A Tandem Algorithm for Singing Pitch Extraction and Voice Separation From Music Accompaniment

Chao-Ling Hsu, DeLiang Wang, *Fellow, IEEE*, Jyh-Shing Roger Jang, and Ke Hu, *Student Member, IEEE*

***Abstract*—Singing pitch estimation and singing voice separation are challenging due to the presence of music accompaniments that are often nonstationary and harmonic. Inspired by computational auditory scene analysis (CASA), this paper investigates a tandem algorithm that estimates the singing pitch and separates the singing voice jointly and iteratively. Rough pitches are first estimated and then used to separate the target singer by considering harmonicity and temporal continuity. The separated singing voice and estimated pitches are used to improve each other iteratively. To enhance the performance of the tandem algorithm for dealing with musical recordings, we propose a trend estimation algorithm to detect the pitch ranges of a singing voice in each time frame. The detected trend substantially reduces the difficulty of singing pitch detection by removing a large number of wrong pitch candidates either produced by musical instruments or the overtones of the singing voice. Systematic evaluation shows that the tandem algorithm outperforms previous systems for pitch extraction and singing voice separation.**

***Index Terms*—Computational auditory scene analysis (CASA), iterative procedure, pitch extraction, singing voice separation, tandem algorithm.**

I. Introduction

WHILE separating target voice from a monaural mixture of different sound sources appears effortless for the human auditory system, it is very difficult for machines and has been extensively studied for decades. In particular, separating singing voice from music accompaniment remains a major challenge.

Singing voice separation is, in a sense, a special case of speech separation and has many similar applications. For example, automatic speech recognition corresponds to automatic lyrics recognition [24], automatic speaker identification to automatic singer identification [30], and automatic subtitle alignment which aligns speech and subtitle to automatic lyric alignment [28] which can be used in a karaoke system. These applications also encounter similar problems. They perform substantially worse in the presence of background noise or music accompaniment. Compared to speech separation, separation of singing voice could be simpler with less pitch variation. On the other hand, there are several major differences. For speech separation, or the cocktail party problem, the goal is to separate the target speech from various types of background noise which can be broadband or narrowband, periodic or aperiodic. In addition, the background noise is independent of speech in most cases so that their spectral contents are uncorrelated. For singing voice separation, the goal is to separate singing voice from music accompaniments which in most cases are broadband, periodic, and strongly correlated to the singing voice. Furthermore, the upper pitch boundary of singing can be as high as 1400 Hz for soprano singers [29] while the pitch range of normal speech is between 80 and 500 Hz. These differences make the separation of singing voice and music accompaniment potentially more challenging.

For monaural singing voice separation, existing methods can be generally classified into three categories depending on their underlying methodologies: spectrogram factorization (e.g., [19], [22]), model-based methods (e.g., [16], [19]), and pitch-based methods (e.g., [13], [23]). Spectrogram factorization methods utilize the redundancy of the singing voice and music accompaniment by decomposing the input signal into a pool of repetitive components. Each component is then assigned to a sound source. Model-based methods learn a set of spectra from music accompaniment only segments. Spectra of the vocal signal are then learned from the sound mixture by fixing accompaniment spectra. Pitch-based methods use extracted vocal pitch contours as the cue to separate the harmonic structure of the singing voice.

These methods have their limitations. Spectrogram factorization methods encounter difficulties in assigning repetitive components or bases to the corresponding sound sources. The performance drops significantly when the number of musical instruments increases. Furthermore, it is difficult to separate singing voice from a short mixture since vocal signals typically have more diverse spectra than that of each instrument [23]. Model-based methods require a considerable amount of music accompaniment only segments so that they can model the characteristics of the background music.

