Towards Solving the Bottleneck of Pitch-based Singing Voice Separation

Bilei Zhu[1], Wei Li[1,2*], Linwei Li[1]

[1] School of Computer Science and Technology, Fudan University, Shanghai, China
[2] Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, Shanghai, China
bileizhu@gmail.com, weili-fudan@fudan.edu.cn, lwli11@fudan.edu.cn

ABSTRACT

Singing voice separation from accompaniment in monaural music recordings is a crucial technique in music information retrieval. A majority of existing algorithms are based on singing pitch detection, and take the detected pitch as the cue to identify and separate the harmonic structure of the singing voice. However, as a key yet undependable premise, vocal pitch detection makes the separation performance of these algorithms rather limited. To overcome the inherent weakness of pitch-based inference algorithms, two novel methods based on non-negative matrix factorization (NMF) are devised in this paper. The first one combines NMF with the distribution regularities of vocals under different time frequency resolutions, so that many vocal unrelated portions are eliminated and the singing voice is hence enhanced. In consequence, the accuracy of vocal pitch detection is significantly improved. The second method applies NMF to decompose the spectrogram into non-overlapping and indivisible segments, which can be used as another cue besides the pitch to help identify the vocal harmonic structure. The two proposed methods are integrated into the framework of pitch-based inference. Extensive testing on the MIR-1K public dataset shows that both of them are rather effective, and the overall performances outperform other state-of-the-art singing separation algorithms.

Categories and Subject Descriptors

H.5.5 [Information Systems]: Information Interfaces and Presentation-*Sound and Music Computing*

General Terms

Algorithms; Experimentation

Keywords

Singing voice separation; pitch-based inference; singing pitch detection; non-negative matrix factorization (NMF)

MM'15, October 26-30, 2015, Brisbane, Australia
© 2015 ACM. ISBN 978-1-4503-3459-4/15/10…$15.00
DOI: http://dx.doi.org/10.1145/2733373.2806257

1. INTRODUCTION

Most music pieces people listen to in daily life are mixtures of singing voice and accompaniment. Compared with the instrumental accompaniment, the singing voice carries important information such as the main melody and lyrics, and is usually more impressive. Many applications in music information retrieval (MIR), e.g., melody extraction [1], singer identification [2], and automatic lyrics recognition [3], are mainly related to the singing voice, and in this case the musical accompaniment is treated as noise. Therefore, singing voice separation has emerged as a crucial technique in recent years.

Music recordings can be either monaural or stereo, and in the former case singing voice separation is generally considered more challenging because of the absence of spatial information. To tackle this problem, a few algorithms have been proposed in recent years and most of them are within the framework of pitch-based inference. It is known that singing voice is primarily comprised of voiced sounds (about 90%) [4], which are roughly harmonic, with frequencies of concurrent overtones being approximately integer multiples of the fundamental frequency. Pitch-based inference algorithms utilize the harmonic structure of the singing voice, and generally first extract the singing pitch from sound mixtures as the cue for subsequent separation.

Unfortunately, pitch-based inference algorithms have several intrinsic limitations. First of all, they highly rely on the technique of singing pitch detection from polyphonic music signals, which, however, remains an open problem and has not been maturely solved so far [5, 6, 7, 8]. As a premise, if the detected pitch is inaccurate, the harmonic structure of the singing voice cannot be correctly identified, so that the quality of vocal separation will be low. Next, singing voice is almost always accompanied by instrumental sounds, which in most cases are harmonic, broadband and strongly correlated with the vocals [9]. Spectral contents of the singing voice and accompaniment overlap in many time-frequency positions, which bring about great difficulties to vocal separation since it is nearly impossible to discriminate overlapping spectral contents using pitch as the only cue. Finally, although the majority of singing voice is voiced sounds, a small part of unvoiced sounds do exist. Having no underlying periodicity, unvoiced sounds cannot be characterized by pitch and further effectively separated from accompaniment using most of existing pitch-based inference algorithms [10].

This paper aims to solve the first two limitations of pitch-based inference to a certain degree, using two methods based on non-negative matrix factorization (NMF). In the first method, NMF is used to decompose the spectrogram of the input sound mixture, with long and short frames successively. Owing to the property of being fluctuant in frequency and short in time, singing voice exhibits certain continuity in the frequency axis and time axis when long and short frames are adopted, respectively. In consequence, NMF components reflecting the above phenomenon can be picked out and resynthesized to obtain an enhanced vocal signal, which will benefit for the subsequent singing pitch detection. In the second method, NMF is applied to the input sound mixture to obtain a set of non-overlapping time-frequency segments, each of which approximately originates from a single sound source. By virtue of the indivisible property, these segments provide additional information besides the periodicity information provided by the singing pitch, which is often inaccurate, for the identification of the vocal harmonic structure. The two proposed methods are integrated into the framework of pitch-based inference. Experiments carried out on the MIR-1K public dataset show the effectiveness of both methods, and the overall performances outperform two state-of-the-art methods.

The remainder of the paper is organized as follows. Section 2 reviews some related works on monaural singing voice separation. Section 3 introduces NMF briefly. Section 4 describes our singing voice separation algorithm. Experiment results and performance analysis are presented in section 5. Finally, section 6 concludes this paper.

2. RELATED WORK

In this section, related works on monaural singing voice separation are briefly reviewed. To our knowledge, the first attempt to comprehensively solve the problem of singing voice separation was made by Li and Wang [9]. Their algorithm is a typical method of pitch-based inference under the framework of computational auditory scene analysis (CASA) [11], where the predominant pitch is first extracted from the input sound mixture, and then applied as the cue to label the time-frequency units (T-F units). T-F units are obtained by decomposing the sound mixture via an auditory filterbank, and labeled as singing dominant if their local pitch values match that of the singing voice. Based on the labeled T-F units, the algorithm forms a binary mask and applies it to resynthesize the singing voice. This algorithm is later extended by Hsu and Jang [10], by combining an unvoiced sounds separation method and a spectral subtraction approach.

Obviously, the accuracy of singing pitch detection is critical for the performance of pitch-based singing separation. And meanwhile, it is also evident that enhanced singing voice is beneficial for singing pitch detection [12]. To utilize this interdependency, Hsu et al. proposed a tandem algorithm, which estimates the singing pitch and separates the singing voice jointly and iteratively [13]. Specifically, the algorithm first has a rough estimation for the pitch of the singing voice and then applies it to separate the singing voice by considering harmonicity and temporal continuity. The separated singing voice and estimated pitch are then iteratively fed back to each other for further refinement.

In contrast with the above three methods that represent the singing voice with a group of singing dominant T-F units, Ryynanen et al. proposed an alternative scheme which models the singing voice as a sum of time varying sinusoids [14]. Using this sinusoidal model, the task of singing voice separation resolves into first estimating the model parameters, i.e., frequency, amplitude and phase of each sinusoid per frame, and then interpolating them over time. To estimate these parameters, the singing pitch is first extracted, with its integer multiples as the constraints imposed on the frequencies of sinusoids. The amplitude and phase of each sinusoid are then estimated from the normalized cross-correlation between the analysis frame and a complex exponential having the frequency of the sinusoid. As above, singing pitch detection is the foundation of separation performance.

