SINGING VOICE SEPARATION FROM MONAURAL MUSIC BASED ON KERNEL BACK-FITTING USING BETA-ORDER SPECTRAL AMPLITUDE ESTIMATION

Hye-Seung Cho, Jun-Yong Lee, Hyoung-Gook Kim
Kwangwoon University, Seoul, Rep. of Korea
{hye_seung401,jasonlee88,hkim}@kw.ac.kr

ABSTRACT

Separating the leading singing voice from the musical background from a monaural recording is a challenging task that appears naturally in several music processing applications. Recently, kernel additive modeling with generalized spatial Wiener filtering (GW) was presented for music/voice separation. In this paper, an adaptive auditory filtering based on β-order minimum mean-square error spectral amplitude estimation (bSA) is applied to the kernel additive modeling for improving the singing voice separation performance from monaural music signal. The proposed algorithm is composed of five modules: short time Fourier transform, music/voice separation based on bSA, determination of back-fitting, back-fitting, and inverse short time Fourier transform. In the proposed method, the Singular Value Decomposition (SVD)-based factorized spectral amplitude exponent β for each kernel component is adaptively calculated for effective bSA-based auditory filtering performance during kernel back-fitting. Using a back-fitting threshold, the kernel back-fitting process can automatically be iteratively performed until convergence. Experimental results show that the proposed method achieves better separation performance than GW based on kernel additive modeling.

1. INTRODUCTION

A singing voice in a music signal contains useful information for a song, as it embeds the singer, the lyrics, and the emotion of the song. Therefore, vocal or singing voice separation from monaural music signal is an important task in many applications, such as automatic karaoke [1], instrument/vocalist identification [2], music/voice transcription, music remixing [3] and audio restoration.

So far, numerous vocal separation algorithms have been proposed with various approaches, such as non-negative matrix factorization [4], adaptive Bayesian modeling [5], and pitch-based interference [6-7]. These methods usually first map signals onto a feature space, then detect singing voice segments, and finally apply source separation.

Recently, a relatively promising approach using kernel additive modeling (KAM) was proposed [8], wherein the spectrogram of each source is modeled only locally. This approach encompasses a large number of recently proposed methods for source separation [9-14]. KAM permits the use of different proximity kernels for different sources, with separation using an iterative kernel back-fitting (KBF) algorithm. In the kernel back-fitting, generalized Wiener filtering (GW) is used for the step of mixed music signal separation, and two-dimensional median filtering is applied to the power spectrogram of each source estimate for kernel spectrogram model fitting at each iteration. The GW requires good models of the spectrograms of each proximity source along with its spatial characteristics and permits very good separation provided these parameters are well estimated.

In spoken speech enhancement, one source may be the target voice, while others correspond to background noise which must be filtered out. Among the vast amount of single channel speech enhancement algorithms based on minimum mean-square error (MMSE) estimation of short-time spectral amplitude (STSA) published in the literature, it is well-known that the Bayesian STSA estimation methods [15] outperform the Wiener filtering, spectral-subtraction, and subspace approaches. In addition, among the Bayesian STSA estimation methods, β-order MMSE spectral amplitude estimation [15-17] achieved better enhancement performance than the existing Bayesian estimators, such as those based on the MMSE of the short-time spectral amplitude [15-17], and the MMSE of the logarithm of the STSA (LSA) [15-17].

In this paper, an advanced music/voice separation method is proposed, in which β-order MMSE spectral amplitude estimation and kernel spectrogram back-fitting are combined for improvement of the separation performance. In addition, the parameter β concerned in β-order MMSE spectral amplitude estimation is adaptively estimated according to the masking mechanism of human auditory system, the compressive nonlinearities of the cochlea and the critical sub-band SNR.

