Music/Voice Separation Based on Kernel Back-Fitting Using Weighted β-Order MMSE Estimation

Hyoung-Gook Kim and Jin Young Kim

Recent developments in the field of separation of mixed signals into music/voice components have attracted the attention of many researchers. Recently, iterative kernel back-fitting, also known as kernel additive modeling, was proposed to achieve good results for music/voice separation. To obtain minimum mean square error (MMSE) estimates of short-time Fourier transforms of sources, generalized spatial Wiener filtering (GW) is typically used. In this paper, we propose an advanced music/voice separation method that utilizes a generalized weighted β-order MMSE estimation (WbE) based on iterative kernel back-fitting (KBF). In the proposed method, WbE is used for the step of mixed music signal separation, while KBF permits kernel spectrogram model fitting at each iteration. Experimental results show that the proposed method achieves better separation performance than GW and existing Bayesian estimators.

Keywords: Kernel back-fitting, separation, generalized weighted β-order MMSE estimation, singular value decomposition.

Manuscript received Mar. 16, 2015; revised Dec. 17, 2015; accepted Dec. 28, 2015.

Hyoung-Gook Kim (corresponding author, hkim@kw.ac.kr) is with the Department of Electronics Convergence Engineering, Kwangwoon University, Seoul, Rep. of Korea.

Jin Young Kim (beyondi@jnu.ac.kr) is with the Department of Electronics and Computer Engineering, Chonnam National University, Gwangju, Rep. of Korea.

I. Introduction

Music/voice separation of mixed music signals refers to the problem of trying to separate vocals from instrumentals in a song to produce both an *a cappella* track and an instrumental track. It is a topic that has many applications, such as automatic karaoke [1], instrument/vocalist identification [2], music/voice transcription [3], music remixing [4], [5], and audio restoration [6].

A number of approaches have been applied to the problem of separating the foreground (typically the voice) from the background (the musical accompaniment) components, including spectrogram factorization [7], accompaniment model learning [8], and pitch-based inference techniques [9].

Recently, a relatively promising approach using kernel additive modeling (KAM) was proposed [10], wherein the spectrogram of each source is modeled only locally. This approach encompasses a large number of recently proposed methods for source separation [1], [11]–[15]. KAM permits the use of different proximity kernels for different sources, with separation using an iterative kernel back-fitting (KBF) algorithm. In KBF, generalized Wiener filtering is used for the step of mixed music signal separation, and 2D median filtering is applied to the power spectrogram of each source estimate for kernel spectrogram model fitting at each iteration.

In spoken speech enhancement, one source may be the target voice, while other sources may correspond to background noise — which must be filtered out. Among the vast number of single-channel speech enhancement algorithms based on minimum mean square error (MMSE) estimation of short-time

http://dx.doi.org/10.4218/etrij.16.0115.0256

spectral amplitude (STSA) published in the literature, it is well known that the Bayesian STSA estimation methods [16] outperform the Wiener filtering, spectral-subtraction, and subspace approaches.

In addition, among the Bayesian STSA estimation methods, such as those based on the MMSE of the STSA [17], the MMSE of the logarithm of the STSA (LSA) [17], the weighted Euclidean (WE) error [17], and β-order MMSE STSA (bSA) [17], [18], we find that weighted β-order MMSE estimation (WbE) [17] achieves the best enhancement performance in terms of both objective and subjective measures. WbE [17]–[19] combines the power law of the bSA estimation method and the weighting factor of the WE error estimation method. Its parameters are chosen based on the human auditory system, which provides the advantage of improving noise reduction at high frequencies while limiting speech distortion at low frequencies.

In this paper, an advanced music/voice separation method is proposed, in which WbE and KBF are combined for improvement of the separation performance.

This paper is organized as follows. Section II describes the proposed method, while Section III discusses the experimental results. Finally, the conclusion is presented in Section IV.

II. Proposed Music/Voice Separation Method

Algorithm 1 shows the overall procedure of the proposed advanced music/voice separation method. The algorithm is composed of seven steps.

Let a real-valued monaural music signal in a discrete-time domain, $x(n)$, be defined as $x(n) = v(n) + h(n) + p(n)$, where $v(n)$, $h(n)$, and $p(n)$ denote the singing voice, the stable harmonic elements, and the percussive elements including periodic components, respectively.

