Vocal Separation from Monaural Music Using Adaptive Auditory Filtering Based on Kernel Back-fitting

Jun-Yong Lee, Hye-Seung Cho, Hyoung-Gook Kim

Kwangwoon University, Seoul, Rep. of Korea

{jasonlee88; hye_seung401; hkim}@kw.ac.kr

Abstract

Recently, kernel additive modeling with generalized spatial Wiener filtering (GW) was presented for music/voice separation. In this paper, an adaptive auditory filtering, called generalized weighted β-order MMSE estimation (WbE), is applied to the basic iterative kernel back-fitting algorithm for improving the separation performance of monaural music signal into music/voice components. In the proposed method, the perceptually weighting factor α and the singular value decomposition (SVD)-based factorized spectral amplitude exponent β for each kernel component are adaptively calculated for effective WbE-based auditory filtering performance. Experimental results show that the proposed method achieves better separation performance than GW and the existing Bayesian estimators.

Index Terms: kernel back-fitting, separation, generalized weighted β-order MMSE estimation, singular value decomposition.

1. Introduction

Vocal separation of mixed music signal is a fundamental preprocessing step in various cases of applications in our real lives, such as automatic karaoke [1], instrument/vocalist identification [2], music/voice transcription, music remixing [3] and audio restoration.

Recently, a relatively promising approach using kernel additive modeling (KAM) was proposed [4], wherein the spectrogram of each source is modeled only locally. 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 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 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 [5] outperform the Wiener filtering, spectral-subtraction, and subspace approaches. In addition, among the Bayesian STSA estimation methods, weighted β-order MMSE estimation [5] achieved better enhancement performance than the existing Bayesian estimators, such as those based on the MMSE of the short-time spectral amplitude [5], the MMSE of the logarithm of the STSA (LSA) [5], the Weighted Euclidien (WE) error [5], and β-order MMSE STSA (bSA) [5], in terms of both objective and subjective measures. The weighted β-order MMSE estimation [5], [6] combines the power law of the β-SA and the weighting factor of the WE. In this paper, an advanced music/voice separation method is proposed, in which weighted β-order MMSE estimation and kernel back-fitting are combined for improvement of the separation performance.

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 weighted β-order MMSE estimation (WbE), 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.

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 $X(\omega,t)$ using the short-time discrete Fourier transform (STFT), as shown:

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

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 $X(\omega,t)$, complex spectrogram $O_j(\omega,t)$ for each objective sources is estimated by generalized weighted β-order MMSE estimation.

Each current estimated spectrogram is compared with each previous estimated complex spectrogram. If the difference

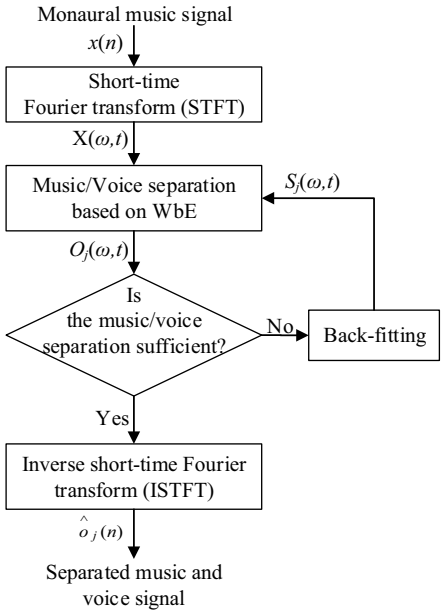

Figure 1: *Overall flow chart of proposed music/voice separation algorithm.*

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.

While the back-fitting processes, the power spectrogram of the estimated spectrogram is filtered by a simple 2-D 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 sources and to separate them in order to perform the estimation. 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 2D median filter is applied to the power spectrogram of the complex spectrogram $V_j(\omega,t)$ with source-specific binary kernels, vocal, harmonic, and percussive. The three kernels [4] 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 vocal kernel is chosen. 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.

(Step 3) Kernel back-fitting using Wiener filtering or the generalized weighted 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 $M_j(\omega,t)$:

$$S_j(\omega,t) = D_j \Sigma_j C_j = \mathrm{SVD}\left[M_j(\omega,t)\right] \tag{5}$$

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 weighted β-order MMSE estimation

In the separation step, generalized weighted β-order MMSE estimation (WbE) of the factorized spectral amplitude is used instead of GW for the kernel back-fitting procedure to achieve better music/voice separation performances. For effective WbE estimation performance, the perceptually weighted order $\alpha_j(\omega,t)$ and the singular value decomposition (SVD)-based factorized spectral amplitude $\beta_j(\omega,t)$ are adaptively calculated.