Compared to spectrogram factorization and model-based methods, pitch-based methods potentially have fewer limitations. The only required cue is the pitch contours of the singing voice which can be extracted from a very short mixture and

Manuscript received December 01, 2010; revised May 23, 2011; accepted December 17, 2011. Date of publication January 02, 2012; date of current version March 02, 2012. Part of the work was conducted while C.-L. Hsu was visiting The Ohio State University. This work was supported in part by the National Science Council, Taiwan, under Grant NSC 96-2628-E-007-141-MY3 and in part by the by the Air Force Office of Scientific Research (AFOSR) under Grant FA9550-08-1-0155. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Tomohiro Nakatani.

C.-L. Hsu is with Mediatek, Inc., Hsinchu 30078, Taiwan (e-mail: leon@mirlab.org).

D. L. Wang and K. Hu are with the Department of Computer Science and Engineering and Center for Cognitive Science, The Ohio State University, Columbus, OH 43210 USA (e-mail: dwang@cse.ohio-state.edu; huk@cse.ohio-state.edu).

J.-S. R. Jang is with the Department of Computer Science, National Tsing Hua University, Hsinchu 30013, Taiwan (e-mail: jang@mirlab.org).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2011.2182510

does not need the accompaniment only parts. In [8], a predominant-F0 estimation method is proposed to estimate fundamental frequency (F0) using maximum *a posteriori* probability estimation and pitch trajectories based on pitch continuity. Li and Wang [13] proposed a computational auditory scene analysis (CASA) system which is effective for singing voice separation. In their system, pitch is detected based on a hidden Markov model (HMM). In [6], a source/filter model based approach is used to model singing voice. This model together with a music accompaniment model is used to perform a maximum likelihood estimation of the mixture. Pitch contours are estimated via a Viterbi-type algorithm based on source model parameters and pitch continuity. Tachibana *et al.* performed pitch estimation based on singing voice enhanced by a harmonic/percussive sound separation. They use matched filtering to model pitch likelihoods and a Gaussian mixture model (GMM) to model pitch transitions [21]. In a model-based method [7], sinusoidal models are used to model harmonic partials given detected pitch. The models are used to create smooth amplitude and phase trajectories over time and sinusoids are generated and summed to produce an estimate of the vocal signal. However, as pointed out in [23], a drawback of this method is that it will overestimate partials in the case of music interference.

In aforementioned methods, either separation of singing voice is based on detected pitch or pitch detection benefits from enhanced singing voices. As pointed out in [4], this interdependency between pitch estimation and pitch-based separation creates a "chicken and egg" problem for pitch estimation and voice separation. In order to escape from this dilemma in the context of speech separation, Hu and Wang [11] recently proposed a tandem algorithm which performs pitch estimation and voice separation jointly and iteratively. It is observed that the target pitch can be estimated from a few harmonics of the target signal. On the other hand, one can separate some target signals without perfect pitch estimation. Thus, their strategy is to have a rough estimate of the target pitch first and then separate the target speech by considering harmonicity and temporal continuity. The separated speech and the estimated target pitch are then used to improve upon each other iteratively. In [11], they show a consistent performance improvement for all types of intrusion except rock music, presumably because of the strong harmonicity of the music accompaniment. This indicates that separating speech from music is challenging to their tandem algorithm.

In this study, we investigate and extend the tandem algorithm to separate the voiced portions of singing voice from music accompaniment. To improve the performance of pitch-based voice separation, the most important issue is to improve pitch estimation. Most of pitch detection methods operate in a plausible pitch range chosen heuristically. The pitch range is usually large to cover most of the possible pitches of singing voice such as from 80 Hz to 500 Hz. However, as mentioned earlier the pitch of singing can be as high as 1400 Hz. On the other hand, it is unlikely that pitch changes in such a wide range in a short period of time.

To address the above problems, we propose a trend estimation algorithm to bound the singing pitch contours in a series of time–frequency (T-F) blocks that have much narrower pitch ranges compared to the entire possible range. The estimated trend substantially reduces the difficulty of singing pitch detection by eliminating a large number of wrong pitch candidates.