First, an input monaural music signal $x(n)$ is transformed into a complex spectrogram, X, using a short-time discrete Fourier transform (STFT), as shown in step 3, as follows:

$$X = X(u,l) = \mathrm{STFT}_l\left[x(n)\right] = \sum_{n=0}^{N-1} x(Rl+n)\,w(n)\exp\left(\frac{-\mathrm{j}2\pi un}{N}\right), \quad (1)$$

where R denotes the frame shift, l is the frame index, $w(n)$ indicates a window function, N is a window size, and $u \in \{0, 1, \ldots, U-1\}$ is the frequency bin index, which is related to the normalized center frequency.

This is followed by step 4, wherein music/voice separation is carried out based on WbE.

Algorithm 1. Whole procedure of advanced music/voice separation method.

1) Input
- Given an input signal $x(n)$
- Kernels Λ_V, Λ_H, Λ_P for vocal, harmonic, percussive components

2) Initialization
- $G_V = G_H = G_P = 1$ ▷ WbE-based gain

3) Preprocessing
a) $X \leftarrow \mathrm{STFT}_l[x(n)]$ ▷ Complex spectrogram

4) Music/voice separation based on WbE : Algorithm 2
a) $S_V \leftarrow G_V \cdot X$ ▷ Complex vocal spectrogram estimation
b) $S_H \leftarrow G_H \cdot X$ ▷ Complex harmonic spectrogram estimation
c) $S_P \leftarrow G_P \cdot X$ ▷ Complex percussive spectrogram estimation

5) Determination of music/voice enhancement
If each estimated complex spectrogram S_V, S_H, S_P, is sufficiently enhanced or separated,
then go to step 7.
else
then go to step 6-4-5.

6) Back-fitting
a) $B_V \leftarrow |S_V|^2$ ▷ Vocal power spectrogram
b) $B_H \leftarrow |S_H|^2$ ▷ Harmonic power spectrogram
c) $B_P \leftarrow |S_P|^2$ ▷ Percussive power spectrogram
d) $M_V \leftarrow \mathrm{median}\,[B_V|\Lambda_V]$ ▷ Median filtering using vocal kernel Λ_V
e) $M_H \leftarrow \mathrm{median}\,[B_H|\Lambda_H]$ ▷ Median filtering using harmonic kernel Λ_H
f) $M_P \leftarrow \mathrm{median}\,[B_P|\Lambda_P]$ ▷ Median filtering using percussive kernel Λ_P
g) $D_V\Sigma_V E_V \leftarrow \mathrm{SVD}\,[M_V]$ ▷ Singular value decomposition
h) $D_H\Sigma_H E_H \leftarrow \mathrm{SVD}\,[M_H]$
i) $D_P\Sigma_P E_P \leftarrow \mathrm{SVD}\,[M_P]$

7) Output
a) $\hat{v}(n) \leftarrow \mathrm{STFT}_l^{-1}[S_V]$ ▷ Waveform synthesis
b) $\hat{h}(n) \leftarrow \mathrm{STFT}_l^{-1}[S_H]$
c) $\hat{p}(n) \leftarrow \mathrm{SFTF}_l^{-1}[S_P]$

From the complex spectrogram X of the input music signal, each complex spectrogram, S_V, S_H, and S_P, for the vocal, harmonic, and percussive components is estimated by each generalized WbE, G_V, G_H, and G_P, of decomposed spectral amplitude by singular value decomposition (SVD) for the vocal, harmonic, and percussive components, respectively. The WbE estimation gain, G_j, for each source j (= 0, 1, 2, … , J) is explained in detail in Algorithm 2.

In the fifth step, 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,

then each complex spectrogram for the vocal, harmonic, and percussive components is converted back to the time domain using an inverse STFT, as step 7 (the sum of the separated harmonic and percussive waveforms can be represented as the musical accompaniment, while the separated vocal waveform is the singing voice). Conversely, if the difference between the two is larger than the back-fitting threshold value, then the KBF process of step 6 is iterated until convergence. Applying the auxiliary function approach [20] to a quadrature form of the spectrogram gradients, the back-fitting threshold value is calculated at each iteration.

During the KBF step of each source, a simple 2D median filter is applied to the power spectrogram of the complex spectrogram (filtered by WbE) with source-specific binary kernels, Λ_V, Λ_H, and Λ_P. The three kernels [10] used for the median filter are as follows: (1) for a percussive and repeating source, the vertical kernel Λ_P is chosen; (2) for a harmonic source, the horizontal kernel Λ_H is chosen; and (3) for a source with only a spectral smoothness assumption, the cross-like vocal kernel Λ_V is chosen. This KBF 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.

