horizontal median filter length in seconds at this stage as shown in Fig. 5. Here we found 2 seconds to be a good value for the overall improvement and balance of the three metrics.

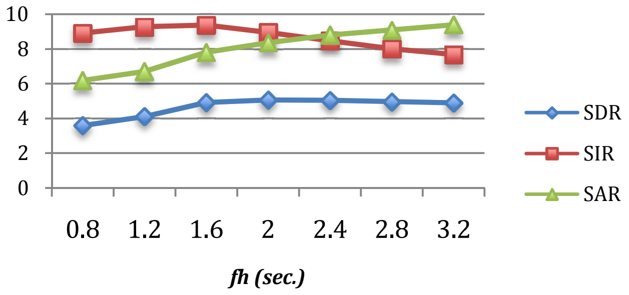

Fig. 5. Vocal separation metrics when changing the horizontal median filter length in seconds at the high frequency resolution stage.

After that we turn to the low frequency resolution stage parameters starting by the vertical median filter length. As Fig. 6 indicates, 250 Hz seems to be a good choice. Finally we changed the horizontal median filter length at this stage and we picked 0.15 seconds from Fig. 7. The summary of all the practical median filters' lengths that are empirically estimated is shown in Table I.

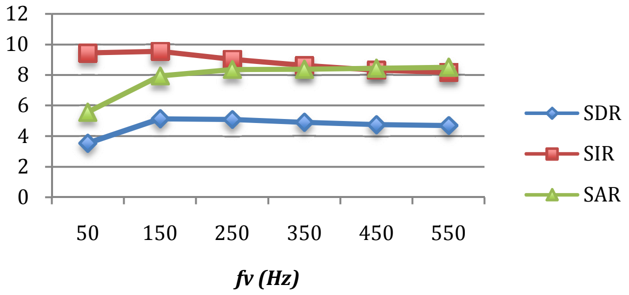

Fig. 6. Vocal separation metrics when changing the vertical median filter length in Hz at the low frequency resolution stage.

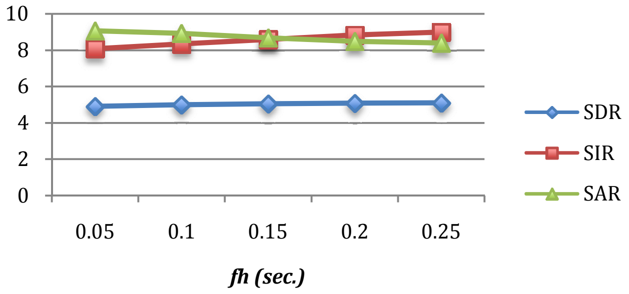

Fig. 7. Vocal separation metrics when changing the horizontal median filter length in seconds at the low frequency resolution stage.

TABLE I. PRACTICAL LENGTHS FOR ALL MEDIAN FILTERS

Spectrogram/Stage	Vertical filter length (Hz)	Horizontal filter length (seconds)
High frequency resolution	20	2
Low frequency resolution	250	0.15

Note that the suggested filters' lengths in Table I are approximate and are not necessarily the optimal for each song. For example, the vertical median filter length of 20 Hz at the high frequency resolution stage is a good compromise between the frequency span of pitched instruments horizontal ridges on one side and the frequency span of percussive instruments and most vocal fluctuations on the other side. However, if pitched instruments horizontal ridges have higher frequency spans, then a higher filter length, say 30 Hz, would probably achieve better separation results. Similar arguments can be made about the other filters' lengths in the table.

B. *Testing on the Beach Boys dataset*

To evaluate the performance of the algorithm with the new parameters as well as the diagonal filters, we tested it on real-world songs from the Good Vibrations album by the Beach Boys [14]. The dataset consists of 12 clips whose lengths range from 31 to 53 seconds, sampled at 44.1 kHz, where vocals and instrumental track were available separately. We mixed the voice and music signals of these songs linearly with equal energy to generate the mixture signal.

The first experiment was run with all median filters' lengths set to 17 bins (or frames) as in the baseline algorithm described in [12]. Note that these lengths correspond to a vertical filter length of about 46 Hz and a horizontal filter length of 0.8 seconds in the high frequency resolution stage, while in the low frequency resolution stage, the vertical filter length was 732 Hz and the horizontal filter length was 0.1 seconds. Obviously, these lengths are quite different from the new ones suggested in Table I.

In the second experiment, all median filters' lengths were set as in Table I. Additionally, the FFT length of the low frequency resolution stage was set to 2048 samples instead of 1024 for better overall separation performance.

Finally, in the third experiment we used the six diagonal median filters in addition to the horizontal one in the low frequency resolution stage as explained earlier in section III. All the lengths of the median filters were kept the same as in the second experiment.

Fig. 8 and Fig. 9 show the three metrics: SDR, SIR and SAR for voice and music respectively for the three experiments. When examining the effect of using the paractical parameters as indicated by the difference between the first (O) and the second (P) boxplots, we notice that most metrics increased significantly for both voice and music.