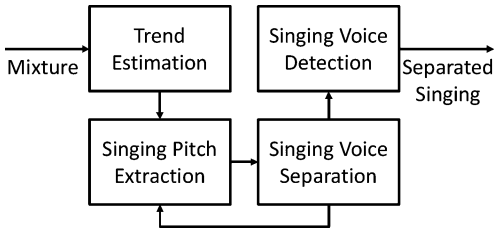

Fig. 1. Schematic diagram of the proposed system.

The rest of this paper is organized as follows. In Section II, we give an overview of the extended tandem algorithm. In Section III, we describe the proposed trend estimation in detail. Two key steps of the tandem algorithm, mask estimation and pitch determination, are then presented in Section IV. The iterative procedure of the tandem algorithm is explained in Section V. Section VI describes singing voice detection. The systematic evaluation on pitch estimation and singing voice separation is given in Section VII. Finally, we conclude this work in Section VIII.

II. System Overview

The proposed system is illustrated in Fig. 1. A trend estimation algorithm first estimates the pitch ranges of the singing voice. The estimated trend is then incorporated in the tandem algorithm to acquire the initial estimate of the singing pitch. Singing voice is then separated according to the initially estimated pitch. The above two stages, i.e., pitch determination and voice separation then iterate until convergence. A post-processing stage is introduced to deal with the "sequential grouping" problem, i.e., deciding which pitch contours belong to the target [2], [27], an issue unaddressed in the original tandem algorithm [11]. Finally, singing voice detection is performed to discard the nonvocal parts of the separated singing voice. As will be clear later, these improvements contribute to significantly better performance compared to the version for speech separation.

Our tandem algorithm detects multiple pitch contours and separates the singer by *estimating* the ideal binary mask (IBM), which is a binary matrix constructed using premixed source signals. In the IBM, 1 indicates that the singing voice is stronger than interference in the corresponding time-frequency unit and 0 otherwise. The IBM has been shown to be a reasonable goal of CASA [25] and used as a measure of ceiling performance for speech separation in many studies.

III. Trend Estimation

This section describes the proposed trend estimation algorithm. The purpose of trend estimation is to bound the singing pitch in a plausible range. As a result of placing a tight bound on the pitch range, we hope to reduce spurious pitches caused by music accompaniments or higher harmonics. First, the singing voice is enhanced by considering temporal and spectral smoothness. As the fundamental frequency of a singing voice tends to be smooth across time, we bound the vocal F0s in a series of time–frequency blocks. These T-F blocks give rough pitch

ranges along time which are much narrower than the possible pitch range.

A. Vocal Component Enhancement

It can be observed in a spectrogram that frequencies of the sounds generated by harmonic musical instruments (e.g., violin) are very smooth along the temporal direction while those from percussive instruments (e.g., drum) are smooth along the spectral direction. Ono *et al.* [15] handled harmonic/percussive source separation (HPSS) by using this characteristic to separate the mixture into harmonic components and percussive components. Specifically, HPSS is devised as an optimization problem that minimizes the following objective function:

$$\int\int\left(\frac{\partial}{\partial m}\left|U(m,f)\right|^{\gamma}\right)^{2}dmdf+\int\int\left(\frac{\partial}{\partial f}\left|V(m,f)\right|^{\gamma}\right)^{2}dmdf \tag{1}$$

under the constraint that $U(m,f)+V(m,f)=W(m,f)$, where $U(m,f)$ and $V(m,f)$ are the complex spectrogram components of harmonic sound and percussive sound to be estimated, $W(m,f)$ is the spectrogram of the input mixture, where m and f index the time frame and frequency bin, respectively. γ is an exponential constant (around 0.6) to imitate the auditory system.

