Technical Report TR61, September 2005
Technical Report: OSU-CISRC-9/05-TR61
Department of Computer Science and Engineering
The Ohio State University, Columbus, OH 43210-1277, USA
Web site: **http://www.cse.ohio-state.edu/research/techReport.html**
Ftp site: **ftp.cse.ohio-state.edu**
Login: **anonymous**
Directory: **pub/tech-report/2005**
File in pdf format: **TR61.pdf**

Separation of Singing Voice from Music Accompaniment for Monaural Recordings

Yipeng Li
Department of Computer Science and Engineering
The Ohio State University
Columbus, OH, 43210-1277, USA
liyip@cse.ohio-state.edu

DeLiang Wang
Department of Computer Science and Engineering
& Center of Cognitive Science
The Ohio State University
Columbus, OH, 43210-1277, USA
dwang@cse.ohio-state.edu

Abstract

Separating singing voice from music accompaniment is very useful in many applications, such as lyrics recognition and alignment, singer identification, and music information retrieval. Although speech separation has been extensively studied for decades, singing voice separation has been little investigated. We propose a system to separate singing voice from music accompaniment for monaural recordings. Our system consists of three stages. The singing voice detection stage partitions and classifies an input into vocal and non-vocal portions. For vocal portions, the predominant pitch detection stage detects the pitch of the singing voice and then the separation stage uses the detected pitch to group the time-frequency segments of the singing voice. Quantitative results show that the system performs the separation task successfully.

I. INTRODUCTION

It is well known that the human auditory system has a remarkable capability in separating sounds originated from different sources. One important aspect of this capability is hearing out singing voice (also called vocal line) accompanied by musical instruments. Although this task seems effortless to humans, it turns out to be very difficult for machines. To date, few systems have addressed the problem of separating singing voice from music accompaniment systematically. A singing voice separation system has its applications in areas such as automatic lyrics recognition and alignment. Automatic lyrics recognition often requires that the input to the system is solo singing voice [39], which is often unrealistic since for almost all songs singing voice is accompanied by musical instruments. However such a requirement can be satisfied if successful separation of singing voice is used for pre-processing. Aligning lyrics to singing voice is a key step for applications such as karaoke [43] and currently it remains labor-intensive work. Automating this process therefore will be of considerable help. An accurate lyrics alignment system will allow listeners to follow singing voice more easily. But the task of aligning lyrics to singing voice becomes difficult when accompaniment is present, and a separation system can be used to alleviate the problem. Singer identification is another promising area for applying such a system. Several studies [3], [19], [45] have addressed the problem of singer identification in real recordings but the attempts so far have not separated a singer's voice. With singing voice separation, the accuracy of singer identification is expected to improve. Another area where singing voice separation can be applied is musical information retrieval. Singing voice carries useful information, such as melody, for identifying a song in a database and singing voice separation can facilitate the extraction of such information.

Although songs today are often recorded in stereo, we focus on singing voice separation for monaural recordings where only one channel is available. This is because a solution for monaural recordings is indispensable in many cases, such as for recordings of live performance (non-studio recordings). Such a solution can also assist in analysis of stereo recordings. It is well known that human listeners have little difficulty in hearing out singing voice even when it is recorded with music accompaniment in a single channel. Therefore a separation system for monaural recordings could also enhance our understanding of how the human auditory system performs this task.

Although speech separation has been extensively studied, few studies are devoted to separating singing voice from music accompaniment. Since singing voice is produced by the speech organ, it may be sensible to explore speech separation techniques for singing voice separation. Before applying such techniques, it is instructive to compare singing voice and speech. Singing voice bears many similarities to speech. For example, they both consist of voiced and unvoiced sounds. But the differences between singing and speech are also significant. A well known difference is the presence of an additional formant, called the singing formant, in the frequency range of 2000–3000 Hz in operatic singing. This singing formant helps the voice of a singer to stand out from the accompaniment [35]. However, the singing formant does not exist in many other types of singing [5], [22], such as the ones in rock and country music we examine in this paper. Another difference is related to the way singing and speech are uttered. During singing, a singer usually intentionally stretches the voiced sound and shrinks the unvoiced sound to match other musical instruments. This has two direct consequences. First, it alters the percentage of voiced and unvoiced sounds in singing. The large majority of sounds generated during singing is voiced (about 90%) [20] while speech has a larger amount of unvoiced sounds [40]. Second, the pitch dynamics (the evolution of pitch in time) of singing voice tends to be piece-wise constant with abrupt pitch changes in between. This is in contrast with the declination phenomenon [30] in natural speech where pitch frequencies slowly drift down with smooth pitch change in an utterance. Besides these differences, singing also has a wider pitch range. The pitch range of normal speech is between 80 and 400 Hz while the upper pitch range of singing can be as high as 1400 Hz for soprano singers [36].

From the sound separation point of view, the most important difference between singing and speech is the nature of other concurrent sounds. In a real acoustic environment, speech is usually contaminated by interference that can be harmonic or nonharmonic, narrowband or broadband. Interference in most cases is independent of speech in the sense that the spectral contents of target speech and interference are uncorrelated. For recorded singing voice, however, it is almost always accompanied by musical instruments that in most cases are harmonic, broadband, and are correlated with singing since they are composed to be a coherent whole with the singing voice. This difference makes the separation of singing voice from the accompaniments potentially more challenging.

In this paper we propose a singing voice separation system. Our system consists of three stages. The first stage performs singing voice detection. In this stage, the input is partitioned and classified into vocal and non-vocal portions. Then vocal portions are used as input to a stage for predominant pitch detection. In the last stage, detected pitch contours are used for singing voice separation where we extend a system for pitch-based separation [18]. The output of the overall system is separated singing voice.

The remainder of this paper is organized as follows. Section II presents related work to singing separation. Section III gives an overview of the system and describes each stage in detail. Section IV presents the systematic evaluation of each stage as well as the overall system. The last section gives further discussion and concludes the paper.

II. RELATED WORK

To our knowledge, only a few systems directly address the separation of singing voice from music accompaniment. Wang [38] developed a system for singing voice separation by using a harmonic-locked loop technique to track a set of harmonically related partials. In his system, the fundamental frequency of the singing voice needs to be known *a priori*. The system also does not distinguish singing voice from other musical sounds, i.e., when the singing voice is absent the system incorrectly tracks partials which belong to some other harmonic source. The harmonic-locked loop requires the estimation of a partial's instantaneous frequency, which is not reliable in the presence of other partials and other sound sources. Therefore the system only works in conditions where the energy ratio of singing voice to accompaniment is high. Another system proposed by Meron and Hirose [27] aims to separate singing voice from piano accompaniment. For the system to work a significant amount of prior knowledge is required, such as the partial tracks of premixing singing voice and piano or the music score for piano sound. This prior knowledge in most cases is not available therefore the system cannot be applied for most real recordings.

Since we pursue a sound separation solution for monaural recordings, approaches to speech separation based on microphone arrays are not applicable. Speech enhancement can be employed for separation for monaural recordings. However, it tends to make strong assumptions about interference, such as stationarity, which generally are not satisfied for music accompaniment. An emerging approach for general sound separation exploits the knowledge gained from the human auditory system. In an influential book [6], Bregman proposed that the auditory system employs a process called *auditory scene analysis* (ASA) to organize an acoustic mixture into different perceptual streams which correspond to different sound sources. This process involves two main stages: Segmentation stage and grouping stage. In the segmentation stage, the acoustic input is decomposed into time-frequency (T-F) segments, each of which likely originates from a single source. In the grouping stage, segments from the same source are grouped according to a set of ASA principles, such as common onset/offset and harmonicity. ASA has inspired researchers to build *computational auditory scene analysis* (CASA) systems for sound separation [7], [12]. Compared to other sound separation approaches, CASA makes minimal assumptions about concurrent sounds; instead it relies on the intrinsic properties of sounds and therefore shows a greater potential in singing voice separation for monaural recordings.

The work by Mellinger [26] represents the first CASA attempt to musical sound separation. His system extracts onset and common frequency variation and uses them to group frequency partials from the same musical instrument together. However these two cues seem not strong enough to separate different sounds apart. The author suggested that other cues, such as pitch, should be incorporated for the purpose of sound separation. The pitch cue, or the harmonicity principle, is widely used in CASA systems. For example, Godsmark and Brown [14] developed a CASA system which uses the harmonicity and other principles in a blackboard architecture for grouping. Goto [15] developed a music-scene-description system which uses the harmonicity principle for melody detection.

Recently a sound separation system developed by Hu and Wang [18] successfully segregates voiced speech from acoustic interference based on pitch tracking and amplitude modulation. The Hu–Wang system employs different segregation methods for resolved and unresolved harmonics. Systematic evaluation over a commonly used database shows that the system performs significantly better over previous systems.