Virtanen et al. [15] first employed a pitch-based approach to extract the harmonic structure of the singing voice, and then applied NMF on the residual spectrogram. The factorization results in an estimation of the musical accompaniment, which is finally subtracted from the original mixture to achieve a better separation of the singing voice. On the surface, this method resembles our algorithm since they both integrate NMF into a pitch-based inference singing separation framework. However, the purposes of NMF are completely irrelevant. In other words, in [15] NMF is used to estimate the accompaniment, while in our algorithm it is used for singing enhancement and more precise identification of the vocal harmonic structure, respectively.

In addition to the above pitch-based inference separation algorithms, several algorithms outside this framework have also been proposed for monaural singing voice separation. Some of these algorithms are in accordance with the line of spectrogram factorization. In [16], Vembu and Baumann proposed using NMF to decompose the spectrogram of the input sound mixture into a set of components, which are then clustered into different sound sources using an unsupervised clustering approach based on spectral features. In [17], Chanrungutai and Ratanamahatana also applied NMF for spectrogram factorization, while the components are assigned to different sound sources by investigating the rhythmic and continuous events. Other algorithms for monaural singing separation include those based on the extraction of the repeating musical structure [18, 19], and those using robust principal component analysis [20, 21], etc.

3. NON-NEGATIVE MATRIX FACTORIZATION

Non-negative matrix factorization (NMF) [22] is an unsupervised technique employed for linear representation of non-negative data. Given a non-negative matrix $\mathbf{X}$ of dimensions $K \times T$ and a positive integer J, NMF finds an approximate factorization

$$\mathbf{X} \approx \mathbf{BG} \qquad (1)$$

where $\mathbf{B}$ and $\mathbf{G}$ are non-negative matrices of dimensions $K \times J$ and $J \times T$ respectively. In contrast to many other linear representations such as independent component analysis (ICA) [23] and principal component analysis (PCA) [24], NMF imposes the non-negativity constraint, which leads to a parts based representation because the constraint allows only additive, not subtractive combinations of the original data.

Recently, NMF and its extensions have been successfully used in the field of audio analysis, for various problems such as polyphonic music transcription [25], audio-to-score alignment [26], and musical instrument classification [27]. In particular, NMF based algorithms have produced attractive results in sound source separation [28].

When using NMF for sound source separation, the observation matrix $\mathbf{X}$ is typically a phase-invariant time-frequency representation (e.g., magnitude spectrogram or power spectrogram), where K is the number of frequency channels and T is the number of time frames. The model matrices, $\mathbf{B}$ and $\mathbf{G}$ are basis matrix and gain matrix respectively, where the columns of $\mathbf{B}$ are basis functions and the rows of $\mathbf{G}$ are their gains in each frame.

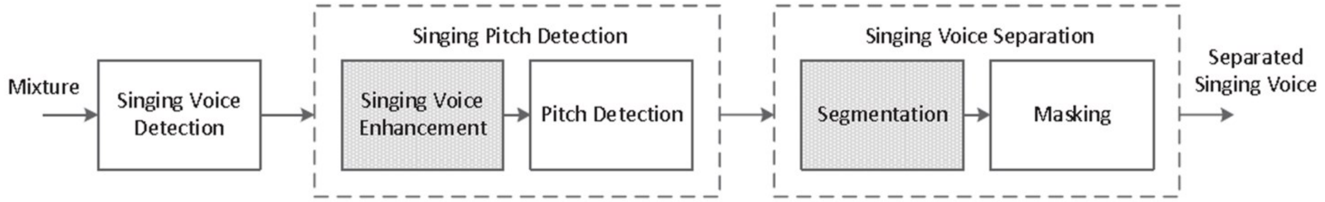

Figure 1: Overview of our singing voice separation algorithm.

The factorization, according to Eq. (1), is usually sought by minimizing a chosen cost function between $\mathbf{X}$ and $\mathbf{BG}$ while restricting their elements to nonnegative values. According to [28], the best performed cost function among several commonly used ones is the Kullback- Leibler divergence of Eq. (2)

$$\mathbf{D}(\mathbf{X}||\mathbf{BG}) = \sum_{k=1}^{K}\sum_{t=1}^{T}[\mathbf{X}]_{k,t}\boldsymbol{log}\frac{[\mathbf{X}]_{k,t}}{[\mathbf{BG}]_{k,t}} - [\mathbf{X}]_{k,t} + [\mathbf{BG}]_{k,t} \quad (\mathbf{2})$$

Compared with other cost functions such as the Euclidean distance, the divergence is more sensitive to low energy observations, making it a better approximation of human auditory perception. To solve the minimization problem with respect to Eq. (2), Lee and Seung proposed a simple method [29], where $\mathbf{B}$ and $\mathbf{G}$ are initialized with random positive values and then alternatively updated with the following multiplicative update rules

$$\mathbf{B} \leftarrow \mathbf{B}\otimes\frac{\frac{\mathbf{X}}{\mathbf{BG}}\mathbf{G}^{\boldsymbol{T}}}{\mathbf{1}\mathbf{G}^{\boldsymbol{T}}} \quad (\mathbf{3})$$

$$\mathbf{G} \leftarrow \mathbf{G}\otimes\frac{\mathbf{B}^{\boldsymbol{T}}\frac{\mathbf{X}}{\mathbf{BG}}}{\mathbf{B}^{\boldsymbol{T}}\mathbf{1}} \quad (\mathbf{4})$$

where $\mathbf{A}\otimes\mathbf{B}$ and $\frac{A}{B}$ are the element-wise multiplication and division of matrices **A** and **B**, respectively, and **1** is an all-one matrix of the same size as **X**. According to the proofs given in [29], the divergence Eq. (2) is non-increasing under the update rules Eq. (3-4). As a result of NMF, the observation matrix of the input sound mixture is decomposed into a set of repetitive components, each of which represents parts of a single sound source. A component here refers to a basis function (a column of **B**) and its time-varying gain (the corresponding row of **G**), and there are thus *J* components in total. Generally, each sound source is modeled as a sum of one or more components, and therefore the separation is done by grouping these components to sound sources.

4. ALGORITHM DESCRIPTION

As shown in Fig. 1, the overall pitch-based inference algorithm consists of three modules, i.e., singing voice detection, singing pitch detection, and singing voice separation. A sound mixture is first input into the singing voice detection module to locate the vocal portions by supervised classification. Then, such portions are used to calculate the singing pitch. To improve the precision of pitch estimation, a novel NMF-based singing enhancement method is designed and adopted as a preprocessing step. Finally, singing voice separation is done by decomposing the input mixture into T-F units and grouping singing dominant units to form a binary mask for resynthesizing. To correct the errors in identifying singing dominant units with only pitch as the cue, another NMF-based method, which decomposes the original mixture into time-frequency segments related to specific sound source, is designed to help improve the identification accuracy.