Tachibana *et al.* [21] extended the original method to a multistage HPSS version to enhance the singing voice by tuning the smoothness along the time and frequency axes. The idea comes from the observation that the partials of singing voice are not as smooth as those of harmonic instruments along the temporal direction but much smoother than those of percussive instruments. They first apply a larger size window (e.g., 256 ms) to compute short time Fourier transform (STFT) so that the spectrum has a high spectral resolution and low temporal resolution. This biased resolution makes the partials of harmonic instruments smooth in the temporal direction since window lengths are long and bandwidths are narrow. Thus, more energy of harmonic instruments is assigned to the harmonic component. In other words, most of the energy of singing voice and percussive sounds are assigned to the percussive component. Afterward, HPSS is applied to the percussive component again but with a shorter window (e.g., 30 ms) to separate the singing voice from percussive sounds.

In this study, we only apply the first stage of the multistage HPSS which attenuates the energy of harmonic instruments. The reason is that the sounds of percussive instruments are aperiodic and do not create much difficulty in estimating target pitch. Fig. 2(a) and (b) shows the spectrograms of an input mixture before and after applying HPSS, respectively, and the solid lines represent the pitch contours of the singing voice. As we can see, the energy of the sounds produced by harmonic instruments is attenuated significantly.

B. Pitch Range Estimation

The goal of this stage is to find a sequence of relatively tight pitch ranges where the F0s of the singing voice are present. The main idea to achieve this goal is to remove unreliable peaks not originating from periodic sounds and then higher harmonics of the singing voice. The remaining peaks approximate fundamentals and we estimate the F0 range by bounding the peaks in a sequence of T-F blocks.