The Hu–Wang system relies heavily on pitch to group segments therefore the accuracy of pitch detection is critical. However, their system obtains its initial pitch estimation from the time lag corresponding to the maximum of a summary autocorrelation function. This estimation of pitch is unreliable for singing voice as shown in [24] and it limits the separation performance of the system. In [24] we proposed a predominant pitch detection algorithm which can detect the pitch of singing voice for different musical genres even when the accompaniment is strong. The Hu–Wang system assumes that voiced speech is always present. For singing voice separation this assumption is not valid. Therefore it is necessary to have a mechanism to distinguish portions where singing voice is present from those where it is not. On the other hand, although their system cannot separate unvoiced speech, this limitation is

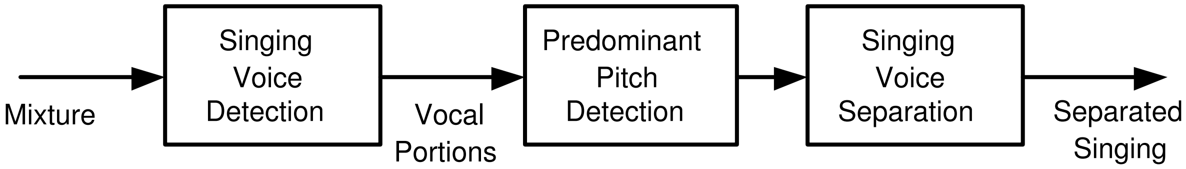

Fig. 1. Schematic diagram of the proposed system

less severe for singing voice separation because unvoiced singing comprises a smaller percentage in terms of time and its contribution to the intelligibility of singing is less than that to the intelligibility of speech.

III. SYSTEM DESCRIPTION

Our system is illustrated in Fig. 1. The input to the system is a mixture of singing voice and music accompaniment. In the singing voice detection stage, the input is first partitioned into spectrally homogeneous portions by detecting significant spectral changes. Then each portion is classified, based on the overall likelihood, as a vocal portion in which singing voice is present, or a non-vocal portion in which singing voice is absent.

The predominant pitch detection stage detects the pitch contours of singing voice for vocal portions. In this stage, a vocal portion is first filtered by a filterbank which simulates the frequency decomposition of the auditory periphery. For high-frequency channels, the envelopes of filter outputs are extracted. After auditory filtering normalized correlograms are calculated. Generally speaking, the peaks in normalized correlograms contain periodicity information of the input. However some peaks may give misleading information because of the presence of music accompaniment. We apply channel/peak selection to normalized correlograms to obtain useful periodicity information. Next the probability of a pitch hypothesis is calculated by integrating the periodicity information across all frequency channels. A hidden Markov model (HMM) is used to model the pitch generation process and form a continuous pitch contour. In order to reduce the interference of other harmonic sounds from accompaniment, the HMM simultaneously tracks up to 2 predominant pitch contours. Finally the most probable pitch hypothesis sequence is identified using the Viterbi algorithm. The first pitch contour of this optimal sequence is considered as the pitch contour of the singing voice.

The last stage of the system, the singing voice separation stage, extends the Hu–Wang system [18]. With more reliable pitch contours, the original system can be simplified. Our separation stage has two main steps: Segmentation step and grouping step. In the segmentation step, a vocal portion is decomposed into T-F units and segments are formed from T-F units based on temporal continuity and cross-channel correlation. T-F units are also labeled as singing dominant or accompaniment dominant using the detected pitch contour. Then additional segments are formed from those T-F units which have high cross-channel correlation of envelopes and are labeled as singing dominant but not belonging to any segments generated before. In the grouping stage, segments in which the majority of T-F units are labeled as singing dominant are grouped to form the foreground stream, which corresponds to the singing voice. Separated singing voice is then resynthesized from the segments belonging to the foreground stream.

The following subsections explain each stage in detail.

A. SINGING VOICE DETECTION

The goal of this stage is to partition the input into vocal and non-vocal portions. Therefore this stage needs to address the classification and partition problem. For the classification problem, the two key components in the system design are features and classifiers. Different features have been explored for singing voice detection. These features include mel-frequency cepstral coefficients (MFCC) [2], [25], linear prediction coefficients (LPC) [25], perceptual linear prediction coefficients (PLP) [3], and the 4-Hz harmonic coefficient [10]. MFCC, LPC and PLP are also widely used for general sound classification tasks and they are the so-called short-term features because they are calculated in short-time windows. Similarly different classifiers have also been explored, including Gaussian mixture models (GMM) [10], support vector machines (SVM) [25], and multi-layer perceptrons (MLP) [3]. As for the partition problem, HMM [2] and rule-based post-processing [10] have been proposed. The underlying assumption of these two methods is that a vocal or non-vocal portion sustains a certain amount of time therefore the short-term classification should not jump back and forth rapidly.

Several studies have shown that MFCC is a good feature for sound classification, even for mixtures. Li et al. [23] compared different features in classifying a sound into seven classes and found that MFCC provides the best classification. In Berenzweig's work [2], MFCC-based classification also performs well compared to other more complicated features. Therefore we use MFCC as the short-term feature for classification and calculate it for each frame. A frame is a block of samples within which the signal is assumed to be near stationary. However, the short-term classification is not reliable since the information within a frame is limited. Observe that, when a new sound enters a mixture, it usually introduces significant spectral changes. As a result, the possible instances of a sound event in a mixture can be determined by identifying significant spectral changes. This idea is more compelling in singing voice detection since a voice more likely joins the accompaniment at beat times in order to conform with the rhythmic structure of a song [31]. Beats are regularly spaced pulses that give the sensation of the rhythm of music. Because beats are usually generated by percussive instruments, they tend to introduce strong spectral perturbations. The portion between two consecutive spectral change instances is relatively homogeneous, and the short-term classification results can then be pooled over the portion to yield more reliable classification. Therefore we propose to first partition the input into portions by detecting instances when significant spectral changes occur, and then pool the likelihoods over all the frames of a portion and classify the portion to the class with a larger overall likelihood.

We use a simple spectral change detector proposed by Duxbury et al. [13]. This detector calculates the Euclidian distance in the complex domain between the expected spectral value and the observed one in a frame:

$$\eta(m) = \sum_k (|\hat{S}_k(m) - S_k(m)|) \tag{1}$$

where $S_k(m)$ is the observed spectral value at frame m and frequency bin k. $\hat{S}_k(m)$ is the expected spectral value of the same frame and the same bin, calculated by:

$$\hat{S}_k(m) = |S_k(m-1)|\hat{\phi}_k(m) \tag{2}$$

where $|S_k(m-1)|$ is the spectral magnitude of the previous frame at bin k. $\hat{\phi}_k(m)$ is the expected phase which can be calculated as the sum of the phase of previous frame and the phase difference between the previous two frames:

$$\hat{\phi}_k(m) = \tilde{\varphi}_k(m-1) + (\tilde{\varphi}_k(m-1) - \tilde{\varphi}_k(m-2)) \tag{3}$$

$\tilde{\varphi}_k(m-1)$ and $\tilde{\varphi}_k(m-2)$ are the unwrapped phases for frame $m-1$ and frame $m-2$, respectively.

A local peak in $\eta(m)$ indicates a spectral change, which can either be that the spectral contents of a sound are changing or a new sound is entering the scene. To accommodate the dynamic range of the spectral change as well as spectral fluctuations, we apply weighted dynamic thresholding to identify the instances of significant spectral changes. Specifically, a frame m will be recognized as an instance of significant spectral change if $\eta(m)$ is a local peak and $\eta(m)$ is greater than the weighted median value in a window of size H:

$$\eta(m) > C_t \text{median}(\eta(m - \frac{H}{2}), ..., \eta(m + \frac{H}{2})) \tag{4}$$

where C_t is the weighting factor. Finally instances that are close are merged – specifically if the enclosed interval is less than T_m. We find this complex-domain spectral change detection gives slightly better results in term of overall classification accuracy than real-domain detection where the magnitude difference is used.

After the input is partitioned, we pool the information in a whole portion to obtain more reliable classification. A portion is classified as vocal if the overall likelihood of the vocal class is greater than that of the non-vocal class. Formally let $\{\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_M\}$ be a set of feature vectors for a portion with M frames. Let $\log p(\boldsymbol{X}|c_v)$ and $\log p(\boldsymbol{X}|c_{nv})$ represent the log likelihood of an observed feature vector $\boldsymbol{X}$ being in the vocal class c_v and the non-vocal class c_{nv}, respectively. Then a portion is classified as vocal if:

$$\sum_{j=1}^{M} \log p(\boldsymbol{X}_j|c_v) > \sum_{j=1}^{M} \log p(\boldsymbol{X}_j|c_{nv}) \tag{5}$$

We choose GMM as the classifier for it has been widely and successfully used with MFCC for audio classification tasks [2], [23].

B. PREDOMINANT PITCH DETECTION

In the second stage, portions classified as vocal are used as input to a predominant pitch detection algorithm we proposed in [24]. This algorithm is extended from the one by Wu et al. [44], which detects multipitch contours for noisy speech. With multipitch detection, pitches from singing voice and other harmonic musical instruments can be tracked separately and therefore the interference of harmonic accompaniment can be effectively reduced. Our algorithm is more suitable for the pitch detection of singing voice than the original algorithm.

Our predominant pitch detection starts with an auditory peripheral model for frequency decomposition. The signal is sampled at 16 kHz and passed through a 128-channel gammatone filterbank. The center frequencies of the channels are equally distributed on the equivalent rectangular bandwidth (ERB) scale between 80 Hz and 5 kHz. Channels with center frequencies lower than 800 Hz are designated as low-frequency channels and others are designated as high-frequency channels. In each high-frequency channel, the envelope of the filter output is extracted using the Teager energy operator and a low-pass filter with the stop frequency 800 Hz [44].