1. WbE of SVD-Based Decomposed Spectral Amplitude

For the music/voice separation from monaural music signals, we propose a new kind of generalized spatial WbE of the decomposed spectral amplitude by SVD. The proposed estimation method takes full advantage of both a generalized weighted β-order spectral amplitude estimator and an SVD-based subspace decomposition.

Let the spectrogram of the monaural music signal X be expressed as $X = X_1 + X_2 + X_3$ for each source, where X_1, X_2, and X_3 are the complex spectrograms for vocal, harmonic, and percussive components, respectively.

The generalized weighted β-order spectral amplitude estimator combines the power law in the β-order spectral amplitude cost function and the weighting of the WE cost function. The cost function of the generalized WbE can be expressed as follows:

$$C\left(X_j,\hat{X}_j\right)=\left(\frac{X_j^{\beta}-\hat{X}_j^{\beta}}{X_j^{\alpha}}\right)^2, \tag{2}$$

where α is the perceptually weighted order and β is the spectral amplitude order.

The above cost function takes advantage of not only perceptual weighting, to allow the estimation error to be penalized more heavily in the spectral valleys than the spectral peaks, but also the cochlea's compressive nonlinearities. It is well known that the WbE method outperforms Wiener filtering in speech enhancement under different noisy environments. It can also be concatenated with kernel spectrogram back-fitting to yield good results for music/voice separation.

However, KBF using either Wiener filtering or the generalized weighted β-order spectral amplitude estimator comes with an important drawback: it requires a full-resolution spectrogram, and storage of a huge number of parameters at each iteration, and for each source. To reduce the memory usage and improve the separation performance while maintaining computational efficiency, SVD is applied to the full-resolution spectrogram X_j, as follows:

$$K_j=\mathbf{D}_j\mathbf{\Sigma}_j\mathbf{E}_j=\mathrm{SVD}\left[X_j\right], \tag{3}$$

where X_j is factored into a matrix product comprising three matrices — an $M \times M$ row basis matrix $\mathbf{D}_j$, an $M \times L$ diagonal singular value matrix $\mathbf{\Sigma}_j$, and an $L \times L$ transposed column basis matrix $\mathbf{E}_j$. Let K_1, K_2, and K_3 represent the decomposed spectral amplitude for vocal, harmonic, and percussive components, respectively. Then, W can be expressed as $W = K_1 + K_2 + K_3$ for each source.

Using the estimated decomposed spectral amplitude $\hat{K}_j$ based on SVD, a new cost function can be obtained as follows:

$$C\left(K_j,\hat{K}_j\right)=\left(\frac{K_j^{\beta}-\hat{K}_j^{\beta}}{K_j^{\alpha}}\right)^2. \tag{4}$$

By minimizing the expectations of the given cost function and substituting β_j for each source j instead of β, the separated spectral amplitude estimation can be obtained as follows:

$$\begin{aligned}\hat{S}_j &= \arg\min\int_0^{\infty} C\left(K_j,\hat{K}_j\right) f_{K_j|X}\left(K_j\,|X\right)\mathrm{d}K_j \\ &= \left(\frac{\mathrm{E}\left\{K_j^{\beta_j-2\alpha_j}\,|X\right\}}{\mathrm{E}\left\{K_j^{-2\alpha_j}\,|X\right\}}\right)^{\frac{1}{\beta_j}} = G_j\cdot X,\end{aligned} \tag{5}$$

where $\mathrm{E}\{\bullet\}$ denotes expectation and G_j is the gain value obtained according to the cost function of the generalized weighted β-order estimator of the SVD-based factorized spectral amplitude.

The resulting gain function for the generalized WbE is derived to be

$$G_j=\frac{\sqrt{\chi_j}}{\gamma_j}\left[\frac{\Gamma\left(\frac{\beta_j-2\alpha_j}{2}+1\right)\Phi\left(-\frac{\beta_j-2\alpha_j}{2},1;-\chi_j\right)}{\Gamma\left(-\alpha_j+1\right)\Phi\left(\alpha_j,1;-\chi_j\right)}\right]^{\frac{1}{\beta_j}}, \tag{6}$$

using

$$\chi_j=\frac{\xi_j}{1+\xi_j}\gamma_j,\qquad \xi_j=\frac{K_j}{W-K_j},\qquad \gamma_j=\frac{|X|^2}{W-K_j}, \tag{7}$$

http://dx.doi.org/10.4218/etrij.16.0115.0256

where ξ_j denotes an a priori signal-to-noise ratio (SNR), γ_j denotes an a posteriori SNR, and χ_j is a function of both ξ_j and γ_j. Furthermore, $\Gamma(\bullet)$ is the gamma function, and $\Phi(\bullet)$ denotes the confluent hypergeometric function, which can be written as

$$\Phi\left(-\frac{m}{2},1;-\chi_j\right)=\frac{\chi_j^{\frac{m}{2}}}{\Gamma\left(\frac{m}{2}+1\right)}. \quad (8)$$

The filtering or separation procedure based on WbE is described in Algorithm 2.