The proposed method has the following four advantages: (1) In the separation step, β-order MMSE estimation (bSA) of the factorized spectral amplitude

Attribution: Hye-Seung Cho, Jun-Yong Lee, Hyoung-Gook Kim. “Singing Voice Separation from Monaural Music Based on Kernel Back-fitting Using Beta-Order Spectral Amplitude Estimation”, 16th International Society for Music Information Retrieval Conference, 2015.

was used instead of GW for the kernel back-fitting procedure to achieve better separation performances. (2) The Singular Value Decomposition (SVD)-based factorized spectral amplitude β_j were adaptively calculated for effective bSA estimation performance. (3) In the back-fitting step, an SVD-based factorization procedure was applied to the power spectrogram filtered by median filter to achieve efficient compression before processing of the next proximity source. (4) Using a back-fitting threshold, the kernel back-fitting process can automatically be iteratively performed until convergence.

This paper is organized as follows. Section 2 describes the proposed method, while Section 3 discusses the experimental results. Finally, the conclusion is presented in Section 4.

2. PROPOSED MUSIC/VOICE SEPARATION ALGORITHM

The proposed algorithm is composed of five modules: short time Fourier transform (STFT), music/voice separation based on β-order MMSE spectral amplitude estimation (bSA), determination of back-fitting, back-fitting, and inverse short time Fourier transform (ISTFT).

Figure 1 denotes the overall procedure of the proposed music/voice separation algorithm.

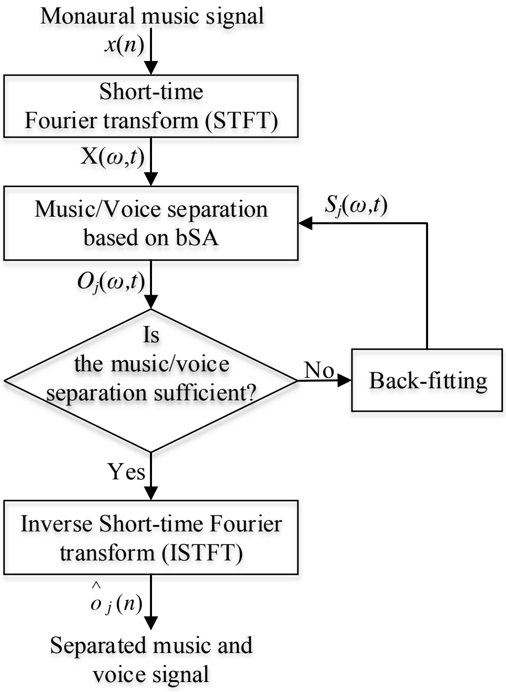
Monaural music signal
x(n)
Short-time Fourier transform (STFT)
X(ω,t)
Music/Voice separation based on bSA
S_j(ω,t)
O_j(ω,t)
Is the music/voice separation sufficient?
No
Back-fitting
Yes
Inverse Short-time Fourier transform (ISTFT)
ô_j(n)
Separated music and voice signal

Figure 1. Overall flow chart of proposed music/voice separation algorithm.

We assume that the mixture music signal, $x(n)$, is taken as the sum of j underlying sources that are composed of some of percussive elements, one of the stable harmonic elements, and one of the singing voice. Let a real-valued monaural music signal in discrete-time domain $x(n)$ be assumed as:

$$x(n) = \sum_{j=1}^{J} o_j(n) \tag{1}$$

where j (= 1, 2, … J) is index of each objective sources, n is sample index, and $o_j(n)$ denotes an objective source in mixture music signal.

First, an input monaural music signal $x(n)$ is transformed into the complex spectrogram $\mathrm{X}(\omega,t)$ using the short-time discrete Fourier transform (STFT), as shown:

$$\mathrm{X}(\omega,t) = \sum_{n=0}^{N-1} x(Rt+n)w(n)\exp\left(\frac{-i2\pi\omega n}{N}\right) \tag{2}$$

where R denotes the frame shift, t is the frame index, $w(n)$ indicates a window function, N is size of window, and ω is the frequency bin index, which is related to the normalized center frequency.

From the input complex spectrogram $\mathrm{X}(\omega,t)$, complex spectrogram $O_j(\omega,t)$ for each objective sources is estimated by β-order MMSE spectral amplitude estimation.