After peripheral processing, a normalized correlogram is computed for each channel c with a frame length of 16 ms and a frame shift of 10 ms:

$$A(c,m,\tau) = \frac{\sum_{n=-N/2}^{N/2} r(c,m+n)r(c,m+n+\tau)}{\sqrt{\sum_{n=-N/2}^{N/2} r^2(c,m+n)}\sqrt{\sum_{n=-N/2}^{N/2} r^2(c,m+n+\tau)}} \tag{6}$$

where r is the filter output for low-frequency channels and the envelope of the filter output for high-frequency channels. Here, $N = 256$ corresponds to the frame length of 16 ms and the normalized correlogram is calculated for time lag τ from 0 to 200. m is the frame index and n is the time step index.

The peaks in the normalized correlograms indicate the periodicity of the input. However, the presence of accompaniment makes the peaks in some channels misleading. Percussive accompaniment usually, in particular, has significant energy in the low-frequency channels, which makes the peaks in those channels particularly unreliable. Consequently we apply channel selection to the low-frequency channels. Specifically, a channel is selected if the maximum value of its normalized correlogram in the plausible pitch range (80–500 Hz) exceeds a threshold $\theta = 0.945$. Note that in a selected channel, usually only one harmonic is dominant and this harmonic may or may not belong to singing voice. For a selected low-frequency channel, the time lags of peaks are included in the set of peaks Φ. For high-frequency channels, unlike in [44], we retain all the channels and apply peak selection to conform with the beating phenomenon [17]. Specifically, only the first peak at a non-zero lag in the plausible pitch range of the normalized correlogram is retained and the corresponding time lag is included in Φ.

Not applying channel selection in the high-frequency channels may introduce noisy peaks, whose time lags do not correspond well to the fundamental period of the singing voice. But we have found experimentally that the time lag of the first peak within the pitch range in a noisy high-frequency channel is still a good indicator of the true pitch of singing voice in many cases. We emphasize that this is not caused by the singing formant since for the genres tested the singing formant is not present. It is, however, possible that the high frequency components become more salient because of singing. More importantly, keeping all the high-frequency channels makes more channels available, which is important for distinguishing different harmonic sources as well as for reducing pitch–halving errors. The peak selection method in high-frequency channels is motivated by the beating phenomenon, i.e., high-frequency channels respond to multiple harmonics and the envelope of the response fluctuates at the fundamental frequency [17], [18]. Therefore the selected time lag of the first peak corresponds to the fundamental period of some harmonic source. Compared to [44], this channel/peak selection method is tailored for the pitch detection of singing voice in the presence of music accompaniment. As a result, our method along with the following statistical cross-channel integration substantially improves the performance of pitch detection for singing voice.

Next, the probability of a pitch hypothesis is evaluated. Notice that, if voiced singing is dominant in a channel, the distance, δ, between the true pitch period d and the time lag of the closest observed peak l tends to be small. With clean singing voice available, the statistics of δ can be extracted. This statistic can be quantitatively described by a Laplacian distribution [24], which centers at zero and exponentially decreases as $|\delta|$ increases:

$$L(\delta;\lambda_c) = \frac{1}{2\lambda_c}\exp\left(-\frac{|\delta|}{\lambda_c}\right) \tag{7}$$

where the distribution parameter λ_c indicates the spread of the Laplacian distribution. The probability distribution of δ in a channel c is defined as:

$$p_c\left(\delta\right) = \left(1-q\right)L\left(\delta;\lambda_c\right) + qU\left(\delta;\eta_c\right) \tag{8}$$

where the uniform distribution $U(\delta;\eta_c)$ is used to model background noise and η_c indicates the possible range of pitch periods. q is the partition factor $(0 < q < 1)$.

The Laplacian distribution parameter λ_c gradually decreases as the channel center frequency increases. When estimated for each frequency channel, we approximate this relation by $\lambda_c = a_0 + a_1 c$. A maximum likelihood method is used to estimate the parameters a_0, a_1, and q according to the statistics of δ collected from singing voice alone. Due to the different properties of low- and high-frequency channels, the parameters are estimated in these two ranges separately.

The statistics of δ when singing voice is accompanied by musical instruments can also be extracted. Since the sound of each musical instrument in the accompaniment is not available, we only collect δ from channels where singing voice is dominant, i.e., the energy of singing voice is stronger than that of accompaniment. Here we assume that δ in these channels is similarly distributed. The probability distribution of δ is denoted as $p'_c(\delta)$ and has the same form as in (8). Distribution parameters are also estimated using the maximum likelihood method based on the statistics collected from mixtures of singing voice and accompaniment. The resulting parameters for $p_c(\delta)$ and $p'_c(\delta)$ are similar to those in [44] therefore are not listed here. For more details about parameter estimation as well as the probability formulation described in (7)–(8) and in the following (9)–(13), the interested reader is referred to [44], [24]. All the statistics used to train the model is collected from a small database which consists of clips different from those used for testing.

With the distribution of δ available, the channel conditional probability for 1- and 2-pitch hypotheses can be formulated. For the 1-pitch hypothesis:

$$p_c(\Phi_c|d) = \begin{cases} p_c(\delta), \text{if channel } c \text{ is selected} \\ q_1(c)U(0;\eta_c), \text{otherwise} \end{cases} \tag{9}$$

where d is the hypothesized pitch and δ is the difference between d and the time lag of the closest peak in Φ_c, which is the set of peaks selected for channel c. $q_1(c)$ is the partition factor for channel c in the 1-pitch case. If a channel is not selected, then the probability of background noise channels is used.

The channel conditional probability of a 2-pitch hypothesis can be formulated as:

$$p'_c(\Phi_c|\left(d_1,d_2\right)) = \begin{cases} q_2(c)U(0;\eta_c) \quad \text{if channel } c \text{ is not selected} \\ p'_c(\Phi_c|d_1) \quad \text{if channel } c \text{ belongs to } d_1 \\ max(p'_c(\Phi_c|d_1), p'_c(\Phi_c|d_2)) \quad \text{else} \end{cases} \tag{10}$$

where d_1 and d_2 are the hypothesized pitches. $q_2(c)$ is the partition factor for channel c in the 2-pitch case. $p'_c(\Phi_c|d)$ is the same as $p'_c(\delta)$ mentioned before. Channel c belongs to d_1 if the distance between d_1 and the time lag corresponding to the closest peak in that channel is less than $5\lambda_c$. This condition essentially tests whether channel c is dominated by d_1. In this way, the formulation distinguishes $p'_c(\Phi_c|(d_1,d_2))$ from $p'_c(\Phi_c|(d_2,d_1))$. In the former case, the dominance of d_1 is first tested while in the latter case the dominance of d_2 is first tested. If the hypothesized pitch d_1 dominates channel c, $p'_c(\Phi_c|(d_1,d_2))$ exceeds $p'_c(\Phi_c|(d_2,d_1))$, and vice versa. In other words, the first pitch in a 2-pitch hypothesis is the dominant one.

Due to the wideband nature of singing voice, the responses of different channels are correlated therefore the statistical independence assumption is generally invalid. However, according to [16], this can be partially remedied by smoothing the combined probability estimates by taking a root greater than 1. Hence, the probability of the 1-pitch and 2-pitch hypothesis across all the frequency channels can be obtained by

$$p\left(\Phi|d\right) = k_1\sqrt[b]{\prod_c p_c\left(\Phi_c|d\right)} \tag{11}$$

TABLE I
TRANSITION PROBABILITIES BETWEEN STATE SUBSPACES OF PITCH

	$\rightarrow\Omega_0$	$\rightarrow\Omega_1$	$\rightarrow\Omega_2$
Ω_0	0.2875	0.7125	0.0000
Ω_1	0.0930	0.7920	0.1150
Ω_2	0.0000	0.0556	0.9444

$$p'\left(\Phi|\left(d_1,d_2\right)\right)=k_2\sqrt[b]{\prod_c p'_c\left(\Phi_c|\left(d_1,d_2\right)\right)} \tag{12}$$

where k_1 and k_2 are the normalization factors. b is used to compensate for statistical dependency among channels. Note that the combined probability estimate preserves the dominance of the first pitch in a 2-pitch hypothesis.