4.1 Singing Voice Detection

As a prerequisite of the following vocal separation, singing voice detection is first performed to partition the input sound mixture into vocal and nonvocal portions. To achieve this goal, we follow the typical solution of supervised classification. In training, each song in the training set is first divided into short-time overlapping frames, then Mel-frequency cepstral coefficients (MFCCs) are extracted from each frame. All pairs of MFCCs and pre-labeled tags (vocal/nonvocal) are assembled together frame by frame, and fed into the hidden Markov model (HMM) classifier to train its parameters. In testing, given a new song outside the training dataset, the MFCCs of each unlabeled frame are extracted and input to the classifier, with a decision made on whether vocal is present or not.

4.2 Singing Pitch Detection

The vocal portions obtained from the above module are normally mixed with music accompaniment, which brings great disturbance to singing pitch detection. Therefore, to weaken the negative effects of accompaniment, we first design a novel method to enhance the singing voice based on its unique time-frequency characteristic, and then integrate an existing method described in [30] to perform singing pitch detection.

The proposed method for singing voice enhancement is inspired by the observation that singing voice is a temporally variable signal with distinctive characteristics, i.e., fluctuation and shortness [12]. Due to the intrinsic features, singing voice exhibits distinct temporal and spectral properties in spectrograms calculated with different frame lengths. To be specific, in the case of long (short) frame length, the frequency (time) resolution is relatively high, thus the vocal signals appear more smoothly and continuously in the frequency (time) axis, as illustrated in Fig. 2.

Motivated by the above observation, a two-stage NMF based procedure is naturally conceived to screen out the vocal portions from spectrograms of different time-frequency resolutions. As shown in Fig. 3, a magnitude spectrogram is first constructed from the input song using long-frame (e.g., 256 ms) short-time Fourier transform (STFT), and then decomposed into a set of NMF components. Generally speaking, each component approximately originates from a single sound source. As a result, the problem of how to select and retain vocal portions that are relatively continuous along the frequency axis is equivalently converted to

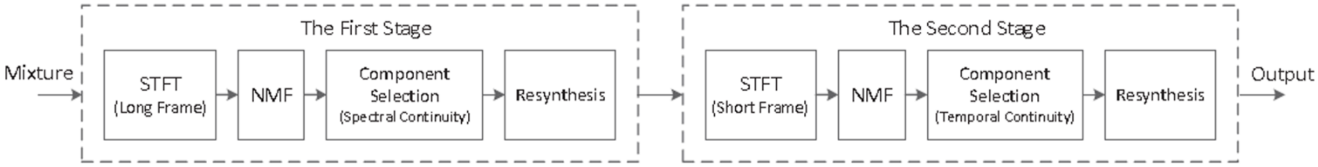

Figure 3: Overview of the proposed singing voice enhancement method.

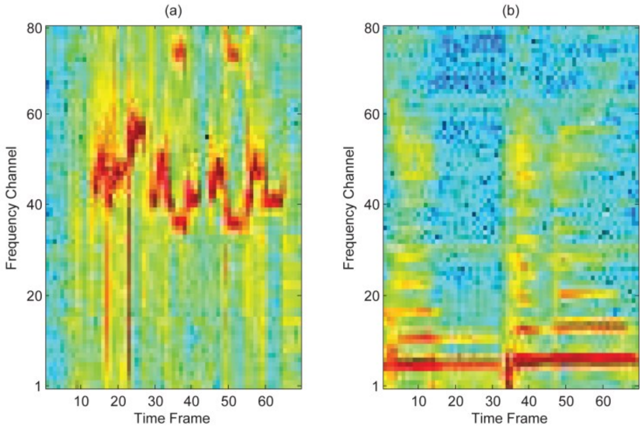

Figure 2: Magnitude spectrograms of a singing voice signal. (a) A part of a long-frame spectrogram (frame length = 256 ms, frame overlap = 128 ms). (b) A part of a short-frame spectrogram (frame length =32 ms, frame overlap = 16 ms).

how to pick out the NMF components which are continuous in the same direction. After eliminating unwanted components, a new spectrogram is reconstructed from the retained ones, and used to resynthesize the preliminarily enhanced singing signal via inverse STFT. Next, the resynthesized signal is input into the second stage which is completely similar to the above procedures except that the short-frame (e.g., 32 ms) STFT is adopted to pick out vocal signals that are relatively continuous along the time axis. After the two-stage of filtering, accompaniment is significantly attenuated, namely the singing content is greatly enhanced which will be beneficial for the subsequent singing pitch calculation.

As a key technical problem, the selection of NMF components is solved by applying a spectral or temporal continuity thresholding method, formalized as below. Given the spectrogram **X** of dimensions $K \times T$, the number of components J, and the factorization **X** $\approx$ **BG**, spectral continuity of each component is measured by assigning a cost to large changes between adjacent elements in the corresponding column of the basis matrix **B**. Specifically, for component $\boldsymbol{X}^j$, j = 1,…, J being the component index, the spectral continuity measure $c_s(\boldsymbol{X}^j)$ is defined as

$$\boldsymbol{c_s(X^j) = \frac{1}{\sigma_j^2}\sum_{k=2}^{K}([B]_{k,j} - [B]_{k-1,j})^2},\qquad (5)$$

where $\sigma_j = \sqrt{\frac{1}{K}\sum_{k=1}^{K}[B]_{k,j}^2}$ is a normalization factor. And if $c_s(\boldsymbol{X}^j)$ satisfies

$$\boldsymbol{c_s(X^j) \leq \theta_s},\qquad (6)$$

where θ_s is a threshold, the component is considered to be continuous in the frequency axis.

As for temporal continuity of each component, it is measured using the criteria described in [28], which assigns a cost to large changes between adjacent elements in the corresponding row of the gain matrix **G**. Specifically, for component $\boldsymbol{X}^j, j = 1, \ldots,$ J being the component index, the temporal continuity measure $c_t(\boldsymbol{X}^j)$ is written as

$$\boldsymbol{c_t(X^j) = \frac{1}{\epsilon_j^2}\sum_{t=2}^{T}([G]_{j,t} - [G]_{j,t-1})^2},\qquad (7)$$

where $\epsilon_j = \sqrt{\frac{1}{T}\sum_{t=1}^{T}[\boldsymbol{G}]_{j,t}^2}$ is a normalization factor. And if $c_t(\boldsymbol{X}^j)$ satisfies

$$\boldsymbol{c_t(X^j) \leq \theta_t},\qquad (8)$$

where θ_t is a threshold, the component is considered to be continuous in the time axis.