2.2.1. Weighted β-order MMSE Estimation

The WbE 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 $\alpha_j(\omega,t)$ and $\beta_j(\omega,t)$, and gain function. Figure 2 shows the weighted β-order MMSE estimation.

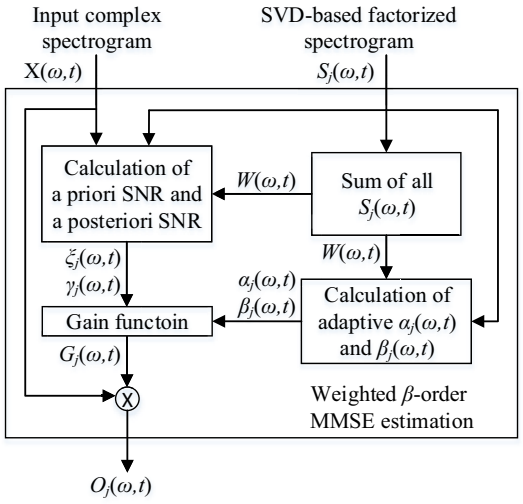

Figure 2: Overall flow chart of the weighted β-order MMSE estimation.

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

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

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

$$\begin{aligned}\xi_j(\omega,t)&=\frac{S_j(\omega,t)}{W(\omega,t)-S_j(\omega,t)};\\ \gamma_j(\omega,t)&=\frac{\left|\mathrm{X}(\omega,t)\right|^2}{W(\omega,t)-S_j(\omega,t)};\\ \chi_j(\omega,t)&=\frac{\xi_j(\omega,t)}{1+\xi_j(\omega,t)}\gamma_j(\omega,t);\end{aligned} \tag{7}$$

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

The gain function for the WbE is given by:

$$G_j(\omega,t)=\frac{\sqrt{\chi_j(\omega,t)}}{\gamma_j(\omega,t)} \cdot\left[\frac{\Gamma\left(\frac{\beta_j(\omega,t)-2\alpha_j(\omega,t)}{2}+1\right)}{\Gamma\left(-\alpha_j(\omega,t)+1\right)}\cdot\frac{\Phi\left(-\frac{\beta_j(\omega,t)-2\alpha_j(\omega,t)}{2},1;-\chi_j(\omega,t)\right)}{\Phi\left(\alpha_j(\omega,t),1;-\chi_j(\omega,t)\right)}\right]^{\frac{1}{\beta_j(\omega,t)}} \tag{8}$$

where $\alpha_j(\omega,t)$ and $\beta_j(\omega,t)$ the parameters based on the human auditory system. And $\Gamma(\bullet)$ is the gamma function, $\Phi(\bullet)$ is the confluent hypergeometric function.

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

$$O_j(\omega,t)=G_j(\omega,t)\cdot\mathrm{X}(\omega,t) \tag{9}$$

2.2.2. Calculation of adaptive $\alpha_j(\omega,t)$ and $\beta_j(\omega,t)$

Since the perceptually weighted order $\alpha_j(\omega,t)$ and the spectral amplitude order $\beta_j(\omega,t)$ 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 $\alpha_j(\omega,t)$ and $\beta_j(\omega,t)$ can result in better enhancement or separation performance. The processing of calculation of $\alpha_j(\omega,t)$ and $\beta_j(\omega,t)$ are as follow:

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

$$Z_j(t)=10\log_{10}\frac{\sum_{\omega=0}^{\Omega-1}\left|W(\omega,t)-\sqrt{W(\omega,t)-S_j(\omega,t)}\right|^2}{\sum_{\omega=0}^{\Omega-1}\left(W(\omega,t)-S_j(\omega,t)\right)} \tag{10}$$

Combining the sub-band SNR $Z_j(t)$ with the masking threshold T, the parameter $\hat{\alpha}_j(t)$ can be described as a function of the two variables in polynomial form. The parameter $\hat{\alpha}_j(t)$ is approximated as follows:

$$\begin{aligned}\hat{\alpha}_j(t)&=\mathrm{F}\left(Z_j(t),\mathrm{T}_j\right)\cong\sum_{p=0}^{\infty}\sum_{q=0}^{\infty}c_{pq}Z_j^p(t)\mathrm{T}_j^q\\ &\cong e_0+e_1Z_j(t)+e_2\mathrm{T}_j+e_3Z_j(t)\mathrm{T}_j\end{aligned} \tag{11}$$

where c_{pq} 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 [6].