The final part of our pitch detection algorithm performs pitch tracking by an HMM, which models the pitch generation process. The pitch state space is a union of three i-dimensional subspaces $\bigcup_{i=0}^{2}\Omega_i$, each of which represents the collection of hypotheses with i pitches. In each frame, a hidden node represents the pitch state space and the observation node represents the set of observed peaks Φ. The observation probability is calculated as (11) and (12). The pitch transition between consecutive frames, i.e., between different states in the pitch state space, is described by pitch dynamics, which has two components: the transition probability between different pitch configurations in the same pitch subspace and the jump probability between different pitch subspaces. The transition behavior within Ω_1 is well described by a Laplacian distribution:

$$p(\Delta)=\frac{1}{2\lambda}\exp(-\frac{|\Delta-\mu|}{\lambda}) \tag{13}$$

where Δ is the change of pitch periods in two consecutive frames of a pitch contour and μ is the mean of the changes. We extract Δ from the true pitch contours of clean singing voice and estimate μ and λ using the maximum likelihood method. For singing voice, the estimated values are $\lambda=0.7$ and $\mu=0$, respectively. The zero value of μ indicates that pitch contours of singing voice do not exhibit systematic drift. This is different from natural speech where μ is estimated to be 0.4 [44]. Compared to [44], the value of λ for singing voice is also smaller (0.7 vs. 2.4), which indicates that the distribution is more narrow. The transition behavior within Ω_2 can be described as $p(\Delta_1)p(\Delta_2)$ by assuming the two pitch contours evolve independently. Δ_i is the change of pitch periods in two consecutive frames of the i^{th} pitch contour. The transition probability between different pitch subspaces is determined by examining the pitch contours of singing voice and the pitch contours of the dominant sound in the accompaniment. The later one can be obtained by inspecting the spectrogram of the accompaniment. Table I shows the estimated transition probability between different pitch subspaces using the training data.

The Viterbi algorithm is used to decode the most likely sequence of pitch hypotheses. If a pitch hypothesis in the optimal sequence contains two pitches, the first pitch is considered as the pitch of singing voice. This is because, as mentioned before, the first pitch is the dominant one in our formulation.

C. SINGING VOICE SEPARATION

For our separation task, as mentioned earlier, we apply the speech-separation algorithm of Hu and Wang [18], originally proposed to separate voiced speech from interference based on pitch tracking and amplitude modulation.

To apply this algorithm, we first decomposes the input into T-F units. A T-F unit corresponds to a time frame and a frequency channel. Within each T-F unit, we extract the following features: autocorrelation of a filter response, autocorrelation of the envelope of a filter response, cross-channel correlation, and cross-correlation of envelopes. Then contiguous T-F units are merged into segments if their energy and cross-channel correlation are both high. The Hu–Wang system uses an iterative method to estimate the pitch contour of the target signal. Since pitch contours have been obtained in the second stage, we supply our detected pitch contours to label each T-F unit as singing dominant or accompaniment dominant. Additional segments are formed for those T-F units if they are labeled as singing dominant but do not belong to any segments generated before and they have high cross-channel correlation of envelopes. This completes the segmentation step. The grouping step simply groups the segments where the majority of T-F units are labeled as singing dominant to form the foreground stream, which corresponds to the singing voice. Our system finally outputs separated singing voice resynthesized from the segments in the foreground stream. For details of the Hu–Wang algorithm, see [18], and for program code see http://www.cse.ohio-state.edu/pnl/software.html.

IV. EVALUATION AND COMPARISON

Systematic evaluation is important for gauging the performance of a sound separation system. Although several common databases currently exist for speech separation, there is none for singing voice separation. The difficulty of constructing such a database mainly lies in getting separately recorded singing voice and music accompaniment. In modern studios, singing voice and accompaniment are usually recorded separately and then mixed together. However such separate recordings are not accessible due to copyright issues. On the other hand some modern commercial karaoke compact disks (CDs) are recorded with multiplex technology in which singing voice and accompaniment are multiplexed and stored in a single file. With proper de-multiplexing software, separate singing voice and accompaniment can be extracted. We extracted 10 songs from karaoke CDs obtained from [1] to construct a database for singing voice detection. These songs are sampled at 16 kHz with 16 bit resolution. Among these 10 songs, 5 are rock music and the other 5 are country music. Clips are extracted to form another database for singing voice pitch detection and separation. We refer to the energy ratio of singing voice to accompaniment as signal to noise ratio (SNR) as in speech separation studies. In the following subsections, we evaluate the performance of each stage as well as the performance of the whole separation system.

A. SINGING VOICE DETECTION

With separate singing voice available, vocal and non-vocal portions can be easily labeled for training and testing purposes. We apply a simple energy-based silence detector on clean singing voice signals to distinguish vocal portions from non-vocal portions. Few systems developed for singing voice detection consider the effect of SNRs on classification. We found that a classifier trained at one SNR often performs poorly when tested at another SNR because of the mismatch between training and testing. Nwe et al. [31] pointed out that different sections of a song (intro, verse, chorus, bridge, and outro) have different SNRs and a singing voice detector needs to handle different sections properly. To address this problem, we train a classifier with samples mixed in different SNRs. In this way the classifier is trained over a range of SNRs. Specifically, we mix the singing voice track and the accompaniment track of each song at SNRs of 10 dB and 0 dB and then use the mixtures to train the classifier. As mentioned in Section III-A, we choose the MFCC as the feature vector and the GMM as the classifier. A 13-dimensional MFCC feature vector is calculated for each frame of 16 ms with a frame shift of 10 ms using the auditory toolbox by Slaney [34]. A Gaussian mixture model with 4 components, each having a diagonal covariance matrix, is used to model the MFCC distribution of the two classes: c_v and c_{nv}. The parameters of the GMMs are estimated using the toolbox by Murphy [29].

For spectral change detection, we calculate the spectrogram with a frame size 16 ms and a frame shift of 10 ms, the same as the calculation for MFCC. After the distance between two consecutive frames is calculated for all frames, we apply a 11-tap symmetric median filter $(H = 5)$. The weighting factor C_t is set to 1.5 which allows spectral contents to fluctuate to some extent. Smaller values of H and C_t can also be used as they result in more spectrally homogeneous partitions. To avoid over-partitioning, the minimum spectral change interval T_m is set to be 100 ms, which is found to yield the best result.

Fig. 2 shows the classification result for a clip of rock music. The clean singing voice is shown in Fig. 2(a) and in Fig. 2(b) it is mixed with music accompaniment to give an overall SNR of 0 dB. The thick line above the waveform in Fig. 2(a) shows the vocal portions obtained from silence detection. In Fig. 2(c), the spectrogram of the mixture is plotted. The vertical lines in Fig. 2(d) show the instances of significant spectral changes identified by our spectral change detector. The input is over-partitioned to some extent, but the beat times and the time instances when the singing voice enters are well captured except at times around 0.7 and 1.1 seconds. Fig. 2(e) gives the

TABLE II
CLASSIFICATION ACCURACY FOR DIFFERENT METHODS (% FRAMES)

	-5 dB	0 dB	5 dB	10 dB
proposed method	**80.3**	**85.0**	**90.2**	**91.1**
frame-level classification	71.3	77.4	81.7	83.8
HMM	79.0	83.5	87.5	88.8
spectral change detection + majority vote	80.4	84.5	89.1	90.2

result of frame-level classification, i.e., a frame is classified as vocal (indicated as a high value) if its likelihood of c_v is greater than that of c_{nv}, and vice versa. As can be seen, frame-level classification is not very reliable. Fig. 2(f) shows the final classification, which matches the reference labeling indicated in Fig. 2(a) well except at around 1.5 seconds for a very short nonvocal portion. Many frames around 2.7 seconds are misclassified as vocal in the frame-level classification but are correctly classified as non-vocal as shown in Fig. 2(f).

We perform 10-fold cross validation to access the overall performance of the proposed detection method. Each time 90% of the data is used for training and the rest is used for testing. This process is repeated 10 times and the average of classification accuracy (percentage of frames) is taken as the performance of the method. The total amount of data for training and testing is about 30 minutes. As shown in the last row of Table II, the classification accuracies of the proposed method are 80.3%, 85.0%, 90.2% and 91.1% for -5, 0, 5 and 10 dB, respectively. Also the performance decreases gradually as the SNR decreases.

For comparison purposes, the classification accuracies of three other different methods are also obtained as shown in Table II. Frame-level classification is as described before. The HMM method is similar to the one used in [2]. Each class is modeled as a one-state HMM using the trained GMM as the observation distribution. The exiting probability from the state is the inverse of the average duration of portions of each class. Another reasonable method combines the spectral change detection and a majority vote to determine the labeling of a portion. By a majority vote we mean that if the majority of frames of a portion is classified as vocal the portion is classified as vocal, and vice versa.

The HMM method imposes a temporal continuity constraint and it outperforms frame-level classification by at least 5.0% in each case. The majority vote scheme gives similar results as the proposed method that is based on overall likelihood. Besides the temporal continuity constraint, these two methods also impose that a class change can only occur at times when spectral contents change. This additional constraint makes the proposed method outperform the HMM method by 1.3%, 1.5%, 2.7% and 2.3% for -5, 0, 5, and 10 dB cases, respectively.

B. PREDOMINANT PITCH DETECTION

In order to evaluate the applicability of the proposed system to a wide range of polyphonic audio for singing voice detection and separation, we further extract a total of 25 clips from the 10 songs used in the singing voice detection. The average length of each clip is 3.9 seconds and the total length of all the clips is 97.5 seconds. The clips include both male and female singers. In some clips, singing voice is present all the time; in some other clips, singing voice is present either at the beginning, the middle or the end of a clip. For each clip, the singing voice and the accompaniment are mixed at 4 different SNRs: -5, 0, 5, and 10 dB. The variety in the testing database is designed to better access the proposed system.