Each current estimated spectrogram is compared with each previous estimated complex spectrogram. If the difference between the current and previous estimated spectrograms is not larger than the back-fitting threshold value, each complex spectrogram is converted back to the time domain using an inverse STFT. Conversely, if the difference between the two is larger than back-fitting threshold value, the kernel back-fitting process is iterated until convergence.

During the back-fitting processes, the power spectrogram of the estimated spectrogram is filtered by a simple two dimensional median filter with source-specific binary kernels. The source-specific binary kernels are explained in detail in next sub-section.

This kernel back-fitting proceeds in an iterative fashion, with alternate performance of separation and re-estimation (back-fitting) of the parameters to obtain new spectrogram estimates for each source.

2.1 Re-estimation using back-fitting

The re-estimation using back-fitting permits one to use different proximity kernels for each source and to separate them in order to perform the estimation. It assumes that vertical lines in a spectrogram correspond to percussive events; horizontal lines are typically associated with harmonics of pitched instruments, while cross-like forms correspond to singing voice events. In this case, peaks due to pitched harmonics can be regarded as outliers on the vertical lines associated with percussive events. Similarly, peaks due to the percussion events can be regarded as outliers on the horizontal lines associated with pitched harmonic instruments. Median filters used extensively in image processing are good at eliminating outliers. That is, median filtering each time frame will suppress harmonics in this frame resulting in a percussion enhanced frame, while median filtering each frequency slice will suppress percussion events. This brings to the concept of using median filters individually in the horizontal, vertical, and cross-like directions to separate harmonic, percussive and vocal events.

The process is as follows:

(Step 1) Using the estimated complex spectrogram $O_j(\omega,t)$, the power spectrogram of the complex spectrogram is calculated as:

$$V_j(\omega,t) = \left|O_j(\omega,t)\right|^2 \tag{3}$$

(Step 2) A simple two dimensional median filter is applied to the power spectrogram $V_j(\omega,t)$ of the complex spectrogram with source-specific binary kernels, vocal, harmonic, and percussive. The different three proximity kernels [8] used for the median filter are as follows: (1) For a percussive and a repeating source, the vertical kernel is chosen; (2) For a harmonic source, the horizontal kernel is chosen; (3) Finally, for a source with only a spectral smoothness assumption, the cross-like kernel is chosen as vocals. The detailed three kernels are explained in the source separation using kernel additive models [8].

The median filtered kernel spectrogram is given by:

$$M_j(\omega,t) = median[V_j(\omega,t)\,|\,\mathrm{K}_j(\omega,t)] \tag{4}$$

where $\mathrm{K}_j(\omega,t)$ is a kernel which includes percussive elements of periodic components (j = 1, 2, ... J-2), the stable harmonic elements ($j = J$-1), and the singing voice ($j = J$), respectively. In effect, the original sample of the power spectrogram $V_j(\omega,t)$ of the complex spectrogram is replaced with the middle value obtained from a sorted list of the samples in neighborhoods of the original sample according to each kernel.

(Step 3) Kernel back-fitting using Wiener filtering or the β-order spectral amplitude estimator comes with an important drawback: it requires the full-resolution spectrogram, and storage of a huge amount of parameters in each iteration, and for each source. To reduce the memory usage and improve the separation performance while maintaining computational efficiency, Singular Value Decomposition (SVD) is applied to the full-resolution spectrogram $M_j(\omega,t)$:

$$S_j(\omega,t) = D_j\Sigma_jC_j = \mathrm{SVD}\left\lfloor M_j(\omega,t)\right\rfloor \tag{5}$$

where $M_j(\omega,t)$ is factored into the matrix product of three matrices: the $M \times M$ row basis D_j matrix, the $M \times L$ diagonal singular value matrix Σ_j and the $L \times L$ transposed column basis functions C_j.

2.2 Separation using β-order MMSE spectral amplitude estimation

In the separation step, β-order MMSE spectral amplitude estimation of the factorized spectral amplitude is used instead of GW for the kernel back-fitting procedure to achieve better music/voice separation performances. In the β-order MMSE spectral amplitude estimation, the spectral amplitude order β is quite important for singing voice enhancement or separation from monaural music signal. For the different β values, the gain values are different, and noise or other source reduction obtained is also different. In this way, the appropriated gain can be obtained by adaptively choosing right β.