Algorithm 2. Filtering based on WbE.

1) Input

$M_{\mathrm{V}}, M_{\mathrm{H}}, M_{\mathrm{P}}$

2) Initialization

- $K_1 \leftarrow D_{\mathrm{V}}\Sigma_{\mathrm{V}}E_{\mathrm{V}} = \mathrm{SVD}[M_{\mathrm{V}}]$ ▷ SVD of M_{V}
- $K_2 \leftarrow D_{\mathrm{H}}\Sigma_{\mathrm{H}}E_{\mathrm{H}} = \mathrm{SVD}[M_{\mathrm{H}}]$ ▷ SVD of M_{H}
- $K_3 \leftarrow D_{\mathrm{P}}\Sigma_{\mathrm{P}}E_{\mathrm{P}} = \mathrm{SVD}[M_{\mathrm{P}}]$ ▷ SVD of M_{P}
- $W \leftarrow K_1 + K_2 + K_3$
- $G_{\mathrm{V}} \leftarrow G_1$,
- $G_{\mathrm{H}} \leftarrow G_2$,
- $G_{\mathrm{P}} \leftarrow G_3$

3) Computation of beta and alpha

compute α_j and β_j

4) WbE computation for each source *j*

for $0 \le j \le J$ do

a) $\xi_j \leftarrow \frac{K_j}{W-K_j}$ ▷ Priori SNR

b) $\gamma_j \leftarrow \frac{|X|^2}{W-K_j}$ ▷ Posteriori SNR

c) $\chi_j \leftarrow \frac{\gamma_j \cdot \xi_j}{1+\xi_j}$

d) $\Psi_j \leftarrow \Gamma\left(\frac{\beta_j-2\alpha_j}{2}+1\right)\cdot\Phi\left(-\frac{\beta_j-2\alpha_j}{2},1;-\chi_j\right)$

e) $\Theta_j \leftarrow \Gamma(-\alpha_j+1)\cdot\Phi(\alpha_j,1;-\chi_j)$

f) $G_j \leftarrow \frac{\sqrt{\chi_j}}{\gamma_j}\left[\frac{\Psi_j}{\Theta_j}\right]^{\frac{1}{\beta_j}}$

2. Appropriate Calculation of β_j and α_j

In single-channel speech enhancement methods based on WbE, the cost function includes both a power law β and a weighting factor α. To obtain a significant noise reduction and an improved speech estimation of weak speech components both in terms of objective and subjective measures, adaptive calculation methods of parameters α and β, of the cost function, have been proposed and published in the literature [19]–[21]. We think that the perceptually weighted order α_j and the SVD-based factorized spectral amplitude β_j for K_j are also important for enhancement or separation of voice/music based on the WbE.

Since α_j and β_j are based on characteristics of the human auditory system, including the compressive nonlinearities of the cochlea, the perceived loudness, and the ear's masking properties, the choosing of appropriate values for α_j and β_j can result in better enhancement or separation performance. In this subsection, the adaptive estimation of parameters α_j and β_j is described.

An adaptive calculation method of parameter α_j is given as follows:

$$\alpha_j = \alpha_{\mathrm{low}} + a\cdot\frac{(f_u-2{,}000)(\alpha_{\mathrm{high}}-\alpha_{\mathrm{low}})}{F_s/2-2{,}000}+(1-a)\cdot\hat{\alpha}_j, \quad (9)$$

where $\alpha_{\mathrm{low}} = 0.25$ and $\alpha_{\mathrm{high}} = 0.94$ are used for the trade-off between target source enhancement and other source reduction; a ($0 < a < 1$) is a smoothing parameter; and f_u is the frequency in hertz corresponding to spectral component u; that is, $f_u = uF_s/N$, where F_s is the sampling frequency.