Since separate singing voice tracks are available, accurate reference pitch contours can be determined. The reference pitch contours are calculated using Praat [4], which is a standard system of pitch detection for clean signals. The clean singing voice is processed by Praat and the detected pitch contour is visually inspected to correct obvious pitch halving and doubling errors.

Fig. 3 shows the result of the pitch detection for the same clip in Fig. 2(b). The clip is partitioned into vocal and non-vocal portions by the first stage. The cochleagram of the clip is shown in Fig. 3(a). Unlike the spectrogram

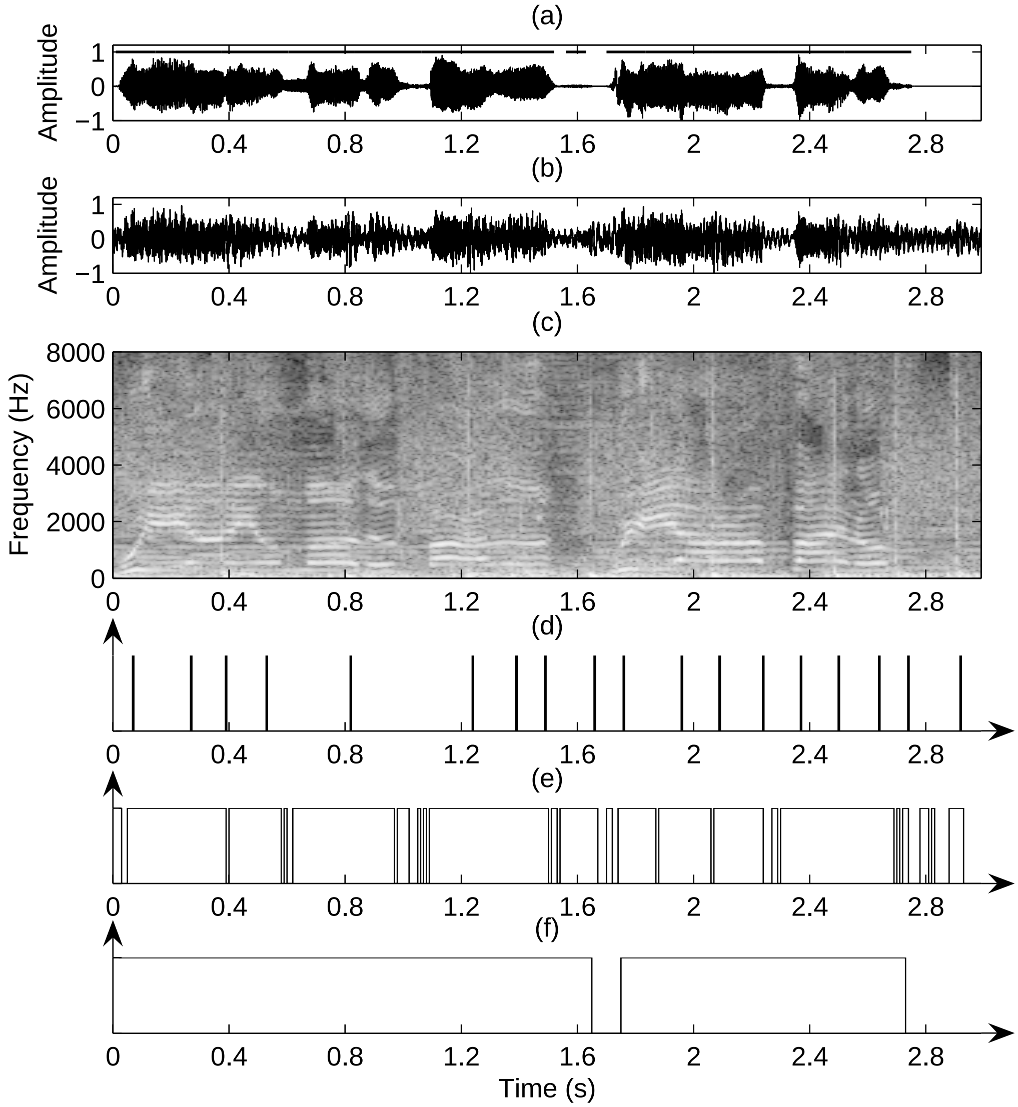

Fig. 2. Singing voice detection for a clip of rock music. (a) The waveform of the singing voice signal. The thick lines above the waveform indicate vocal portions. (b) The mixture of the singing voice and the accompaniment in 0 dB SNR. (c) The spectrogram of the mixture. Brighter area indicates stronger energy. The vertical lines in (d) indicate the spectral change moments identified by the spectral change detector. (e) The frame-level classification of the clip. A high value indicates the frame is classified as vocal and a low value as non-vocal. (f) The final classification using the spectral change detection and the overall likelihood.

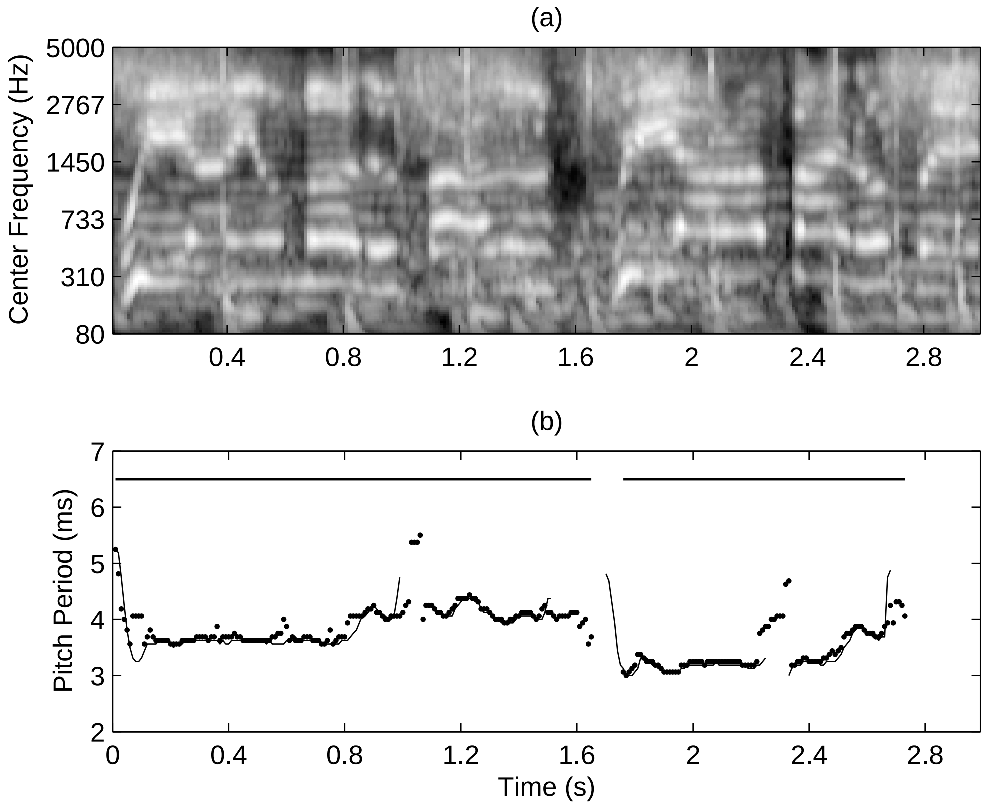

Fig. 3. Predominant pitch detection on the clip of rock music. (a) Cochleagram of the clip. Brighter area indicates stronger energy. The vertical axis shows the center frequencies of frequency channels. (b) Results of pitch detection. The thin solid lines indicate the reference pitch contours and the dots represent the detected pitches. The thick lines at the top indicate the detected vocal portions.

as in Fig. 2(c), the cochleagram is an auditory spectrogram of a signal with a quasi-logarithmic frequency scale similar to that of the human auditory system. In this case, the cochleagram is calculated over 20-ms time frames with 10 ms frame shift and a gammatone filterbank with 128 frequency channels whose center frequencies are distributed on the ERB scale. It can be seen that the singing voice is dominant in high frequency channels while the low frequency channels are severely corrupted by the accompaniment. The predominant pitch detection algorithm is applied to the detected vocal portions. In Fig. 3(b) the detected pitches are plotted as dots against the reference pitch contours which are plotted as solid lines. In this example, the detected pitches well match the reference most of the time. For unvoiced singing, such as the portion from 2.2 to 2.3 seconds, the pitch detector gives pitches belonging to some other source. The thick lines in Fig. 3(b) indicate the detected vocal portions.

Since pitch detection depends on classification, we consider three cases to evaluate different aspects of the predominant pitch detection stage:

1) No classification: no classification is used and the predominant pitch detector is applied to the whole clip. The results in this case should demonstrate the value of singing voice detection.
2) Ideal classification: the reference classification is used and the predominant pitch detector is applied to vocal portions only. This evaluates the performance of the pitch detector alone.
3) Actual classification: the classification obtained in the first stage is used and the predominant pitch detector

TABLE III
PREDOMINANT PITCH DETECTION ERROR RATES WITH NO CLASSIFICATION (%)

	-5 dB	0 dB	5 dB	10 dB
Proposed	**56.9**	**50.0**	**46.0**	**45.2**
Correlogram	88.4	85.2	81.8	78.3
Klapuri	70.1	60.3	53.6	50.8
Wu et al.	69.7	57.8	50.8	45.9

TABLE IV
PREDOMINANT PITCH DETECTION ERROR RATES WITH IDEAL CLASSIFICATION (%)

	-5 dB	0 dB	5 dB	10 dB
Proposed	**18.5**	**11.5**	**7.6**	**6.8**
Correlogram	49.7	46.6	43.1	39.6
Klapuri	31.7	21.9	15.2	12.4
Wu et al.	31.2	19.4	12.3	7.5

is applied to the detected vocal portions. This gives the combined result for the first two stages.