However, the traditional calculation method about β is based on overall Signal-to-Noise Ratio (SNR) of each frame. That is, their values are fixed and not vary with frequency in each frame. Furthermore, the human auditory system has different sensitivity for different frequency components. Therefore, the b-th critical sub-band SNR is employed to calculate β values. For more effective bSA estimation performance, the Singular Value Decomposition (SVD)-based factorized spectral amplitude order $\beta_j(b,t)$ is adaptively calculated. Using adaptive β values and Singular Value Decomposition (SVD)-based factorized spectral amplitude, we can yield effective music/voice separation and obtain a good enhancement performance.

2.2.1 β-order MMSE spectral amplitude estimation

The β-order MMSE spectral amplitude estimation is composed of following four modules: sum of all $S_j(\omega,t)$, calculation of a priori SNR and a posteriori SNR, calculation of adaptive $\beta_j(b,t)$, and bSA-based gain function.

Figure 2 shows the β-order MMSE spectral amplitude estimation.

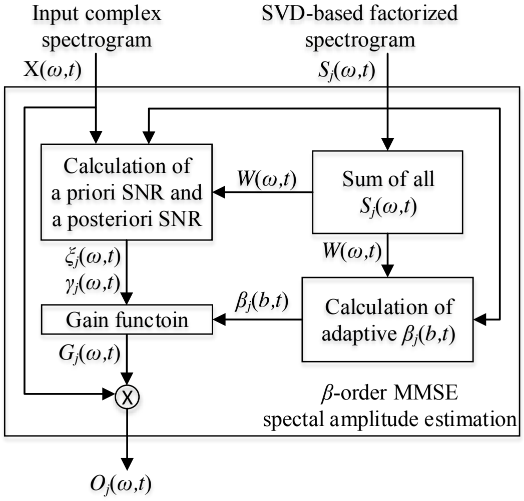
Input complex spectrogram
X(ω,t)
SVD-based factorized spectrogram
S_j(ω,t)
Calculation of a priori SNR and a posteriori SNR
W(ω,t)
Sum of all S_j(ω,t)
W(ω,t)
ξ_j(ω,t)
γ_j(ω,t)
β_j(b,t)
Gain functoin
Calculation of adaptive β_j(b,t)
G_j(ω,t)
X
β-order MMSE spectal amplitude estimation
O_j(ω,t)

Figure 2. Overall flow chart of the β-order MMSE spectral amplitude estimation.

Before to obtain the estimated complex spectrum $O_j(\omega,t)$ from SVD-based factorized $S_j(\omega,t)$, the sum $W(\omega,t)$ of all $S_j(\omega,t)$ is defined by:

$$W(\omega,t) = S_1(\omega,t) + S_2(\omega,t) + \ldots + S_J(\omega,t) \tag{6}$$

Then, the a priori SNR $\xi_j(\omega,t)$ and the a posteriori SNR $\gamma_j(\omega,t)$ of each objective proximity sources are calculated as follows:

$$\xi_j(\omega,t) = \frac{S_j(\omega,t)}{W(\omega,t) - S_j(\omega,t)}; \tag{7}$$

$$\gamma_j(\omega,t) = \frac{|X(\omega,t)|^2}{W(\omega,t) - S_j(\omega,t)}; \tag{8}$$

$$\chi_j(\omega,t) = \frac{\xi_j(\omega,t)}{1 + \xi_j(\omega,t)} \gamma_j(\omega,t); \tag{9}$$

where $\chi_j(\omega,t)$ is the function of $\xi_j(\omega,t)$ and $\gamma_j(\omega,t)$.

The gain function $G_j(\omega,t)$ for the bSA is given by:

$$G_j(\omega,t) = \frac{\sqrt{\chi_j(\omega,t)}}{\gamma_j(\omega,t)} \left[\Gamma\left(\frac{\beta_j(b,t)}{2} + 1 \right) \cdot \Phi\left(\frac{\beta_j(b,t)}{2}, 1; -\upsilon_j(\omega,t) \right) \right]^{\frac{1}{\beta_j(b,t)}} \tag{10}$$

where Γ(•) is the gamma function, Φ(•) is the confluent hypergeometric function. And $\beta_j(b,t)$ denotes the parameter based on the human auditory system.