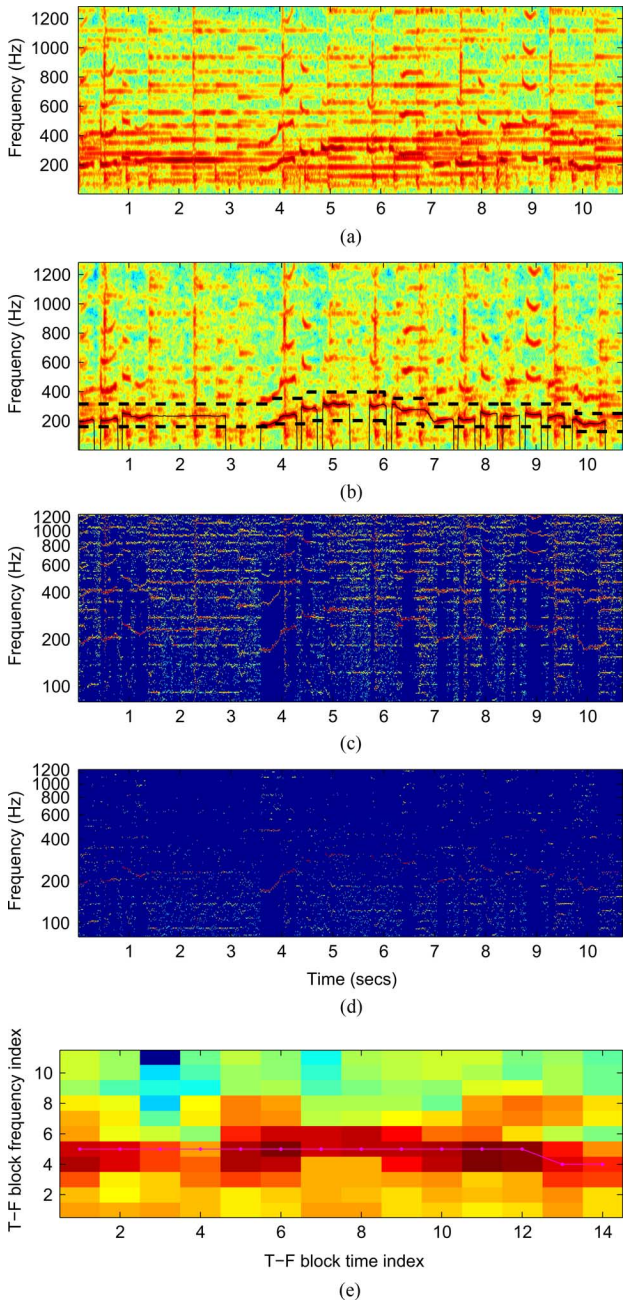

Fig. 2. Example of trend estimation. (a) The spectrogram of a song mixture. (b) The spectrogram after applying HPSS. Dashed lines show the extended boundaries of the estimated trend and solid lines show the ground truth pitch. (c) Results of MR-FFT. (d) Result after deleting harmonic peaks. (e) Magnitude-downsampled diagram. Color indicates the energy of a T-F block. The solid line indicates the optimal path found by DP.

First, we apply the multi-resolution fast Fourier transform (MR-FFT) proposed by Dressler [5]. The MR-FFT method analyzes sounds in different time-frequency resolutions by using different window lengths. Based on the local sinusoidality criteria, it deletes unreliable peaks which do not originate from periodic sounds by considering local characteristics of phase spectrum, or more precisely, the instantaneous frequencies of

neighboring frequency bins; a detailed discussion on phase and the instantaneous frequency can be found in [1].

Because some peaks in the same time frame may correspond to the same sinusoidal component, we first check the instantaneous frequencies of the peaks in each time frame. If their instantaneous frequencies are close enough (less than 0.2 semitone), the one with the largest magnitude is selected. Harmonics are then deleted based on the observation that a vocal F0 can only be the lowest frequency within a frame. Fig. 2(c) shows the result of MR-FFT and the result of deleting harmonics is shown in Fig. 2(d).

The remaining peaks are used to estimate the pitch range of singing voice. We first downsample the magnitudes of the peaks by summing the largest peak values in the frames within a T-F block, which is a rectangular area whose vertical side represents a frequency range and horizontal side a time duration. The entire plane is divided into a fixed set of T-F blocks with 50% overlap in time and in frequency. Then, given a multi-resolution spectrogram $x[m, f]$ produced by MR-FFT, the downsampled magnitude of the T-F block structure is defined as

$$b(T,F) = \sum_{m=0}^{M_T-1} \max_{f\in[0,M_F-1]} x\left[m+TL_T, f+FL_F\right], \quad T = 0,1,\ldots P-1 \text{ and } F = 0,1,\ldots Q-1 \tag{2}$$

where T and F indicate the time and frequency indices of a T-F block, respectively. P and Q indicate the ranges of T-F blocks in time and frequency, respectively. M_T and M_F are the numbers of time frames and frequency bins of a T-F block, respectively, and L_T and L_F are the shifts in frame and frequency bins of a T-F block, respectively.

Note that, although M_F is fixed for all T-F blocks, the bandwidths are different for T-F blocks with different frequency indices. This is because frequency bins in MR-FFT are spaced by 0.25 semitone. In other words, a T-F block with a smaller frequency index has a narrower bandwidth, consistent with human audition.

Finally, we find an optimal path consisting of a sequence of T-F blocks that contain the largest magnitudes by using dynamic programming (DP). The problem is defined as finding an optimal path $[F_0, F_1, \ldots, F_{P-1}]$ that maximizes the following score:

$$\sum_{T=0}^{P-1} b(T,F_T) - \theta \sum_{T=1}^{P-1} |F_T - F_{T-1}| \tag{3}$$

where the first term is the sum of strengths of the T-F blocks along the path, and the second term controls the smoothness of the path with the use of a penalty coefficient θ. The larger is θ, the smoother is a computed path. In this study, θ is set to half of the average strength of the largest $b(T,F)$ for each time index of T-F blocks:

$$\theta = \frac{1}{2}\left(\frac{\sum_{T=0}^{P-1} \max_{F\in[0,Q-1]} b(T,F)}{P}\right) \tag{4}$$

and the Viterbi algorithm [18] is used to find the optimal path. Fig. 2(e) shows an example of a magnitude-downsampled diagram. A color represents the energy strength of a T-F block. The solid line indicates the optimal path found by DP.