For comparison purposes, we implemented a multipitch detection algorithm designed for musical applications by Klapuri [21]. To ensure the quality of our implementation, we consulted the author and obtained comparable results when our implemented algorithm was tested under similar conditions as in [21]. To perform predominant pitch detection using Klapuri's algorithm, we first calculate the most dominant pitch in a short window (20 ms). Then we impose the continuity constraint of a pitch contour by applying median filtering to the resulting pitch contour. This simple technique brings the gross error rate down by 4% for Klapuri's algorithm; a gross error occurs if the detected pitch exceeds 10% of the reference pitch in frequency. Tables III–V show the performance of different pitch detection algorithms in the 3 cases. The numbers in the tables represent the gross error rates at the frame level. The correlogram algorithm for predominant pitch detection has been used in several studies [42], [33] and is used in [18] to get the initial pitch estimation. The performance of the original algorithm by Wu et al. [44] is also listed. Since the pitch range for the types of singing examined in this study is relatively small compared to that of operatic singing, we set 80–500 Hz as the plausible pitch range for all algorithms.

As can be seen in Table III, when no classification is applied, a large number of detected pitch points do not match reference pitches even when SNR is high. This is true of all the algorithms since they give a pitch estimate whether or not singing is present.

Table IV shows the error rates with the ideal classification. The proposed algorithm has a gross error rate of 18.5%, 11.5%, 7.6%, and 6.8% for the -5, 0, 5, and 10 dB cases, respectively. It also can be seen that the performance of the proposed algorithm degrades smoothly as the SNR decreases. The correlogram algorithm performs considerably worse. This indicates that the correlogram alone is not suitable for pitch detection of singing voice when accompaniment is present. Klapuri's algorithm performs reasonably well considering the nature of polyphonic audio. But compared to the proposed algorithm, the error rates of Klapuri's algorithm are 13.2%, 10.4%, 7.6%, and 5.6% higher for the -5, 0, 5, and 10 dB cases, respectively. The Wu et al. algorithm has comparable performance with ours in the 10 dB case. But as SNR decreases, the error rate grows significantly faster in their algorithm. This shows the effectiveness of the new channel/peak selection method introduced in this paper.

TABLE V
PREDOMINANT PITCH DETECTION ERROR RATES WITH ACTUAL CLASSIFICATION (%)

	-5 dB	0 dB	5 dB	10 dB
Proposed	**44.2**	**31.7**	**24.3**	**21.6**
Correlogram	72.4	65.7	59.6	53.9
Klapuri	55.5	41.7	31.7	26.5
Wu et al.	55.1	39.0	29.0	22.4

When actual classification is applied, shown in Table V, for the proposed method the errors introduced by wrong classification can be approximately obtained by calculating the difference of error rates under the ideal classification and actual classification condition. They are 25.7%, 20.2%, 16.7%, and 14.8% for the -5, 0, 5, and 10 dB cases, respectively, showing a similar pattern as in Table II. Although the overall error rates are higher, they are still significantly lower than those when no classification is applied. We also find by inspecting the errors that many errors occur at boundaries of the two classes where singing voice is either weak or cannot be heard at all. Due to classification errors, the gross error rates of other three algorithms increase similarly. As Table V shows, the performance of other three algorithms is inferior to that of the proposed one.

C. SINGING VOICE SEPARATION

As mentioned in Section II, few systems are devoted to singing voice detection. As a result no criterion has been established for evaluating the separation of singing voice. A fundamental question related to evaluation criteria is what the computational goal of a singing voice separation system should be. The Hu–Wang system [18] uses a notion called ideal binary mask to quantify the computational goal. The ideal binary mask is defined as follows: a T-F unit in the mask is assigned 1 if the energy of the target source in the unit is stronger than that of the total interference, and 0 otherwise. This notion is grounded on the well-established auditory masking phenomenon [28]. Human speech intelligibility experiments show that target speech reconstructed from the ideal binary mask gives high intelligibility scores, even in very low SNR conditions [32], [9], [8]. More discussion of the ideal binary mask as the computational goal of CASA can be found in [41].

For musical applications, the perceptual quality of the separated sound is emphasized in some cases. However, perceptual quality is not well defined and hard to quantify. Our informal listening experiments show that the quality of singing voice reconstructed from the ideal binary mask is close to the original one when SNR is high and it degrades gradually with decreasing SNR. Consistent with speech separation, we suggest to use the ideal binary mask as the computational goal for singing voice separation.

To quantify the performance of the system, we then calculate the SNR before and after the separation using the singing voice resynthesized from the ideal binary mask as the ground truth [18]:

$$SNR = 10 \log 10 [\frac{\sum_n I^2(n)}{\sum_n (I(n) - O(n))^2}] \tag{14}$$

$I(n)$ is the resynthesized singing voice from the ideal binary mask, which can be obtained from the premixing singing voice and accompaniment. In calculating the SNR after separation, $O(n)$ is the output of the separation system. In calculating the SNR before separation, $O(n)$ is the mixture resynthesized from an all-one mask, which compensates for the distortion introduced in the resynthesis.

Fig. 4 shows a separation example of the same clip used in Fig. 2 and Fig. 3. Fig. 4(a) is the clean singing voice resynthesized from the all-one mask. Fig. 4(b) is the mixture resynthesized from the all-one mask. Fig. 4(c) shows the resynthesized waveform from the ideal binary mask and Fig. 4(d) is the output of our separation system. As can be seen, the output waveform well matches that from the ideal mask. It also matches the original signal shown in Fig. 4(a) well.

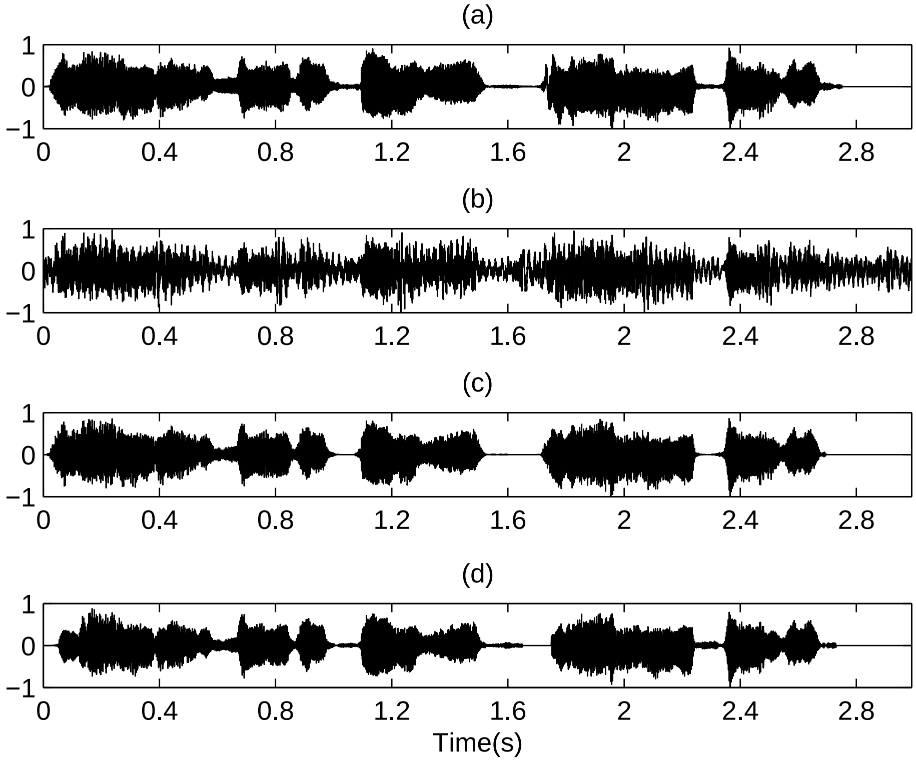

Fig. 4. Waveform comparison. (a) The singing voice. (b) The mixture. (c) The ground truth resynthesized from the ideal binary mask. (d) The output of the proposed separation system. The vertical axis in each plot indicates the amplitude of the waveform.

Fig. 5(a) and (b) show the ideal binary mask and the mask estimated by the separation system, respectively. It is clear that the estimated mask is similar to the ideal mask and retains most of the energy of the singing voice. Fig. 5(c) and (d) plot the cochleagram of the mixture masked by the ideal mask and the estimated mask, respectively.

Since separation depends on classification and pitch detection, we consider three cases in the evaluation, each characterizing a different aspect of the system:

1) Ideal pitch: the reference pitch contour is used for separation. This gives the ceiling performance of the separation system.
2) Ideal classification with pitch detection: use the reference classification but use detected pitch for separation. This isolates the classification stage and gives the performance of the last two stages.
3) Actual classification with pitch detection: this gives the performance of the whole system.