To calculate $\beta_j(b,t)$, we employ the critical sub-band SNR. The b critical bands are divided for each speech frame, where a non-linear mel-frequency scale is used, which approximates the behavior of the auditory system. The mel-scale is a scale of pitches judged by listeners to be equal in distance one from another. The reference point between this scale and normal frequency measurement is defined by equating a 1000 Hz tone, 40 dB above the listener's threshold, with a pitch of 1000 mels. To convert a frequency ω in hertz into its equivalent in mel, the following formula is used:

$$pitch(mel) = 1127.0148 \log\left(1 + \frac{\omega(Hz)}{700} \right) \tag{11}$$

The spectrum is then processed by a mel-filter bank. The signal energy of the spectrum within b-th critical frequency sub-bands by means of a series of triangular filters whose center frequency are spaced according to the mel-scale. Thereafter, the critical sub-band SNR $Z_j(b,t)$ is calculated in the b-th band.

Finally, the estimated complex spectrogram from the gain function is defined as:

$$O_j(\omega,t) = G_j(\omega,t) \cdot X(\omega,t) \tag{12}$$

2.2.2 Calculation of adaptive $\beta_j(b,t)$

Since the spectral amplitude order $\beta_j(b,t)$ is based on characteristics of the human auditory system, including the compressive nonlinearities of the cochlea, and the perceived loudness, the choosing of adequate value for $\beta_j(b,t)$ can result in better enhancement or separation performance.

First, using $W(\omega,t)$ and $S_j(\omega,t)$, the sub-band SNR $Z_j(b,t)$ is calculated as:

$$Z_j(b,t) = 10 \log_{10} \frac{\sum_{\omega=B_{low}(b)}^{B_{up}(b)} \left| W(\omega,t) - \sqrt{W(\omega,t) - S_j(\omega,t)} \right|^2}{\sum_{\omega=B_{low}(b)}^{B_{up}(b)} \left(W(\omega,t) - S_j(\omega,t) \right)} \tag{13}$$

where $b \in [0, 23]$ denotes the index of critical band. $B_{up}(b)$ and $B_{low}(b)$ denote the upper and lower frequency bound of the b-th critical band, respectively.

To obtain $\beta_j(b,t)$, the compression rate $\hat{\beta}_j(b,t)$ at intermediate frequencies can be calculated through linear interpolation between β_{low} and β_{high}. That is,

$$\hat{\beta}_j(b,t) = \beta_{high} - d(b,t)(\beta_{high} - \beta_{low}) \text{ for } 1 \le j \le J \tag{14}$$

using

$$d(b,t) = \frac{1}{B_{up}(b) - B_{low}(b)} \sum_{\omega=B_{low}(b)}^{B_{up}(b)} \left\{ \frac{1}{\eta} \log_{10}\left(\frac{f_\omega}{A} + l \right) \right\} \tag{15}$$

where $d(b,t)$ is the frequency-position function to the critical band, β_{high} = 0.2 and β_{low} = 1 denote the low-frequency and high-frequency of the compression rate, η = 0.06 mm, l = 1, and A = 165.4 Hz are the parameters set in paper [18], and f_ω is the frequency in Hz corresponding to spectral component ω, i.e., $f_\omega = \omega F_s/N$, where F_s is the sampling frequency.

By limiting the range of $\check{\beta}_j(b,t)$ as [β_{min}, β_{max}] in order to obtain a better trade-off between target source enhancement and other source reduction, $\check{\beta}_j(b,t)$ can be calculated through the following relationship:

$$\check{\beta}_j(b,t) = \min\{\max[\mu \cdot Z_j(b,t) + \lambda, \beta_{\min}], \beta_{\max}\} \tag{16}$$

where μ = 0.45, λ = 1.3, β_{min} = 0.4, and β_{max} = 4.0.