While the optimal path is a sequence of T-F blocks, the neighbors of these blocks also provide useful information which reveals the energy distribution of vocal partials since they are half overlapped. Specifically, given a T-F block $b(T,F)$ in the optimal path with frequency boundaries $f_{T,F}^{lower}$ and $f_{T,F}^{upper}$, we extend the lower and upper boundaries as follows:

$$\begin{cases} \left[f_{T,F}^{lower}, f_{T,F}^{upper}+\Omega\right], & \text{if } \frac{b(T,F-1)}{b(T,F+1)}(T,F+1) < 0.8 \\ \left[f_{T,F}^{lower}-\Omega, f_{T,F}^{upper}\right], & \text{if } \frac{b(T,F+1)}{b(T,F+1)}(T,F-1) < 0.8 \\ \left[f_{T,F}^{lower}-\frac{\Omega}{2}, f_{T,F}^{upper}+\frac{\Omega}{2}\right], & \text{else} \end{cases} \tag{5}$$

where Ω is the amount of extension and is set to 4 semitones in this study. The upper boundary of the selected T-F block $B_{T,F}$ is extended if its upper neighbor $B_{T,F+1}$ has much larger energy than its lower neighbor $B_{T,F-1}$. In this case, most of the energy of a vocal partial is likely concentrated at higher frequencies and we therefore extend the upper boundary to cover possible dynamics of vocal partials. The same procedure is also applied to extend a lower boundary. The estimated trend is then formed by these extended pitch ranges. This makes trend estimation capable of locating dominant partials in the first step and then extending its boundary to tolerate the possible pitch changes of the singing voice. Fig. 2(b) shows the extended boundaries of the estimated trend (dashed lines). As can be seen, the estimated trend bounds the singing F0s successfully.

IV. Mask Estimation and Pitch Determination

Two key steps of our tandem algorithm are: 1) IBM estimation given target pitch and 2) pitch determination given a binary mask. Since results of one stage are inputs to the other one, the two stages are used to improve each other iteratively. This section describes these two stages in detail.

A. Front-End Processing

We first perform time–frequency decomposition by using similar settings to those in [11]. The input song mixture is decomposed into 128 channels using the gammatone filterbank [17] whose center frequencies are quasi-logarithmically spaced from 50 Hz to 8 kHz. The signals in different frequency channels are then split into 40-ms frames with 20-ms overlap. Let u_{cm} denote a T-F unit at channel c and frame m, and $y(c,t)$ the filtered signal at channel c and time t. The corresponding normalized correlogram $A(c,m,\tau)$ at $u_{c,m}$ is computed by the following autocorrelation function (ACF):

$$A(c,m,\tau) = \frac{\sum_n y(c,mT_m - nT_n)\, y(c,mT_m - nT_n - \tau T_n)}{\sqrt{\sum_n y^2(c,mT_m - nT_n)\sum_n y^2(c,mT_m - nT_n - \tau T_n)}} \tag{6}$$

where τ is the time delay. T_m is the frame shift and T_n is the sampling time. The above summation is over 40 ms, the length of a time frame. The peaks of the ACF indicate the periodicity of the filter response, and the corresponding delays indicate the periods.

Two adjacent channels triggered by the same harmonic have highly correlated responses [3], [26], and we compute the cross-channel correlation between u_{cm} and $u_{c+1,m}$ by

$$C(c,m) = \frac{\sum_{\tau}[A(c,m,\tau)-\bar{A}(c,m)][A(c+1,m,\tau)-\bar{A}(c+1,m)]}{\sqrt{\sum_{\tau}[A(c,m,\tau)-\bar{A}(c,m)]^2 \sum_{\tau}[A(c+1,m,\tau)-\bar{A}(c+1,m)]^2}} \tag{7}$$

where $\bar{A}$ denotes the average of A over τ.