4.3 Singing Voice Separation

In this stage, singing voice is to be separated from the input sound mixture based on the detected singing pitch. First, the input song is passed through a 128-channel Gammatone filterbank, whose center frequencies are equally distributed on the equivalent rectangular bandwidth (ERB) scale between 80 Hz and 5 kHz. The output signal of each filter is then divided into short-time overlapping frames. As a result, the input sound mixture is decomposed into a collection of T-F units. A T-F unit here is denoted as u_{cm}, where c and m are the indexes of filter channel and time frame, respectively. Given these T-F units, the singing voice separation is done by first estimating the ideal binary time frequency mask and then resynthesizing.

An ideal binary mask is a binary matrix, where 1 means that the energy of the singing voice is stronger than that of the accompaniment within the corresponding T-F unit and 0 indicates weaker [31]. To estimate the mask, a natural choice is to perform unit labeling, i.e., to label each individual T-F unit with 1 if it is identified as singing dominant or 0 if otherwise. Given the detected singing pitch, this is usually done by matching the local periodicity of each T-F unit, obtained by finding the maximum of the autocorrelation within plausible pitch range, with the pitch period of the present frame. If the match occurs, the T-F unit is identified as singing dominant [9, 10]. However, due to the inaccuracy of detected singing pitch, the pitch-based unit labeling often results in errors, especially false negatives for labeling singing-dominant T-F units.

To deal with the errors occurred in pitch-based unit labeling, we devise a NMF-based rectification method specified as follows.

1. Construct an energy matrix $\mathbf{E}$ of T-F units, the matrix element $[\mathbf{E}]_{c,m}$ is calculated as

$$[\mathbf{E}]_{c,m} = \sum_{n=1}^{N} u_{cm}^{2}(n), \tag{9}$$

   where $u_{c,m}(n)$ is the nth sample in $u_{c,m}$, N is frame length in number of samples. Obviously, $\mathbf{E}$ is a nonnegative matrix of dimensions $C \times M$, where C = 128 is the number of filter channels, and M is the number of time frames.

2. Next, perform NMF on the obtained matrix $\mathbf{E}$ to decompose it into a set of components. Given the factorization $\mathbf{E} \approx \mathbf{WH}$ and the number of components R, a component here is denoted as $\boldsymbol{E}^r$, r = 1, R, and represented as a time-frequency matrix. Specifically, the matrix element at position (c, m) is calculated as

$$[\boldsymbol{E}^r]_{c,m} = [\mathbf{W}]_{c,r}[\mathbf{H}]_{r,m} \tag{10}$$

   Generally, each of the obtained components originates from a single sound source.

3. Finally, a segment is generated from each component obtained above. Specifically, for a given component $\boldsymbol{E}^r$, its time-frequency representation is compared with those of other components, with the matrix elements satisfying Eq. (11) selected.

$$[\boldsymbol{E}^r]_{c,m} = \max_{i=1\sim R} [\boldsymbol{E}^i]_{c,m} \tag{11}$$

   In general, each selected element $[\boldsymbol{E}^r]_{c,m}$ corresponds to a T-F unit $u_{c,m}$, and all these T-F units form a segment S^r corresponding to $\boldsymbol{E}^r$.

   Fig. 4 gives an illustration of how to generate segments from NMF components. For a given component, red elements in its time-frequency representation are those satisfying Eq. (11), meaning that they are larger than all the green elements in the same positions of other time-frequency representations. Typically, each red element corresponds to a T-F unit with the same time frequency index with it, and all these units form the segment corresponding to the given component.

As a result of the above procedure, the input sound mixture is decomposed into a set of time-frequency segments, each of which is indivisible, with primary energy originating from a single sound source. With the constraint of Eq. (11), these segments are non-overlapping, i.e., a T-F unit only belongs to a single segment. In other words, segments are disjoint clusters of T-F units, and all the T-F units included in a segment are dominated by the same sound source. Given this property, a T-F unit can be labeled based on not only the periodicity information provided by the singing pitch, but also the origination of the segments that it belongs to.

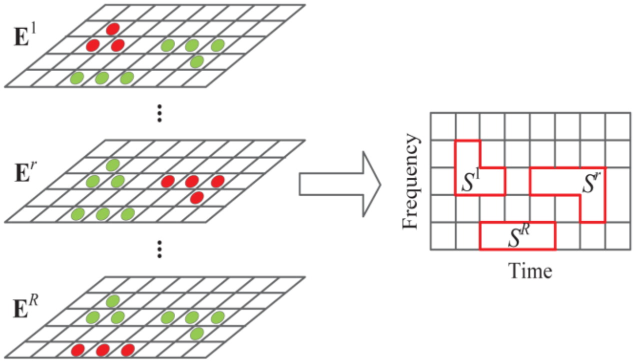

Figure 4: Illustration of segment generation.

Given the time-frequency segments, we now describe how to estimate the ideal binary mask using pitch and segment as two complementary cues. First, let $\boldsymbol{M}^0$ be the masking result for singing voice of conventional pitch-based unit labeling method. Then, the segment cue is considered to get additional masking information. To be specific, for each segment, if more than a certain percentage (20% for example) of its units have been identified as singing dominant, the whole segment is considered to be originated from vocals. All these segments are assembled together and form another masking matrix, denoted as $\boldsymbol{M}^1$. The final estimation result $\mathbf{M}$ is the combination of $\boldsymbol{M}^0$ and $\boldsymbol{M}^1$, i.e.,

$$\mathbf{M} = \boldsymbol{M}^0 || \boldsymbol{M}^1, \tag{12}$$

where $\mathbf{A}||\mathbf{B}$ is the element-wise OR operation of matrices $\mathbf{A}$ and $\mathbf{B}$. In this way, provided located in a same segment, scattered unit label mistakes in $\boldsymbol{M}^0$ have a great chance to be rectified by the successive unit labels in $\boldsymbol{M}^1$.

Given $\mathbf{M}$, the singing voice is finally resynthesized from the input sound mixture, by applying the inverse of the Gammatone filterbank and the technique of overlap and addition.

5 EVALUATION

In this section, we thoroughly evaluate each of the above three modules as well as the overall singing voice separation algorithm. The dataset used for evaluation is MIR-1K published in [10], which contains 1000 song clips recorded at 16 kHz sampling rate and 16 bit quantization, with durations ranging from 4 to 13 s. These clips are extracted from 110 karaoke Chinese pop songs performed by male and female amateurs, with accompaniment and vocals recorded in the left and right channels, respectively. To provide the ground truth, some useful information such as manual annotations of the singing pitch contours, indices of the vocal/nonvocal frames, indices and types for unvoiced frames, and lyrics etc., are included in the dataset. On the basis of MIR-1K dataset, we create three sets of sound mixtures at different qualities for evaluation. To be exact, for each original music clip in MIR-1K, the singing voice and music accompaniment are mixed into a monaural signal under three different signal-to-noise ratios (SNRs), i.e., -5 dB (accompaniment is louder), 0 dB (same level), and 5 dB (singing voice is louder). Note that in this circumstance, signal refers to the singing voice, while the accompaniment is deemed as noise.