According to sub-band SNR, the compressive nonlinearities of the cochlea, and perceived loudness, a parameter $\beta_j(b,t)$ is given as follows:

$$\beta_j(b,t) = q \cdot \hat{\beta}_j(b,t) + (1\text{-}q) \cdot \check{\beta}_j(b,t) \tag{17}$$

where q ($0 < q < 1$) is a smoothing parameter.

3. EXPERIMENTAL RESULTS

In this section, the performance of the proposed bSA-KBF algorithm is evaluated for the separation of background music and singing voice.

For experiments, 100 full-length song tracks were used (50 songs from the ccMixter database containing many different musical genres, 50 songs from a self-recording studio music database), where all singing voices and music accompaniments were recorded separately. All of the song data were stored in PCM format with mono, 16-bit depth, and 44.1 kHz sampling rate.

For each track, the accompaniment of 6 repeating patterns along with a 2 second steady harmonic source was determined. Vocals were modeled using a cross-like kernel with a height of 15 Hz and width of 20 ms. The frame length was set to 90 ms, with 80% overlap. Six to eight iterations were performed for the back-fitting algorithm (approximately until convergence).

For the performance measures, performance was evaluated in terms of Normalized Source-to-Interference Ratio (NSIR) and Normalized Source-to-Distortion Ratio (NSDR) by Blind Source Separation Evaluation (BSS Eval) metrics [19]. NSDR and NSIR for singing voice are defined as:

$$\begin{aligned} \mathrm{NSDR}(v_r, v, x) &= \mathrm{SDR}(v_r, v) - \mathrm{SDR}(x, v) \\ \mathrm{NSIR}(v_r, v, x) &= \mathrm{SIR}(v_r, v) - \mathrm{SIR}(x, v) \end{aligned} \tag{18}$$

where v_r is the synthesized singing voice, v is the original clean singing voice, and x is the mixture. NSDR is for estimating the improvement of the SDR between the processed mixture x and the separated singing voice v_r. Higher values indicate better separation.

The performance of the proposed bSA algorithm was compared with those of GW, LSA based on KAM.

Table 1 presents the experimental results of comparative performance for music/voice separation of the four methods:

- STFT-GW-KAM: As a basic KAM algorithm, the generalized Wiener filter was applied to the power spectrogram based on STFT.
- SVD-GW-KAM: SVD was performed on the power spectrogram based on STFT. To the SVD-based decomposed power spectrogram, the generalized Wiener filter was applied.
- SVD-LSA-KAM: The MMSE of the logarithm of the STSA was applied to the SVD-based decomposed power spectrogram.
- SVD-bSA-KAM: β-order MMSE STSA was applied to the SVD-based decomposed power spectrogram.

Methods	Separation Performance for Music		Separation Performance for Voice	
	NSDR	NSIR	NSDR	NSIR
STFT-GW-KAM	6.37	9.18	1.89	5.76
SVD-GW-KAM	6.83	9.65	2.35	6.23
SVD-LSA-KAM	7.36	10.48	2.87	6.74
SVD-bSA-KAM	8.25	12.13	3.12	6.88

Table 1. Comparative performance for music/voice separation.

As shown in Table 1, the best separation performance of the music from the mixed music signal is obtained with the proposed method, SVD-bSA-KAM, in terms of NSDR and NSIR. Compared to the other three methods, the basic method, STFT-GW-KAM, attains the worst results. And the proposed bSA delivers high performance result in the separation of vocal components.

4. CONCLUSIONS

In this paper, we proposed a β-order MMSE spectral amplitude estimation method based on kernel back-fitting for music/voice separation. The proposed algorithm enhances the basic kernel back-fitting algorithm through application of β-order MMSE spectral amplitude estimation considering the perceptual properties of human auditory system. The experimental results show that the proposed method obtained better results compared to other existing methods.

In future work, we will apply the method to spatial audio reproduction applications running on smart phones.