In high frequencies, a filter responds to multiple harmonics and the filtered signal is amplitude-modulated [10]. We thus calculate the ACF $A_E(c,m,\tau)$ using the envelopes of filtered signals. The corresponding cross-channel correlation is calculated similarly to (7). Here, we extract envelopes by halfwave rectification and bandpass filtering [11].

B. IBM Estimation Given Target Pitch

To estimate the IBM given the pitch of the target (singing voice), we first obtain a statistical model for generating the likelihoods of target-dominant and interference-dominant T-F units by a supervised learning scheme. Following [11], 6 features are extracted from each T-F unit, resulting in the following feature vector, shown in (8) at the bottom of the page, where the first 3 correspond to the filter response and the last 3 the envelope response. The function *int* (x) returns the nearest integer. Previous work [10], [13] shows that $A(c,m,\tau_S(m))$ is a good indicator of similarity between estimated pitch $\tau_S(m)$ and the response period of u_{cm}. $\overline{f}(c,m)$ is the average instantaneous frequency of the filter response within u_{cm}. If the filter response has a period close to $\tau_S(m)$, $\overline{f}(c,m)\cdot\tau_S(m)$ is close to an integer that indicates the harmonic number of the pitched sound. The second feature shows the difference of $\overline{f}(c,m)\cdot\tau_S(m)$ and its nearest integer, which measures how likely u_{cm} corresponds to a harmonic of the estimated pitch. The last 3 features are obtained similarly to the first 3 by replacing filter responses with response envelopes.

Let H_0 be the hypothesis that a T-F unit is target dominant and H_1 otherwise. u_{cm} is labeled as target if

$$P(H_0|r_{cm}(\tau_S(m))) > P(H_1|r_{cm}(\tau_S(m))). \tag{9}$$

Since $P(H_0|r_{cm}(\tau_S(m)))$ and $P(H_1|r_{cm}(\tau_S(m)))$ add to 1, we can perform the classification by

$$P(H_0|r_{cm}(\tau_S(m))) > 0.5. \tag{10}$$

For classification, we obtain $P(H_0|r_{cm}(\tau_s(m)))$ by training a multilayer perceptron (MLP) [20] to estimate the posterior probability with one hidden layer and 5 units for each filter channel, the same configuration as in [11]. During MLP training, ground truth pitch is used.

If more than one pitch candidate are detected at a frame, an additional probability comparison at the target pitch period and the interference pitch period is made within u_{cm}. Furthermore, the neighborhood of u_{cm} is considered for unit labeling. See [11] for details.

C. Pitch Estimation Given Binary Mask

A common way to estimate target pitch is to sum autocorrelations across all the channels and then identify the most dominant peak in the summary correlogram [4]. This can be improved by calculating the summary correlogram only from target-dominant T-F units according to the given binary mask $L(c,m)$ where a value of 1 indicates that $u_{c,m}$ is dominated by the target and 0 otherwise. Also, as shown in [11], replacing ACF with $P(H_0|r_{cm}(\tau))$ improves pitch estimation so we estimate the pitch period by

$$SP_m(\tau) = \sum_c P(H_0|r_{cm}(\tau))L(c,m). \tag{11}$$

We have found that only 0.07% (about 0.5% for speech in [11]) of consecutive frames in singing voice have more than 20% relative pitch changes in our training set. Hence, temporal continuity is used to check the reliability of the estimated pitch. If pitch changes of three consecutive frames are less than 20%, the estimated pitch periods of these three frames are considered reliable. Otherwise, an unreliable pitch is reestimated by limiting the plausible pitch range to 20% of a neighboring reliable pitch.