5.1 Evaluation of Singing Voice Detection

(1) Dataset Description

All 1000 song clips of the MIR-1K dataset are used to evaluate the performance of singing voice detection described in subsection 4.1. Since this approach is based on supervised classification, the whole dataset is further partitioned into two nonintersecting subsets with nearly equal size (483 vs. 517) for training and testing. The final results are given through a two-fold cross validation.

(2) Performance Measure

The performance of singing voice detection is measured with frame-level precision and recall of the vocal and nonvocal portions. Take vocal frame detection as an example, suppose $N_{\text{classified}}$ be the number of frames that are alleged to be vocal by classifier, N_{correct} be the number of frames that are correctly classified as vocals, and $N_{\text{reference}}$ be the number of total true vocal frames, then the precision and recall are defined as $N_{\text{correct}}=N_{\text{classified}}$ and $N_{\text{correct}}=N_{\text{reference}}$, respectively.

(3) Experimental Settings and Results

In experiment, the input song clip is first segmented into frames of 40 ms long with 50% overlap, and a 39-dimensional MFCC feature vector (12 cepstral coefficients and the log energy, together with their first and second order derivations) is extracted from each frame to characterize the audio content. Then, a Gaussian mixture model (GMM) with 32 components, each having a diagonal covariance matrix, is trained to model the MFCC distribution of vocal and nonvocal frames, respectively. Parameters of the GMMs are initialized using the K-means algorithm and iteratively adjusted via the expectation-maximization (EM) algorithm. Each of the GMMs is considered as a state in a fully connected HMM. The transition probabilities of the HMM are obtained by frame counting in the training set, and the Viterbi algorithm is used to decode the sound mixture into vocal and nonvocal portions.

Fig. 5 illustrates the results of frame-level singing voice detection in the precision-recall space at different SNRs of -5, 0 and 5 dB. As can be seen, precisions and recalls keep going upwards with the increase of SNR, no matter for vocal or nonvocal frames. Meanwhile, the performances of vocal frame detection are notably better than those of nonvocal frame detection in all three SNRs. We claim that both of the observations are reasonable, as song clips in the dataset involve significantly more vocal portions (beyond 70%) than nonvocal portions. In the circumstance of singing separation, what to be considered is mainly the precisions and recalls of vocal frames, and in Fig. 5 they have been shown to be high enough (e.g., 93% precision and 83% recall under 5 dB SNR) for further singing voice processing.

5.2 Evaluation of Singing Pitch Detection

(1) Dataset Description

All 1000 music clips of the MIR-1K dataset are used to evaluate the performance of singing pitch detection.

(2) Performance Measure

The performance of singing pitch detection is measured using the frame-level gross error rate, which occurs if the absolute difference between the detected pitch and the ground truth is more than a semitone.

(3) Experimental Settings and Results

As shown in Fig. 1, the singing pitch detection module consists of two steps. The parameter values used in singing voice enhancement are experimentally set and summarized in Table 1. The parameters θ_t(time axis) and θ_s(frequency axis) are obtained by grid search, that is, adjust the parameters step-by-step within a specific range determined in experiment, and select the ones that make the results best. And the pitch detection step follows the default settings of Praat [30], except that the plausible pitch range is set to [80, 500] Hz, and the frame length is set to 40 ms with an overlap of 20 ms.

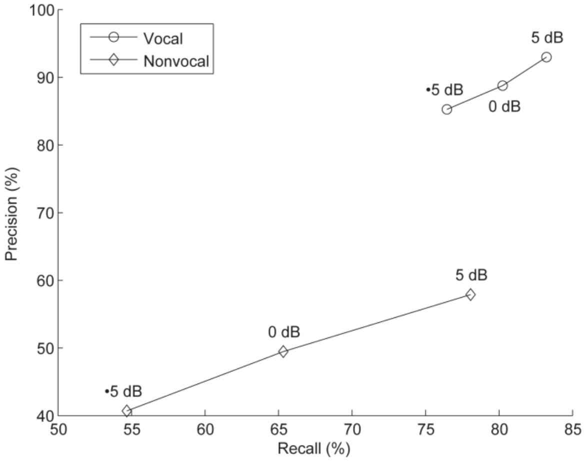

Figure 5: Performance of singing voice detection.

Table 1: Parameter values used for singing voice enhancement

The first stage	The second stage
Frame length = 256 ms Frame overlap = 128 ms # NMF component = 30 $\theta_s = 1200$	Frame length = 32 ms Frame overlap = 16 ms # NMF component = 30 $\theta_t = 300$

To get a direct impression on the effect of singing voice enhancement, Fig. 6 shows a typical example taking a snippet extracted from the MIR-1K dataset as input. The spectrograms of the original song mixture, the original clean singing voice, and the enhanced singing voice are shown in Fig. 6(a), (b), and (c), respectively. As can be obviously seen, Fig. 6(c) resembles Fig. 6(b) pretty much, while is quite dissimilar with Fig. 6(a). In other words, the proposed method retains most of the singing components in the mixture, while eliminates lots of the non-singing ones.

In light of the introduction of section 4.2, singing enhancement includes two NMF-based decompositions that are consecutively performed with long-frame and short-frame spectrograms, respectively. To quantitatively assess the effect of singing enhancement on subsequent pitch detection, four groups of experiments under different conditions are designed as below,

- *Without SVE*: Pitch detection without singing voice enhancement (abbreviated as SVE here).
- *With SVE1*: Pitch detection with only the first stage of singing voice enhancement (long-frame type of NMF based enhancement).
- *With SVE2*: Pitch detection with only the second stage of singing voice enhancement (short-frame type of NMF based enhancement).
- *With SVE*: Pitch detection with two-stage singing voice enhancement.

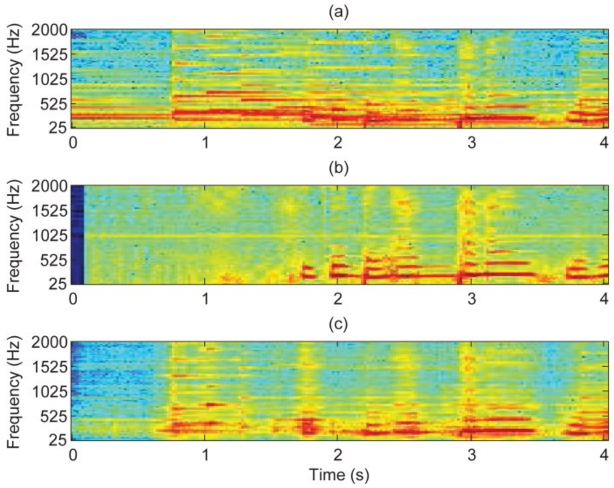

Figure 6: Singing voice enhancement for a snippet of the song clip Ani_4_01 in the MIR-1K dataset at 0-dB SNR. (a) Spectrogram of the original song mixture. (b) Spectrogram of the clean singing voice. (c) Spectrogram of the singing-voice-enhanced signal.