Fig. 6 shows the SNR gains after separation of the proposed system for the 3 cases. When the ideal pitch contour is given (as shown by the Case 1 line), the SNR gains for low SNRs, e.g., -5 dB and 0 dB, are significant. However for the SNR of 10 dB the gain is relatively small. One reason is that in some cases the pitches of singing voice may change rapidly. When the pitches change fast, the separation stage does not group properly. Another reason is the presence of unvoiced consonants. Unvoiced constants cannot be recovered by the pitch-based separation algorithm. Also the Hu–Wang system gives only an estimate of the ideal binary mask, and it makes certain errors in grouping segments belonging to the singing voice. When the original SNR is high, the accompaniment is weak and the energy

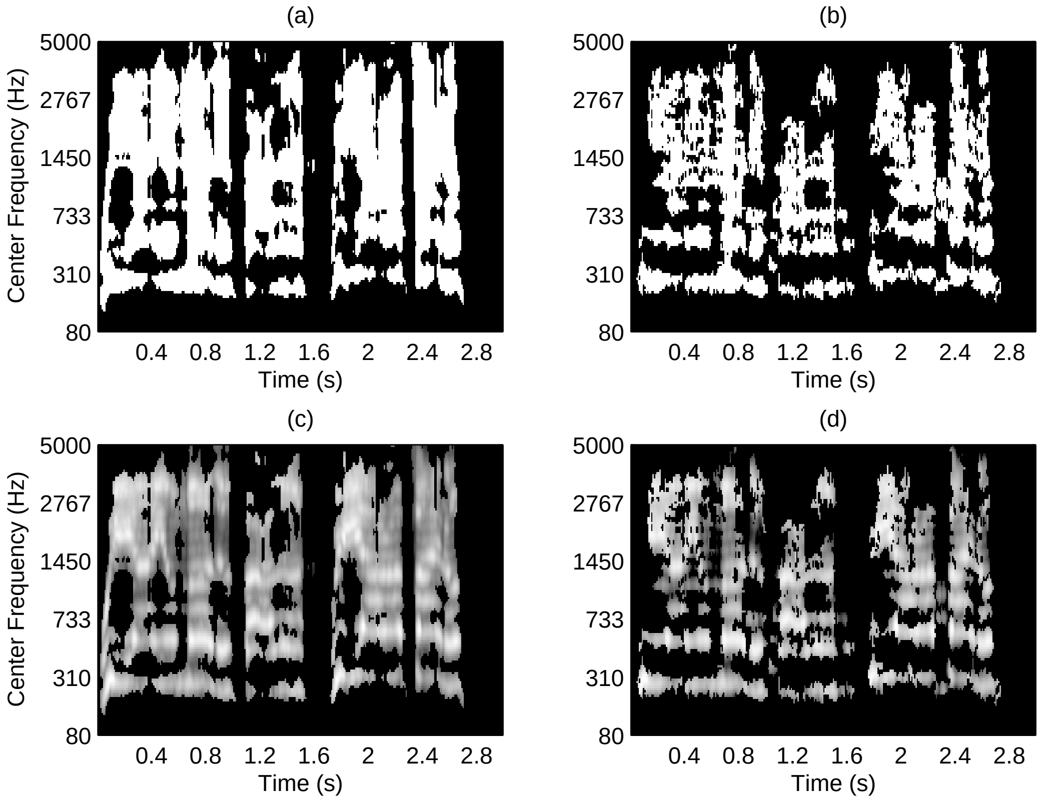

Fig. 5. Mask comparison. (a) Ideal binary mask obtained from the premixed singing voice and accompaniment. White pixels indicate 1 and black pixels indicate 0. (b) The mask of singing voice estimated by the separation system. (c) The cochleagram of the mixture masked by the ideal binary mask. (d) The cochleagram of the mixture masked by the estimated mask.

loss of the singing voice may be comparable to the rejected energy of the accompaniment rejected. As the SNR decreases, the accompaniment becomes stronger and the energy loss of the singing energy becomes less compared to the rejected energy of the accompaniment. Therefore the separation stage works better when the original SNR is lower. For example, the system achieves an SNR gain of 11.4 dB for the input SNR of -5 dB. We note that for many applications, such as those mentioned in Section I, singing separation is particularly needed for low SNR situations.

The use of the pitch detection algorithm given the ideal classification is subject to pitch detection errors. Erroneous pitch estimates make some segments group incorrectly. As a result, the overall performance (Case 2 in Fig. 6) is worse than that with ideal pitch contours. For the SNR of 10 dB, the SNR after separation is even slightly lower than that of the original mixture. However, as the SNR decreases, the SNR after separation is consistently higher. When the classification stage is also included, i.e., the entire system is evaluated, the SNR gains (Case 3 in Fig. 6) are slightly lower than those in the second case. Although the SNR after separation for the 10 dB case is not improved, the system achieves SNR improvements of 7.1, 5.5, and 3.7 dB for the input SNR of -5 dB, 0 dB, and 5 dB, respectively. This demonstrates that the proposed method works well for low SNR situations.

For Case 3 where both actual classification and pitch detection are used, we compare the proposed separation method with a standard comb filtering method [11], which extracts the spectral components at the multiples of a

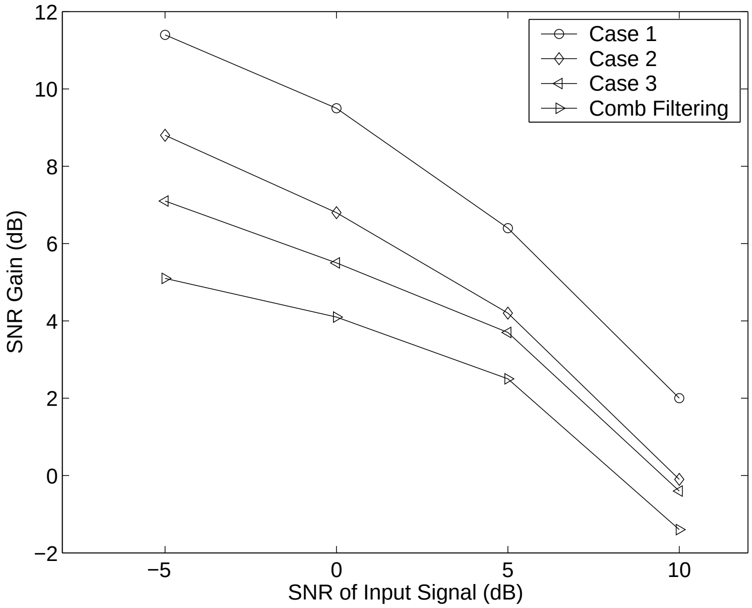

Fig. 6. SNR gain comparison. The SNR gains for the three cases are displayed (See Section IV-C). Also displayed are the SNR gains of a comb filtering method.

given pitch. The comb filter implemented has 3 coefficients and is applied to every frame. After comb-filtering, the obtained signal is passed through an all-one mask. This step makes the comb-filtered signal comparable to the resynthesized sinal from a binary mask. The performance of the comb filtering method, shown in Fig. 6, is consistently worse than that of our approach. For example, the SNR gain is 1 dB lower in the 10 dB case and 2 dB lower in the -5 dB case. The worse performance is mainly caused by the fact that the comb filter passes all frequency components close to the multiples of a given pitch, which include those belonging to music accompaniment.

The classification stage alone is expected to contribute to the SNR gain by rejecting the energy from accompaniment. To quantify this contribution, we calculate the SNR gains resulted from classification alone. More specifically, after classification, the vocal portions of the input are retained while the non-vocal portions are rejected. The retained signal is used in (14) for the SNR calculation. The SNR gains from the classification stage alone are 1.4, 1.0, 1.1, and 0.2 dB for -5, 0, 5, and 10 dB cases, respectively. Therefore, except for the 10 dB case, the contribution of the classification stage to the overall SNR gain is small. In other words, the overall system is responsible for the performance improvements.

We have also directly applied the original Hu–Wang system to the vocal portions obtained from the first stage. In this case, the pitch contour of singing voice is iteratively refined starting from the estimates obtained from the correlogram pitch detection algorithm. It is found that the resulted SNR gains are lower. This indicates that the proposed predominant pitch detection stage is important for the performance of the overall system.

V. DISCUSSION AND CONCLUSION

As mentioned in Section I, few systems have been proposed for singing voice separation. Our system represents the first general framework for singing voice separation. This system is also extensible. Currently, we use pitch as the only organizational cue. Other ASA cues, such as onset/offset and common frequency modulation, can also be incorporated into our system, which would be able to separate not only voiced singing but also unvoiced singing.

Another important aspect of the proposed system is its adaptability to different genres. Currently our system is genre independent, i.e., rock music and country music are treated in the same way. This, in a sense, is a strength of the proposed system. However, considering the vast variety of music, a genre dependent system may achieve better performance. Given the genre information, the system can be adapted to the specific genre. The detection stage can be retrained using genre-specific samples. The observation probability and the transition probability of the HMM in the pitch detection stage are also retrainable. The genre information can be obtained from the metadata of a musical file or by automatic genre classification [37].