Combining the masking threshold T with the sub-band SNR Z, the parameter $\hat{\alpha}_j$ can be expressed as a function of the two variables in polynomial form, approximately as follows:

$$\begin{aligned}\hat{\alpha}_j &= F(Z_j,T_j)\\ &\cong \sum_{q=0}^{\infty}\sum_{r=0}^{\infty}c_{qr}Z_j^qT_j^r\\ &\cong e_0+e_1Z_j+e_2T_j+e_3Z_jT_j,\end{aligned} \quad (10)$$

using

$$\mathrm{Z}_j = 10\log_{10}\frac{\sum_{u=0}^{U-1}\left|W-\sqrt{W-K_j}\right|^2}{\sum_{u=0}^{U-1}(W-K_j)}, \quad (11)$$

where c_{qr} represents the polynomial coefficients, and the empirical values $e_0 = 0.765$, $e_1 = -0.123$, $e_2 = -0.265$, and $e_3 = -0.07$ were obtained through simulation.

The frequency masking threshold T was derived in [19]. Higher masking thresholds will result in larger α_j values, corresponding to higher gain; whereas, lower masking thresholds will result in a smaller α_j values, corresponding to lower gain.

An adaptive calculation method for parameter β_j is given as follows:

$$\beta_j = b\cdot\hat{\beta}_j+(1-b)\overset{\nabla}{\beta}_j, \quad (12)$$

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

According to the frequency-position function (FPF) d_u, the compression rate $\hat{\beta}_j$ at intermediate frequencies can be

calculated through linear interpolation between β_{low} and β_{high}. That is,

$$\hat{\beta}_j = \beta_{\text{high}} - \frac{d_u}{\hat{d}_u}\left(\beta_{\text{high}} - \beta_{\text{low}}\right) \quad \text{for} \quad 0 \le j \le J, \tag{13}$$

where $\beta_{\text{high}} = 0.2$ and $\beta_{\text{low}} = 1$ denote the low-frequency and high-frequency of the compression rate, respectively.

Considering the fact that each frequency corresponds to a position on the basilar membrane, the FPF is given by

$$d_u = \frac{1}{\eta}\log_{10}\left(\frac{f_u}{A} + t\right), \text{ and } \hat{d}_u = \frac{1}{\eta}\log_{10}\left(\frac{F_s}{2A} + t\right), \tag{14}$$

where d_u is the position on the basilar membrane in millimeters, while $\eta = 0.06$ mm, $A = 165.4$ Hz, and $t = 1$ are the parameters set in [22].

By limiting the range of $\overset{\nabla}{\beta}_j$ as $[\beta_{\min}, \beta_{\max}]$ to obtain a better trade-off between target source enhancement and other source reduction, $\overset{\nabla}{\beta}_j$ can be calculated through the following relationship:

$$\overset{\nabla}{\beta}_j = \min\left\{\max\left[\mu \cdot \mathrm{Z}_j + \lambda, \beta_{\min}\right], \beta_{\max}\right\}, \tag{15}$$

where $\mu = 0.45$, $\lambda = 1.3$, $\beta_{\min} = 0.4$, and $\beta_{\max} = 4.0$.

The parameters α_j and β_j are estimated from the input spectrogram and used as initial values to obtain a flexible, effective gain function, which improves better enhancement or separation performance. Therefore, the adaptively estimated parameters α_j and β_j for each source at each iteration guarantee the decrease of the quadrature form of the spectrogram gradients for the effective convergence of the proposed separation algorithm.

III. Experimental Results

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

For the first experiment, 150 full-length song tracks [23] were used (50 songs from the ccMixter database containing many different musical genres, 50 songs from a self-recording studio music database, and 50 songs from the MIR-1 K 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, an accompaniment of six repeating patterns along with a 2 s 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" stage of Algorithm 1 (approximately until convergence).

For the performance measures, performance was evaluated in terms of source-to-interference ratio (SIR) and source-to-distortion ratio (SDR) by blind-source-separation evaluation metrics [24]:

- SDR measures the amount of distortion introduced by the output signal, and is defined as the ratio between the energy of the clean signal and that of the distortion. SDR gives an overall score for separation.
- SIR is defined as the ratio of the target signal power to that of the interference signal, and measures the amount of undesired interference signal still remaining in the separated signal.

The normalized SDR (NSDR) and the normalized SIR (NSIR) for singing voice are defined as

$$\begin{aligned} \mathrm{NSDR}\left(v_{\mathrm{r}}, v, x\right) &= \mathrm{SDR}\left(v_{\mathrm{r}}, v\right) - \mathrm{SDR}\left(x, v\right), \\ \mathrm{NSIR}\left(v_{\mathrm{r}}, v, x\right) &= \mathrm{SIR}\left(v_{\mathrm{r}}, v\right) - \mathrm{SIR}\left(x, v\right), \end{aligned} \tag{16}$$

where v_{r} is the resynthesized 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 NSDR values indicate better separation.

The performance of the proposed WbE algorithm was compared with that of GW, LSA, bSA, and WE, based on KAM.