Results of the above comparative experiments are illustrated in Fig. 7. Obviously, all three enhancement mechanisms, i.e., experiments of *With SVE1*, *With SVE2*, and *With SVE*, show their effectiveness in improving the performance of singing pitch detection. Particularly, applying the entire two-stage process (*With SVE*) achieves the best results (or the minimum pitch detection error rates) in all SNRs, indicating that the two stages are complementary. As for the comparison of the two individual stages, applying only the first stage of the enhancement method (*With SVE1*) achieves better results than applying only the second stage (*With SVE2*). This result is reasonable and expectable. The long-frame based enhancement stage gives emphasis to the spectral smoothness, and therefore attenuates the energy of harmonic chordal sounds which usually create much difficulty for singing pitch detection. In contrast, the short-frame based enhancement stage highlights the temporal smoothness, and hence weakens the energy of percussive sounds, which are aperiodic and do not do as much damage as chordal sounds for singing pitch detection.

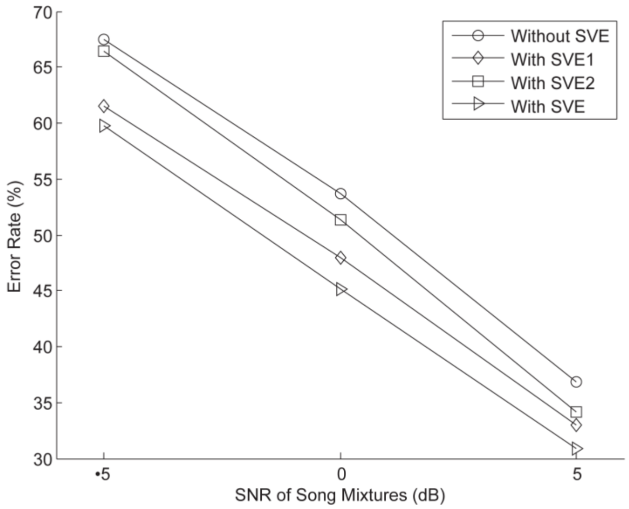

Figure 7: Performance of singing pitch detection.

5.3 Evaluation of Singing Voice Separation

(1) Dataset Description

All 1000 song clips of the MIR-1K dataset are used to evaluate the performance of singing voice separation.

(2) Performance Measure

Given a resynthesized singing voice $\hat{v}$ and the original clean singing voice $\boldsymbol{v}$, the signal-to-distortion ratio (SDR) is first defined as below to measure the quality difference between them.

$$\mathbf{SDR}(\hat{\boldsymbol{v}},\boldsymbol{v}) = \mathbf{10}\ \boldsymbol{log}_{\mathbf{10}}\left[\frac{\langle\hat{\boldsymbol{v}},\boldsymbol{v}\rangle^{2}}{\|\hat{\boldsymbol{v}}\|^{2}\|\boldsymbol{v}\|^{2}-\langle\hat{\boldsymbol{v}},\boldsymbol{v}\rangle^{2}}\right], \qquad (\mathbf{13})$$

where $\langle\hat{\boldsymbol{v}},\boldsymbol{v}\rangle$ is the inner product of vectors $\hat{\boldsymbol{v}}$ and $\boldsymbol{v}$, and $\|\boldsymbol{v}\|^2$ is the energy of $\boldsymbol{v}$.

Next, normalized SDR (NSNR) is defined in Eq. (14) following the suggestion in [10]. It is the improvement of the SDR between the original mixture $\mathbf{x}$ and the estimated voice $\hat{\boldsymbol{v}}$, and used to assess the separation performance of each mixture.

$$\mathbf{NSDR}(\hat{\boldsymbol{v}},\boldsymbol{v},\boldsymbol{x}) = \mathbf{SDR}(\hat{\boldsymbol{v}},\boldsymbol{v}) - \mathbf{SDR}(\boldsymbol{x},\boldsymbol{v}) \qquad (\mathbf{14})$$

Finally, the global NSDR (GNSDR) is calculated in terms of Eq. (15) as the final measure of evaluating the overall separation performance.

$$\mathbf{GNSDR} = \frac{\sum_{n=1}^{N}\boldsymbol{\omega}_{\boldsymbol{n}}\mathbf{NSDR}(\hat{\boldsymbol{v}}_{\boldsymbol{n}},\boldsymbol{v}_{\boldsymbol{n}},\boldsymbol{x}_{\boldsymbol{n}})}{\sum_{n=1}^{N}\boldsymbol{\omega}_{\boldsymbol{n}}} \qquad (\mathbf{15})$$

where *n* is the index of a song, *N* is the total number of the songs, and ω_n is the length of the n^{th} song. Generally speaking, a higher GNSDR value indicates better separation quality.

(3) Experimental Settings and Results

The frame length used in the separation is 40 ms with an overlap of 50%.

To intuitively show the effectiveness of our singing separation algorithm, a song clip in the MIR-1K dataset at 0-dB SNR is taken as an example for testing and illustration. The waveform of the original clean singing voice, and three signals that are separately resynthesized from the ideal binary mask, the result of pitch-based unit labeling, and our separation algorithm (with 60 segments used in experiment) are drawn together in Fig. 8. Apparently, compared with the signal resynthesized from the result of unit labeling, the output waveform of our algorithm better matches that resynthesized from the ideal binary mask. And it also matches the original clean singing voice better.

Apart from observing the visualized time-domain separated waveforms, the binary masking matrices related to Fig. 8 (b, c, d) are also drawn in Fig. 9 for a more specialized comparison. Specifically, Fig. 9(a) is the ideal binary mask estimated from the premixed singing voice and musical accompaniment, Fig. 9(b) is the mask estimated by pitch based unit labeling, and Fig. 9(c) is the mask estimated by our separation algorithm. It can be clearly observed that, the binary mask estimated by our separation algorithm looks more like the ideal binary mask than that estimated by unit labeling. It retains more energy of the singing voice, which indicates the effectiveness of the proposed method.

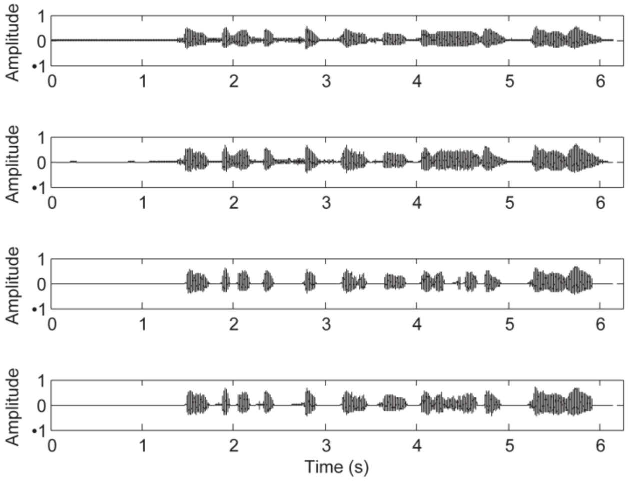

Figure 8: Singing voice separation for the song clip Ani_1_03 in the MIR-1K dataset at 0-dB SNR. (a) Clean singing voice. (b) Signal resynthesized from the ideal binary mask. (c) Signal resynthesized from the result of pitch-based unit labeling. (d) Output of our separation algorithm (number of segments =60).