The SIR of the voice reduced as it was too high in the original setting on the account for the other two metrics of the voice, but it is still resonably good though. Also the SAR of the music reduced due to changing the low frequency resolution stage FFT length from 1024 to 2048 samples, but as indicated earlier this change brings better results overall.

Performing the one-tailed paired T-test on the results of the first and second experiments indicated a statistical significance with t value < 0.05 for all the metrics except for the voice SIR (which was reduced anyway).

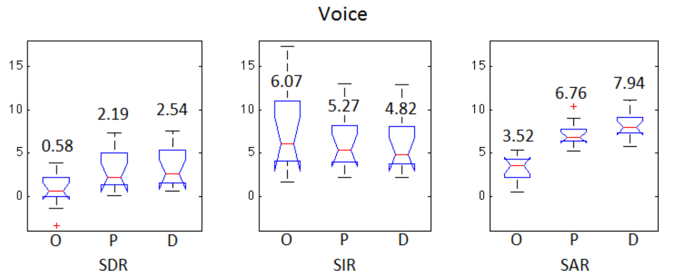

Fig. 8. Separation performance for singing voice using SDR (left), SIR (middle), and SAR (right) metrics. Three boxplots are shown for each metric; the leftmost one is with original parameters (O), followed by the new time-frequency filter parameters (P), then using diagonal filters (D). Median values are displayed.

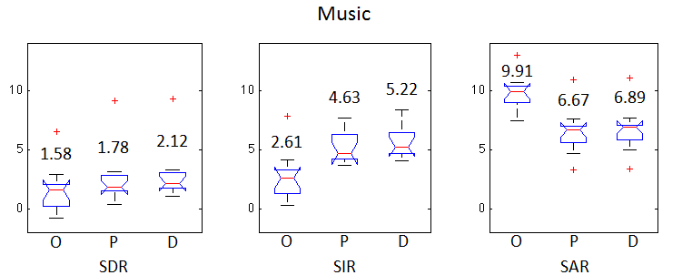

Fig. 9. Separation performance for music instruments using the same metrics as in Fig. 8.

On the other side, when examinging the effect of using the diagonal median filters as indicated by the difference between the second (P) and the third (D) boxplots, we notice that all metrics increased except for the SIR of the voice.

Performing the one-tailed paired T-test on the results of the second and third experiments indicated a statistical significance with t value < 0.0005 for all the metrics. Besides, when preforming the same test to compare the first and third experiments, there was a statistical significance with t value < 0.005 for all the metrics except for the voice SIR. The reader can also check sound samples for the three experiments in [17].

VI. Conclusion

In this paper we increased the capacity of median filters to separate vocals from monaural music accompaniment through the use of diagonal median filters with practical lengths defined in time and frequency units. The new time-frequency filters' lengths led to better separation performance for two sets of songs with different sampling frequencies. In future we will try to use adaptive parameters for these filters and explore the effect of combining the median filtering approach with other source separation techniques.

References

[1] H. Fujihara, M. Goto, J. Ogata, K. Komatani, T. Ogata, and H. G. Okuno, "Automatic synchronization between lyrics and music CD recordings based on Viterbi alignment of segregated vocal signals," in Proc. IEEE Int. Symp. Multimedia, San Diego, CA, 2006, pp. 257–264.

[2] W. Cai, Q. Li, X. Guan, "Automatic singer identification based on auditory features," in Proc. IEEE Int. Conf. Natural Computation, Shanghai, 2011, pp. 1624-1628.

[3] J. Salamon, E. Gomez, D.P.W. Ellis, G. Richard, "Melody Extraction from Polyphonic Music Signals: Approaches, applications, and challenges," IEEE Signal Processing Mag., vol. 31, no. 2, pp.118-134, Mar. 2014.

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

[5] A. Ozerov, E. Vincent, F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," IEEE Trans. Audio, Speech, Lang. Process., vol. 20, no. 5, pp. 1118–1133, May 2012.

[6] 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.

[7] P.-S. Huang, S. D. Chen, P. Smaragdis, and M. Hasegawa-Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, Kyoto, Japan, Mar. 2012, pp. 57–60..

[8] Z. Rafii and B. Pardo, "Repeating pattern extraction technique (REPET): A simple method for music/voice separation," IEEE Trans. Audio, Speech, Lang. Process., vol. 21, no. 1, pp. 73–84, Jan. 2013.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] A. Liutkus, D. Fitzgerald, Z. Rafii, B. Pardo, and L. Daudet, "Kernel Additive Models for Source Separation", IEEE Trans. Signal Process., vol. 62, no. 16, pp. 4298–4310, Aug. 2014.

[14] The Beach Boys, Good Vibrations: Thirty Years Of The Beach Boys, Capitol Records, Capitol C2 0777 7 81294 2 4, 1993.

[15] 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.

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

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