Tables 1 and 2 present the experimental results of seven 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: Instead of the generalized Wiener filter, the MMSE of the LSA was applied to the SVD-based decomposed power spectrogram.
- SVD-bSA-KAM: Instead of the generalized Wiener filter, bSA was applied to the SVD-based decomposed power spectrogram.
- SVD-WE-KAM: Instead of the generalized Wiener filter, WE error was applied to the SVD-based decomposed power spectrogram.
- SVD-GA-KAM: Instead of the generalized Wiener filter, the MMSE STSA with generalized gamma distribution [25] was applied to the SVD-based decomposed power spectrogram.

Table 1. Comparative performance for music separation.

Separation algorithms	Separation performance for music	
	NSDR	NSIR
STFT-GW-KAM	6.38	9.29
SVD-GW-KAM	6.95	9.81
SVD-LSA-KAM	7.45	10.56
SVD-bSA-KAM	8.78	12.67
SVD-WE-KAM	9.12	12.72
SVD-GA-KAM	6.63	9.49
SVD-WbE-KAM	9.54	12.97

Table 2. Comparative performance for vocal separation.

Separation algorithms	Separation performance for vocal	
	NSDR	NSIR
STFT-GW-KAM	2.45	6.67
SVD-GW-KAM	2.99	7.16
SVD-LSA-KAM	3.45	7.54
SVD-bSA-KAM	3.61	7.68
SVD-WE-KAM	2.86	6.45
SVD-GA-KAM	5.19	9.32
SVD-WbE-KAM	5.17	9.56

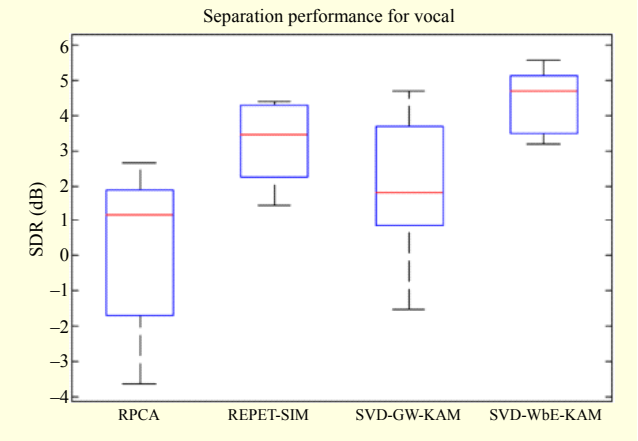

Fig. 1. Boxplot of SDR for vocals.

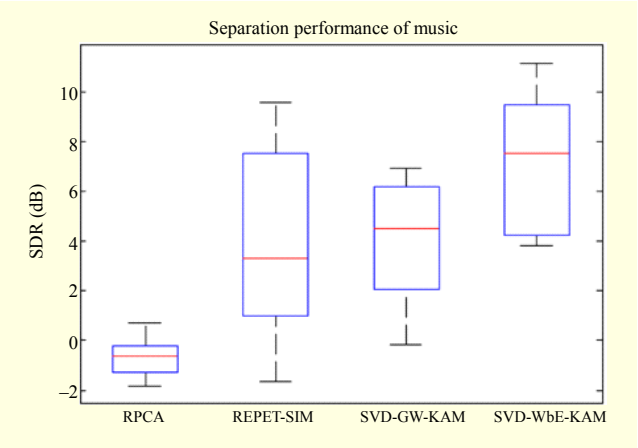

Fig. 2. Boxplot of SDR for music accompaniment.

- SVD-WbE-KAM: Instead of the generalized Wiener filter, WbE was applied to the SVD-based decomposed power spectrogram.

As shown in Table 1, the best separation performance of the music from the mixed music signal is obtained with the proposed method, SVD-WbE-KAM, in terms of NSDR and NSIR. Compared to the other six methods, the basic method, STFT-GW-KAM, attains the worst results.

The experimental results of the seven methods for the separation of vocal components from a monaural signal are depicted in Table 2.

The MMSE STSA with generalized gamma distribution (SVD-GA-KAM) outperforms the other six methods in terms of NSDR, and is slightly lower than the proposed method in terms of NSIR, since speech is well modeled by gamma distribution. However, the proposed method yields better performance than GW, LSA, bSA, and WE.