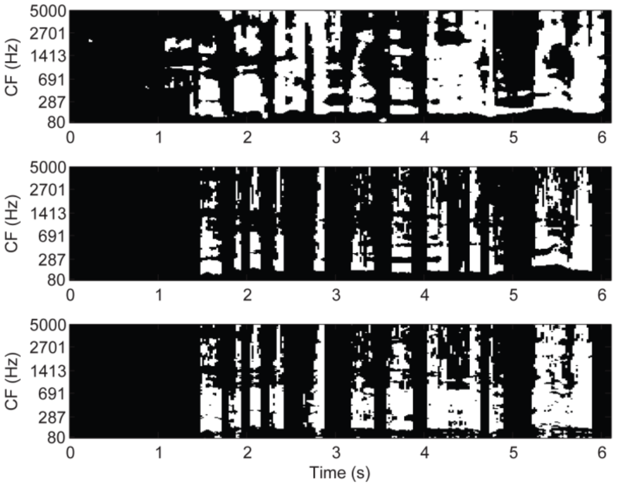

Figure 9: Mask comparison for the song clip Ani_1_03 in the MIR-1K dataset at 0-dB SNR. (a) Ideal binary bask obtained from the premixed singing voice and musical accompaniment. White pixels indicate 1 and black ones indicate 0. (b) The mask estimated by pitch-based unit labeling (3) The mask estimated by our separation algorithm (number of segments = 60).

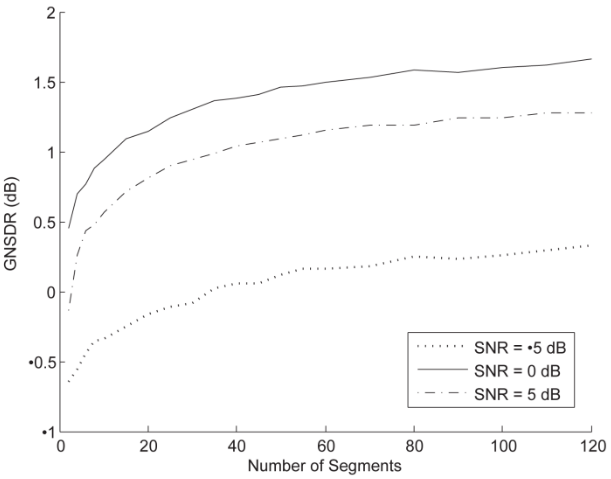

Figure 10: Performance of singing voice separation as a function of the number of segments.

Next experiment is to investigate the relationship between the performance of singing voice separation and the number of NMF segments used in section 4.3. As shown in Fig. 10, increasing the number of segments makes the GNSDRs for all SNRs monotonically rise up until even out after certain points. This phenomenon is as expected. Remind that the task of our segmentation method is to generate a set of indivisible time-frequency segments as another cue to facilitate the singing separation. When the number of segments is small, the size of each segment tends to be large. The larger the segments are, the more difficult to ensure the indivisibility. In contrast, when the number of segments increases, their indivisibility is also increased. And after certain points, the indivisibility converges.

The effects of the two designed improvement approaches, namely singing enhancement and segmentation, are also comparatively studied in the following experiments under different conditions described below, and illustrated in Fig. 10.

- *Without SVE*, without segmentation: Singing voice separation without singing voice enhancement and segmentation. The pitch of the singing voice is extracted from the original sound mixture. The mask for the singing voice is estimated by pitch-based unit labeling only.
- *With SVE*, without segmentation: Singing voice separation with singing voice enhancement, without segmentation. The pitch of the singing voice is extracted from the signal where singing voice is enhanced.
- *Without SVE*, with segmentation: Singing voice separation without singing voice enhancement, with segmentation (number of segments = 60). The mask for the singing voice is estimated by combining $\boldsymbol{M}^0$ and $\boldsymbol{M}^1$.
- *With SVE*, with segmentation: Singing voice separation with both singing voice enhancement and segmentation (number of segments = 60).

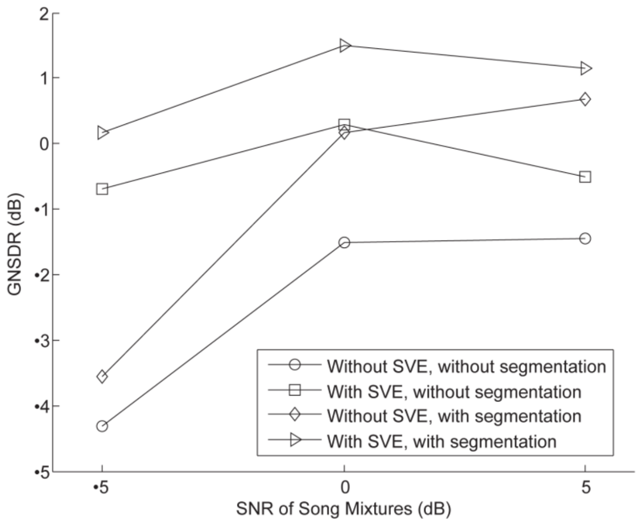

Figure 11: Illustration the effects of the two proposed methods for the performance of singing voice separation.

As demonstrated in the figure, when other factors are fixed (e.g., with or without segmentation), applying the proposed singing voice enhancement method achieves better separation results in all SNRs than not applying it, indicating the importance of accurate singing pitch detection for the performance of pitch-based singing voice separation. Besides, applying the proposed segmentation method also improves the performance of separation in all SNRs, which shows its effectiveness. In other words, the information provided by the segments is reliable, and can be used for more accurate estimation of the ideal binary mask.

Finally, our algorithm is compared with two state-of-the-art singing voice separation algorithms, i.e., the Hsu method [10] and the Rafii method [19]. The result of the comparison is illustrated in Fig. 12. As can be seen, our algorithm significantly outperforms the Hsu method in all three SNRs. It also outperforms the Rafii method in SNRs of 0 dB and 5 dB.

6 CONCLUSION

This paper presents a pitch-based inference algorithm for monaural singing voice separation. To address the limitation caused by inaccurate vocal pitch detection, two methods based on NMF are proposed and integrated into the framework. Specifically, the first method enhances the singing voice in sound mixtures for more accurate singing pitch detection. It utilizes the fluctuation and shortness of the singing voice, and enhances the vocals by sequentially applying NMF on spectrograms of different frame lengths. As for the second method, it decomposes the sound mixture into a set of segments, each of which originates from a single sound source. With these segments, singing-dominant T-F units can be identified based on not only the pitch, but also the origination of the segments they belong to. Quantitative evaluation shows that both of the proposed methods are effective in improving the performance of singing voice separation. In future, experiments should be moved from the manually mixed simple MIR-1K dataset to real CD songs. And in this condition, other long-term mechanisms such as music structure analysis and melody contour similarity etc. can be integrated into the current preliminary framework.