Our classification stage is based on MFCC features. Recently long-term features, such as modulation spectrum [10], have been used with some success in related tasks such as speech/music classification. We have attempted to incorporate the modulation spectrum into the first stage but the overall classification accuracy is not improved. It seems that the modulation spectrum of vocal and non-vocal segments does not have enough discrimination power to produce further improvement.

Our pitch detection system uses an auditory front-end for frequency decomposition and an autocorrelation function for pitch detection. One problem with this autocorrelation-based pitch detection approach is that the frequency resolution in the high-frequency range is limited. As a result the proposed system cannot be used to separate high-pitched singing voice, as encountered in operatic singing. However, most types of singing, such as in pop, rock, and country music, have a smaller pitch range and therefore this system can potentially be applied to a wide range of problems.

VI. ACKNOWLEDGMENT

This research was supported in part by an AFOSR grant (F49620-04-1-0027) and an AFRL grant (FA8750-04-1-0093). We thank G. Hu for many useful discussions.

REFERENCES

[1] TOP tunes karaoke. [Online]. Available: http://www.toptuneskaraoke.com

[2] A. L. Berenzweig and D. P. W. Ellis, "Locating singing voice segments within music signals," in *Proc. IEEE WASPAA*, 2001, pp. 119–122.

[3] A. L. Berenzweig, D. P. W. Ellis, and S. Lawrence, "Using voice segments to improve artist classification of music," in *Proceedings of AES 22nd International Conference on Virtual, Synthetic and Entertainment Audio*, 2002.

[4] P. Boersma and D. Weenink. (2002) Praat: Doing phonetics by computer, version 4.0.26. [Online]. Available: http://www.fon.hum.uva.nl/praat

[5] D. Z. Borch and J. Sundberg, "Spectral distribution of solo voice and accompaniment in pop music," *Logopedics Phoniatrics Vocology*, vol. 27, pp. 37–41, 2002.

[6] A. S. Bregman, *Auditory Scene Analysis*. Cambridge, MA: MIT Press, 1990.

[7] G. J. Brown and D. L. Wang, "Separation of speech by computational auditory scene analysis," in *Speech Enhancement*, J. Benesty, S. Makino, and J. Chen, Eds. New York: Springer, 2005, pp. 371–402.

[8] D. Brungart, P. S. Chang, B. D. Simpson, and D. L. Wang, "Isolating the energetic component of speech-on-speech masking with an ideal binary time-frequency mask," 2005, submitted.

[9] P. S. Chang, "Exploration of behavioral, physiological, and computational approaches to auditory scene analysis," Master's thesis, The Ohio State University, Department of Computer Science and Engineering, 2004. [Online]. Available: http://www.cse.ohio-state.edu/pnl/theses.html

[10] W. Chou and L. Gu, "Robust singing detection in speech/music discriminator design," in *Proc. IEEE ICASSP*, 2001, pp. 865–868.

[11] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*. New York: Macmillan, 1993.

[12] P. Divenyi, Ed., *Speech Separation by Humans and Machines*. Norwell MA: Kluwer Academic, 2005.

[13] C. Duxbury, J. P. Bello, M. Davies, and M. Sandler, "Complex domain onset detection for musical signals," in *Proc. of the 6th Conference on Digital Audio Effect (DAFx-03)*, London, U.K., 2003.

[14] D. Godsmark and G. J. Brown, "A blackborad architecture for computational auditory scene analysis," *Speech Communication*, vol. 27, no. 4, pp. 351–366, 1999.

[15] M. Goto, "A real-time music-scene-description system: Predominant-F0 estimation for detecting melody and bass lines in real-world audio signals," *Speech Communication*, vol. 43, no. 4, pp. 311–329, 2004.

[16] D. J. Hand and K. Yu, “Idiot’s Bayes – Not so stupid after all?” *Int. Statist. Rev.*, vol. 69, no. 3, pp. 385–398, 2001.

[17] H. Helmholtz, *On the Sensations of Tone*, (A. J. Ellis, Trans., New York: Dover, 1954), Braunschweig, Germany: Vieweg & Son, 1863.

[18] G. Hu and D. L. Wang, “Monaural speech segregation based on pitch tracking and amplitude modulation,” *IEEE Trans. Neural Networks*, vol. 15, pp. 1135–1150, 2004.

[19] Y. E. Kim and B. Whitman, “Singer identification in popular music recordings using voice coding features,” in *Proc. International Symposium on Music Information Retrieval*, 2002.

[20] Y. E. Kim, “Singing voice analysis/synthesis,” Ph.D. dissertation, MIT, Media Lab, 2003.

[21] A. Klapuri, “Multiple fundamental frequency estimation based on harmonicity and spectral smoothness,” *IEEE Trans. Speech Audio Processing*, vol. 11, pp. 204–816, 2003.

[22] G. Kovačić, P. Boersman, and H. Domitrović, “Long-term average spectra in professional folk singing voices: A comparison of the klapa and dozivačcki styles,” in *Proceedings 25, Institute of Phonetic Sciences*, 2003, pp. 53–64.

[23] D. Li, I. K. Sethi, N. Dimitrova, and T. McGee, “Classification of general audio data for content-based retrieval,” *Pattern Recognition Letters*, vol. 22, pp. 533–544, 2002.

[24] Y. Li and D. L. Wang, “Detecting pitch of singing voice in polyphonic audio,” in *Proc. IEEE ICASSP*, vol. 3, 2005, pp. 17–20.

[25] N. C. Maddage, C. Xu, and Y. Wang, “A SVM-based classification approach to musical audio,” in *Proc. ISMIR*, 2003.

[26] D. K. Mellinger, “Event formation and separation in musical sound,” Ph.D. dissertation, Stanford University, Department of Computer Science, 1991.

[27] Y. Meron and K. Hirose, “Separation of singing and piano sounds,” in *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP 98)*, 1998.

[28] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, 5th ed. London, U.K.: Academic Press, 2003.

[29] K. Murphy. (2005, June) HMM toolbox for MATLAB. [Online]. Available: http://www.cs.ubc.ca/˜murphyk/Software/HMM/hmm.html

[30] S. Nooteboom, “The prosody of speech: melody and rhythm,” in *The Handbook of Phonetic Science*, W. J. Hardcastle and J. Laver, Eds. Oxford, U.K.: Blackwell, 1997, pp. 640–673.

[31] T. L. Nwe, A. Shenoy, and Y. Wang, “Singing voice detection in popular music,” in *Proc. of the 12th Annual ACM International Conference on Multimedia*, 2004, pp. 324–327.

[32] N. Roman, D. L. Wang, and G. J. Brown, “Speech segregation based on sound localization,” *Journal of the Acoustical Society of America*, vol. 114, no. 4, pp. 2236–2252, 2003.

[33] S. K. Shandilya and P. Rao, “Retrieving pitch of the singing voice in polyphonic audio,” in *Proc. of National Conference on Communications*, IIT Madras, India, 2003.

[34] M. Slaney. (1999, Jan.) Auditory toolbox for MATLAB. [Online]. Available: http://rvl4.ecn.purdue.edu/˜malcolm/interval/1998-010/

[35] J. Sundberg, “The acoustics of the singing voice,” *Scientific American*, pp. 82–91, Mar. 1977.

[36] ——, “Perception of singing,” in *Psychology of Music*, 2nd ed., D. Deutsch, Ed. Academci Press, 1999, pp. 171–214.

[37] G. Tzanetakis and P. Cook, “Musical genre classification of audio signals,” *IEEE Trans. Speech Audio Processing*, vol. 10, pp. 293–302, 2002.

[38] A. L.-C. Wang, “Instantaneous and frequency-warped signal processing techniques for auditory source seperation,” Ph.D. dissertation, Stanford University, Department of Electrical Engineering, 1994.

[39] C. K. Wang, R. Y. Lyu, and Y. C. Chiang, “An automatic singing transcription system with multilingual singing lyric recognizer and robust melody tracker,” in *Proc. EUROSPEECH*, 2003.

[40] D. L. Wang, “Feature-based speech segregation,” in *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, D. L. Wang and G. J. Brown, Eds. New York: IEEE Press (dual imprint with Wiley), 2006, to appear.

[41] ——, “On ideal binary mask as the computational goal of auditory scene analysis,” in *Speech Separation by Humans and Machines*, P. Divenyi, Ed. Boston, MA: Kluwer Academic, 2005, pp. 181–197.

[42] D. L. Wang and G. J. Brown, “Separation of speech from interfering sounds based on oscillatory correlation,” *IEEE Trans. Neural Networks*, vol. 10, pp. 684–697, 1999.

[43] Y. Wang, M.-Y. Kan, T. L. Nwe, A. Shenoy, and J. Yin, “LyricAlly: automatic synchronization of acoustic musical signals and textual lyrics,” in *Proceedings of the 12th Annual ACM International Conference on Multimedia*. New York, NY, USA: ACM Press, 2004, pp. 212–219.

[44] M. Wu, D. L. Wang, and G. J. Brown, “A multipitch tracking algorithm for noisy speech,” *IEEE Trans. Speech Audio Processing*, vol. 11, pp. 229–241, 2003.

[45] T. Zhang, “System and method for automatic singer identification,” HP Laboratories, Tech. Rep. HPL-2003-8, 2003.