As shown in Tables 1 and 2, the proposed WbE delivers high performance results in the separation of both music and vocals from the mixed signal. However, the proposed WbE also comes with two disadvantages: (1) it requires a large number of calculations and (2) the perceptually weighted order α_j and the SVD-based factorized spectral amplitude β_j are sometimes sensitive for estimating good models of the source spectrograms. Nevertheless, the adaptively estimated values for α_j and β_j through the iterative KBF process can result in a better source separation than the fixed values for α_j and β_j (although the fixed values are carefully chosen).

For the second performance comparison, the proposed algorithm, SVD-WbE-KAM, was compared with REPET-SIM [26], RPCA [27], and SVD-GW-KAM. To evaluate the separation of background music and singing voice, 40 full-length song tracks [24] were used (20 songs from the ccMixter database containing many different musical genres, and 20 songs from the MIR-1 K database). Figures 1 and 2 show boxplots of the SDR for the vocals and the music accompaniment, respectively.

As can be seen from Figs. 1 and 2, the proposed method shows the highest SDR compared to the other three methods. Sound examples are available at http://imsp.kw.ac.kr/Research.html.

IV. Conclusion

In this paper, a generalized weighted β-order MMSE estimation (WbE) method based on kernel back-fitting (KBF) was proposed and evaluated for the separation of mixed signals into music/voice components. The proposed algorithm enhances the basic KBF algorithm through application of generalized WbE. The proposed method has the following four advantages: (1) in the separation step, generalized WbE of the factorized spectral amplitude is used instead of GW for the KBF procedure to achieve better separation performances; (2) the perceptually weighted order α_j and the SVD-based factorized spectral amplitude β_j are adaptively calculated for effective WbE estimation performance; (3) in the back-fitting step, an SVD-based factorization procedure is applied to the power spectrogram filtered by median filter to achieve efficient compression before processing of the next source; and (4) using a back-fitting threshold, the KBF process can automatically be iteratively performed until convergence. The experimental results show that the proposed method obtained better results compared to other previously reported methods.

In future work, focus will be centered on the optimization of the separation algorithm to allow more effective music/voice separation, along with the kernel characteristics. The method will be applied to music remixing for three-dimensional audio applications.

Acknowledgement

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2015R1D1A1A01059804). 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), and the work reported in this paper was conducted during the sabbatical year of Kwangwoon University in 2013.

References

[1] Z. Rafii and B. Pardo, "REpeating Pattern Extraction Technique (REPET): A Simple Method for Music/Voice Separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 1, Jan. 2013, pp. 73–84.

[2] N.C. Maddage, C. Xu, and Y. Wang, "Singer Identification Based on Vocal and Instrumental Models," *Proc. Int. Conf. Pattern Recogn.*, Cambridge, UK, Aug. 23–26, 2004, pp. 375–378.

[3] M. Ryynanen and A. Klapuri, "Transcription of the Singing Melody in Polyphonic Music," *Int. Conf. Music Inf. Retrieval*, Victoria, Canada, Oct. 8–12, 2006, pp. 222–227.

[4] S. Marchand et al., "DReaM: A Novel System for Joint Source Separation and Multi-track Coding," *133rd AES Conv.*, San Francisco, CA, USA, Oct. 26–29, 2012.

[5] J. Nikunen, T. Virtanen, and M. Vilermo, "Multichannel Audio Upmixing Based on Non-negative Tensor Factorization Representation," *IEEE Workshop Appl. Signal Process. Audio Acoust.*, New Paltz, NY, USA, Oct. 16–19, 2011, pp. 33–36.

[6] U. Simsekli, Y.K. Yilmaz, and A.T. Cemgil, "Score Guided Audio Restoration via Generalized Coupled Tensor Factorisation," *IEEE Int. Conf. Acoust., Speech Signal Process.*, Kyoto, Japan, Mar. 25–30, 2012, pp. 5369–5372.

[7] 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, Oct. 2011, pp. 1180–1191.

[8] 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, Language Process.*, vol. 18, no. 2, Feb. 2010, pp. 310–319.

[9] T. Virtanen, A. Mesaros, and M. Ryynänen, "Combining Pitch-Based Inference and Non-negative Spectrogram Factorization in Separating Vocals from Polyphonic Music," *ISCA Tutorial Res. Workshop Statistical Perceptual Audition*, Brisbane, Australia, Sept. 21, 2008, pp. 17–22.

[10] A. Liutkus et al., "Kernel Additive Models for Source Separation," *IEEE Trans. Signal Process.*, vol. 62, no. 16, Aug. 2014, pp. 4298–4310.

[11] D. Fitzgerald, "Harmonic/Percussive Separation Using Median Filtering," *Int. Conf. Digital Audio Effects*, Graz, Austria, Sept. 6–10, 2010, pp. 1–4.