5. ACKNOWLEDGMENTS

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(NRF-2013R1A1A2007601). And this research was supported by the MSIP(Ministry of Science, ICT and Future Planning), Korea, under the ITRC(Information Technology Research Center) support program (IITP-2015-H8501-15-1016) supervised by the IITP(Institute for Information & communications Technology Promotion)

6. REFERENCES

[1] Z. Rafii and B. Pardo: "REpeating Pattern Extraction Technique (REPET): A Simple Method for Music/Voice Separation," *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 21, No. 1, pp. 73–84, 2012.

[2] N. C. Maddage, C. Xu, and Y. Wang: "Singer identification Based on Vocal and Instrumental Models," in *Proceedings of the 17th International Conference on Pattern Recognition*, Vol. 2, pp. 375–378, 2004.

[3] S. Marchand et al: "DReaM: A Novel System for Joint Source Separation and Multi-Track Coding," in *Proceedings of the 133rd Audio Engineering Society Convention*, 2012.

[4] S. Vembu and S. Baumann: "Separation of Vocals from Polyphonic Audio Recordings," in *Proceedings of the International Society for Music Information Retrieval Conference*, pp. 337–344, 2005.

[5] A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval: "Adaptation of Bayesian Models for Single-Channel Source Separation and its Application to Voice/Music Separation in Popular Songs," *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 15, No. 5, pp. 1564–1578, 2007.

[6] Y. Li and D. Wang: "Separation of Singing Voice From Music Accompaniment for Monaural Recordings," *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 15, No. 4, pp. 1475–1487, 2007.

[7] C. -L. Hsu and J. -S. R. Jang: "On the Improvement of Singing Voice Separation for Monaural Recordings Using the MIR-1K Dataset," *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 18, No. 2, pp. 310–319, 2009.

[8] A. Liutkus, D. Fitzgerald, Z. Rafii, B. Pardo, and L. Daudet: "Kernel Additive Models for Source Separation," *IEEE Transactions on Signal Processing*, Vol. 62, No. 16, pp. 4298–4310, 2014.

[9] Z. Rafii and B. Pardo: "Repeating Pattern Extraction Technique (REPET): A Simple Method for Music/Voice Separation," *IEEE Transaction on Audio, Speech and Language Processing*, Vol. 21, No. 1, pp. 73–84, 2013.

[10] D. Fitzgerald: "Harmonic/Percussive Separation Using Median Filtering," in *Proceedings of the 13th International Conference on Digital Audio Effects (DAFx-10)*, 2010.

[11] Z. Rafii and B. Pardo: "A Simple Music/Voice Separation Method Based on the Extraction of the Repeating Musical Structure," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 221–224, 2011.

[12] A. Liutkus et al: "Adaptive Filtering for Music/Voice Separation Exploiting the Repeating Musical Structure," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 53–56, 2012.

[13] Z. Rafii and B. Pardo: "Music/Voice Separation Using the Similarity Matrix," in *Proceedings of the International Society for Music Information Retrieval Conference*, pp. 583–588, 2012.

[14] O. Yilmaz and S. Rickard: "Blind Separation of Speech Mixtures via Time-Frequency Masking," *IEEE Transaction on Signal Processing*, Vol. 52, No. 7, pp. 1830–1847, 2004.

[15] E. Plourde and B. Champagne: "Auditory-Based Spectral Amplitude Estimators for Speech Enhancement," *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 16, No. 8, pp. 1614–1623, 2008.

[16] F. Deng, F. Bao and C. -C. Bao: "Speech Enhancement Using Generalized β-Order Spectral Amplitude Estimator," in *Proceedings of the Speech Communication*, Vol. 59, pp. 55–68, 2014.

[17] C. H. You, S. N. Koh, and S. Rahardja: "β-Order MMSE Spectral Amplitude Estimation for Speech Enhancement," *IEEE Transactions on Speech and Audio Processing*, Vol. 13, No. 4, pp. 475–486, 2005.

[18] D. D. Greenwood: "A Cochlear Frequency-Position Function for Several Species-29 Years Later," *Journal of Acoustic Society America*, Vol. 87, No. 6, pp. 2592–2605, 1990.

[19] E. Vincent, R. Gribonval, and C. Fevotte: "Performance Measurement in Blind Audio Source Separation," *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 14, No. 4, pp. 1462–1469, 2006.