According to the sub-band SNR and the auditory masking effect, an adaptive calculation method of parameter $\alpha_j(\omega,t)$ is given as follows:

$$\alpha_j(\omega,t)=\alpha_{low}+a\cdot\frac{\left(f_\omega-2000\right)\left(\alpha_{high}-\alpha_{low}\right)}{F_s/2-2000}+\left(1-a\right)\cdot\hat{\alpha}_j(t) \tag{12}$$

where $\alpha_{low} = 0.25$ and $\alpha_{high} = 0.94$ are used for the trade-off between target source enhancement and other source reduction, a ($0 < a < 1$) is a smoothing parameter, and f_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.

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

$$\hat{\beta}_j(\omega)=\beta_{high}-\frac{\frac{1}{\eta}\log_{10}\left(\frac{f_\omega}{A}+l\right)}{\frac{1}{\eta}\log_{10}\left(\frac{F_s}{2A}+l\right)}\left(\beta_{high}-\beta_{low}\right)\ \text{for}\ 1\le j\le J \tag{13}$$

where $\beta_{high} = 0.2$ and $\beta_{low} = 1$ denote the low-frequency and high-frequency of the compression rate, $\eta = 0.06$ mm, $l = 1$, and $A = 165.4$ Hz are the parameters set in paper [7].

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

$$\check{\beta}_j(t)=\min\left\{\max\left[\mu\cdot Z_j(t)+\lambda,\beta_{\min}\right],\beta_{\max}\right\} \tag{14}$$

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

$$\beta_j(\omega,t)=b\cdot\hat{\beta}_j(\omega)+\left(1-b\right)\cdot\check{\beta}_j(t) \tag{15}$$

where $\mu = 0.45$, $\lambda = 1.3$, $\beta_{min} = 0.4$, and $\beta_{max} = 4.0$, b ($0 < b < 1$) is a smoothing parameter.

3. Experimental Results

In this section, the performance of the proposed WbE-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 20ms. The frame length was set to 90ms, 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 [8]. NSDR and NSIR for singing voice are defined as:

$$\begin{aligned} \mathrm{NSDR}(v_r, v, x) &= \mathrm{SDR}(v_r, v) - \mathrm{SDR}(x, v) \\ \mathrm{NSIR}(v_r, v, x) &= \mathrm{SIR}(v_r, v) - \mathrm{SIR}(x, v) \end{aligned} \quad (16)$$

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 values indicate better separation.

The performance of the proposed WbE algorithm was compared with those of GW, LSA, β-SA, and WE, based on KAM. Tables 1 presents the experimental results of comparative performance for music/voice separation of the six 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.
- SVD-WE-KAM: Weighted euclidien (WE) error was applied to the SVD-based decomposed power spectrogram.
- SVD-WbE-KAM: Weighted β-order MMSE estimation was applied to the SVD-based decomposed power spectrogram.

Table 1. *Comparative performance for music/voice separation*

Ratio	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
SVD-WE-KAM	8.52	12.32	2.43	6.23
SVD-WbE-KAM	8.94	12.54	4.75	8.54

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 five methods, the basic method, STFT-GW-KAM, attains the worst results. And the proposed WbE delivers high performance result in the separation of vocal components.

4. Conclusions

In this paper, we proposed a generalized weighted β-order MMSE estimation method based on kernel back-fitting for music/voice separation. The proposed algorithm enhances the basic kernel back-fitting algorithm through application of generalized weighted β-order MMSE 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. Acknowledgements

This research was supported by the MSIP(Ministry of Science, ICT & Future Planning), Korea, under the ITRC(Information Technology Research Center) support program supervised by the NIPA(National IT Industry Promotion Agency) (NIPA-2014-H0301-14-1019)

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, 2013.

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

[3] S. Marchand, R. Badeau, C. Baras, L. Daudet, D. Fourer, L. Girin, S. Gorlow, A. Liutkus, J. Pinel, G. Richard, N. Sturmel, and S. Zhang, "DReaM: A Novel System for Joint Source Separation and Multi-Track Coding," *133rd Audio Engineering Society Convention*, 2012.

[4] A. Liutkus, D. Fitzgerald, Z. Raffi, B. Pardo, and L. Daudet, "Kernel Additive Models for Source Separation," *IEEE Transactions on Signal Processing*, vol. 62, no. 16, pp. 4298–4310, 2014.

[5] 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.

[6] F. Deng, F. Bao, and C.-C. Bao, "Speech Enhancement Using Generalized β-Order Spectral Amplitude Estimator," *Speech Communication*, vol. 59, pp. 55–68, 2014.

[7] 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.

[8] E. Vincent, R. Gribonval, and C. F´evotte, "Performance Measurement in Blind Audio Source Separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.