[12] Z. Rafii and B. Pardo, "A Simple Music/Voice Separation Method Based on the Extraction of the Repeating Musical Structure," *IEEE Int. Conf. Acoust., Speech Signal Process.*, Prague, Czech Republic, May 22–27, 2011, pp. 221–224.

[13] A. Liutkus et al., "Adaptive Filtering for Music/Voice Separation Exploiting the Repeating Musical Structure," *IEEE Int. Conf. Acoust., Speech Signal Process.*, Kyoto, Japan, Mar. 25–30, 2012, pp. 53–56.

[14] Z. Rafii and B. Pardo, "Music/Voice Separation Using the Similarity Matrix," *Int. Conf. Music Inf. Retrieval*, Porto, Portugal, Oct. 8–12, 2012, pp. 583–588.

[15] O. Yilmaz and S. Rickard, "Blind Separation of Speech Mixtures via Time-Frequency Masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, July 2004, pp. 1830–1847.

[16] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32,

no. 6, Dec. 1984, pp. 1109–1121.
[17] E. Plourde and B. Champagne, "Auditory-Based Spectral Amplitude Estimators for Speech Enhancement," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 8, Nov. 2008, pp. 1614–1623.
[18] C.H. You, S.N. Koh, and S. Rahardja, "β-Order MMSE Spectral Amplitude Estimation for Speech Enhancement," *IEEE Trans. Speech, Audio Process.*, vol. 13, no. 4, July. 2005, pp. 475–486.
[19] F. Deng, F. Bao, and C.-C. Bao, "Speech Enhancement Using Generalized β-Order Spectral Amplitude Estimator," *Speech Commun.*, vol. 59, Apr. 2014, pp. 55–68.
[20] C.H. You, S.N. Koh, and S. Rahardja, "Masking-Based β-Order MMSE Speech Enhancement," *Speech Commun.*, vol. 48, no. 1, Jan. 2006, pp. 57–70.
[21] C.H. You, S.N. Koh, and S. Rahardja, "Improved Adaptive β-Order MMSE Speech Enhancement," *APSIPA Ann. Summit Conf.*, Sapporo, Japan, Oct. 4–7, 2009, pp. 797–800.
[22] D.D. Greenwood, "A Cochlear Frequency-Position Function for Several Species-29 Years Later," *J. Acoust. Soc. America*, vol. 87, no. 6, July 1990, pp. 2592–2605.
[23] Multimedia Technology Laboratory homepage, Accessed Nov. 20, 2015. http://imsp.kw.ac.kr/Research.html
[24] E. Vincent, R. Gribonval, and C. Fevotte, "Performance Measurement in Blind Audio Source Separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 4, July 2006, pp. 1462–1469.
[25] R.C. Hendriks et al., "Minimum Mean-Square Error Amplitude Estimators for Speech Enhancement under the Generalized Gamma Distribution," *Int. Workshop Acoust. Echo Noise Contr.*, Paris, France, Sept. 12–14, 2006, pp. 1–4.
[26] Z. Rafii, A. Liutkus, and B. Pardo, "REPET for Background/Foreground Separation in Audio," in *Blind Source Separation: Advances in Theory, Algorithms and Appl.*, Berlin, Germany: Springer, 2014, pp. 395–411.
[27] P.S. Huang et al., "Singing-Voice Separation from Monaural Recordings Using Robust Principal Component Analysis," *IEEE Int. Conf. Acoust., Speech Signal Process.*, Kyoto, Japan, Mar. 25–30, 2012, pp. 57–60.

Hyoung-Gook Kim received a Dr.-Ing. degree in electrical engineering and computer science from the Technical University of Berlin, Germany, in 2002. From 1998 to 2005, he worked on mobile service robots at Daimler Benz, and speech recognition at Siemens, Berlin, Germany. From 2005 to 2007, he was a project leader at the Samsung Advanced Institute of Technology, Suwon, Rep. of Korea. Since 2007, he has been a professor with the Department of Electronics Convergence Engineering, Kwangwoon University, Seoul, Rep. of Korea. His research interests include audio signal processing, audio-visual content indexing and retrieval, and speech enhancement.

Jin Young Kim received his PhD degree in electronic engineering from Seoul National University, Rep. of Korea. He worked on speech synthesis at Korea Telecom, from 1993 to 1994. Since 1995, he has been a professor with the Department of Electronics and Computer Engineering, Chonnam National University, Gwangju, Rep. of Korea. His research interests include speech synthesis, speech and speaker recognition, and audio-visual speech processing.