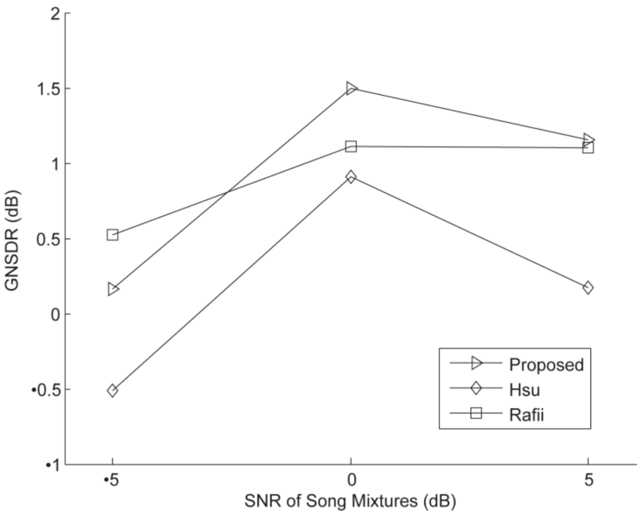

Figure 12: Comparison of our algorithm (number of segments = 60) with the Hsu method and the Rafii method.

7 ACKNOWLEDGEMENTS

This work is supported by NSFC (61171128).

8 REFERENCES

[1]. J. L. Durrieu, G. Richard, B. David and C. Ffievotte. Source/filter model for unsupervised main melody extraction from polyphonic audio signals. IEEE Transactions on Audio, Speech, and Language Processing, 18(3): 564-575, 2010.

[2]. H. Fujihara, M. Goto, T. Kitahara and H. G. Okuno. A modeling of singing voice robust to accompaniment sounds and its application to singer identification and vocal-timbre-similarity-based music information retrieval. IEEE Transactions on Audio, Speech, and Language Processing, 18(3): 638-648, 2010.

[3]. A. Mesaros and T. Virtanen. Automatic recognition of lyrics in singing. EURASIP Journal on Audio, Speech, and Music Processing, 2010, Article ID 546047.

[4]. Y. E. Kim. Singing voice analysis/synthesis. PhD thesis, Massachusetts Institute of Technology, 2003.

[5]. J. Salamon, E. Gomez, D. P. W. Ellis and G. Richard. Melody extraction from polyphonic music signals: Approaches, applications, and challenges. IEEE Signal Processing Magazine, 31(2): 118- 134, 2014.

[6]. J. Salamon and E. Gomez. Melody extraction from polyphonic music signals using pitch contour characteristics. IEEE Transactions on Audio, Speech, and Language Processing, 20(6): 1759-1770, 2012.

[7]. C. L. Hsu and J. S. Roger Jang. Singing pitch extraction by voice vibrato/tremolo estimation and instrument partial deletion. The International Society of Music Information Retrieval Conference (ISMIR), pp. 525-530, 2010.

[8]. T. C. Yeh, M. J. Wu, J. R. Jang, W. L. Chang and I. B. Liao. A hybrid approach to singing pitch extraction based on trend estimation and hidden markov models. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 457-460, 2012.

[9]. Y. Li and D. L. Wang. Separation of singing voice from music accompaniment for monaural recordings. IEEE

Transactions on Audio, Speech, and Language Processing, 15(4): 1475-1487, 2007.

[10]. 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, 18(2): 310-319, 2010.

[11]. D. L. Wang and G. J. Brown, Computational auditory scene analysis: principles, algorithms, and applications. Wiley, New York, 2006.

[12]. H. Tachibana, T. Ono, N. Ono and S. Sagayama. Melody line estimation in homophonic music audio signals based on temporal-variability of melodic source. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 425-428, 2010.

[13]. C. L. Hsu, D. L. Wang, J. S. R. Jang and K. Hu. A tandem algorithm for singing pitch extraction and voice separation from music accompaniment. IEEE Transactions on Audio, Speech, and Language Processing, 20(5): 1482-1491, 2012.

[14]. M. Ryynanen, T. Virtanen, J. Paulus and A. Klapuri. Accompaniment separation and karaoke application based on automatic melody transcription. IEEE International Conference on Multimedia and Expo (ICME), pp. 1417-1420, 2008.

[15]. T. Virtanen, A. Mesaros and M. Ryynanen. Combining pitch-based inference and non-negative spectrogram factorization in separating vocals from polyphonic music. ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing, pp. 17-22, 2008.

[16]. S. Vembu and S. Baumann. Separation of vocals from polyphonic audio recordings. The International Society of Music Information Retrieval Conference (ISMIR), pp. 337-344, 2005.

[17]. A. Chanrungutai and C. A. Ratanamahatana. Singing voice separation in mono-channel music. International Symposium on Communications and Information Technologies, pp. 256-261, 2008.

[18]. A. Liutkus, Z. Rafii, R. Badeau, B. Pardo and G. Richard. Adaptive filtering for music/voice separation exploiting the repeating musical structure. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 53-56, 2012.

[19]. Z. Rafii and B. Pardo. A simple music/voice separation method based on the extraction of the repeating musical structure. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 21-224, 2011.

[20]. P. S. Huang, S. D. Chen, P. Smaragdis and M. Hasegawa-Johnson. Singing-voice separation from monaural recordings using robust principal component analysis. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 57-60, 2012.

[21]. Y. H. Yang. On sparse and low-rank matrix decomposition for singing voice separation. ACM international conference on Multimedia (ACM MM), pp. 757-760, 2012.

[22]. D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. Nature, 401(6755): 788-791, 1999.

[23]. A. Hyvarinen, J. Karhunen and E. Oja. Independent component analysis. Wiley, New York, 2001.

[24]. I. T. Jollifie. Principal component analysis. Springer-Verlag, New York, 1986.

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

[26]. A. Cont. Realtime audio to score alignment for polyphonic music instruments, using sparse non-negative constraints and hierarchical HMMs. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2006.

[27]. E. Benetos, M. Kotti and C. Kotropoulos. Musical instrument classification using non-negative matrix factorization algorithms and subset feature selection. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2006.

[28]. T. Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. IEEE Transactions on Audio, Speech, and Language Processing, 15(3): 1066-1074, 2007.

[29]. D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. Advances in Neural Information Processing Systems, 2001.

[30]. P. Boersma and D. Weenink. Praat: Doing Phonetics by Computer, Ver. 4.0.26. http://www.fon.hum.uva.nl/praat, 2002.

[31]. D. L. Wang. On ideal binary mask as the computational goal of auditory scene analysis. In P. Divenyi, editor, Speech Separation by Humans and Machines, pp. 181-197. Kluwer, Norwell MA, 2005