V. Tandem Algorithm

Our tandem algorithm has the same general steps in [11] but is different in the following ways: 1) the results of the trend estimation are provided to the tandem algorithm; 2) all statistical models used for estimating the IBM are re-trained by using music data instead of speech data; 3) input mixtures are pre-processed by HPSS; 4) since the tandem algorithm usually gives more than one pitch candidate for each frame, we add a sequential grouping step which selects one pitch value as the target as postprocessing. In the following subsections, we describe the tandem algorithm by highlighting the above aspects.

A. Initial Estimation

The iterative procedure starts with pitch estimation. Instead of using the original mixture as in [11], we use the output from HPSS as the input signal.

$$r_{cm}(\tau) = \begin{pmatrix} A(c,m,\tau), & \overline{f}(c,m)\tau - \mathrm{int}(\overline{f}(c,m)\tau), & \mathrm{int}\left(\overline{f}(c,m)\tau\right), \\ A_E(c,m,\tau), & \overline{f}_E(c,m)\tau - \mathrm{int}\left(\overline{f}_E(c,m)\tau\right), & \mathrm{int}\left(\overline{f}_E(c,m)\tau\right) \end{pmatrix} \tag{8}$$

We first treat all T-F units with high cross-channel correlations as dominated by a single source. The estimated pitch period is then selected as the one supported by most active (value 1) T-F units. A T-F unit u_{cm} is considered supporting a pitch period τ if the corresponding $P\left(H_0|r_{cm}(\tau)\right)$ is higher than 0.75; this threshold is relaxed, if needed, to ensure that there is at least one active unit in each frame. The T-F units that do not meet the threshold are then used to estimate the second pitch period if such units exist. Note that the possible pitch periods are now confined to the pitch range of the estimated trend. With the estimated pitch period τ, the corresponding mask is reestimated as the target if $P(H_0|r_{cm}(\tau)) > 0.5$.

After the above estimation, individual pitch periods are combined into pitch contours based on temporal continuity of both pitch periods and corresponding masks. As a result of this step, we obtain multiple pitch contours and their associated T-F masks.

B. Iterative Estimation

The key idea of this step is to expand the pitch contours according to temporal continuity. Let p_k be a pitch contour containing a sequence of pitch points in a continuous set of frames and $L_k(m) = \{L_k(c,m), \forall c\}$ be the associated mask at frame m. We first expand the mask by letting $L_k(m_1 - 1) = L_k(m_1)$ and $L_k(m_2 + 1) = L_k(m_2)$, where m_1 and m_2 are the first and last frame of the pitch contour, respectively. A new p_k is then estimated from this new mask [11]. If the newly estimated pitch points pass the continuity criterion [11], it is considered a reliable pitch. Otherwise, it is discarded.

Since the pitch periods in p_k are reestimated, we update the mask for each pitch contour as described in Section IV-B. The above two steps thus iterate until the estimation of pitch and mask converges, i.e., the estimated pitch contours and the corresponding masks do not change. This typically happens within 20 iterations.

We note here that the estimated pitches are not limited within the estimated trend during the iterative estimation. Thus, it is possible to detect the pitch points outside the trend if they have high temporal continuity. However, most estimated pitch contours are still within the trend because of the temporal continuity criterion.

Fig. 3 shows the pitch estimation result of the tandem algorithm where the input mixture is the same as that in Fig. 2. Fig. 3(a) shows the result without HPSS and trend estimation. Fig. 3(b) shows the result when HPSS is applied and Fig. 3(c) shows the result with both HPSS and trend estimation. Each color represents a pitch contour. Note that some pitch points are outside the trend around 3.8 seconds and are correctly detected because they have high temporal continuity with the previous pitch points. The raw-pitch accuracy of Fig. 3(a)–3(c) is 58%, 75%, and 90%, respectively. In this example, HPSS and trend estimation improve the singing pitch extraction significantly.

C. Postprocessing

Since the pitch estimation algorithm outputs up to two pitch values for each frame, some pitch contours may overlap in time.

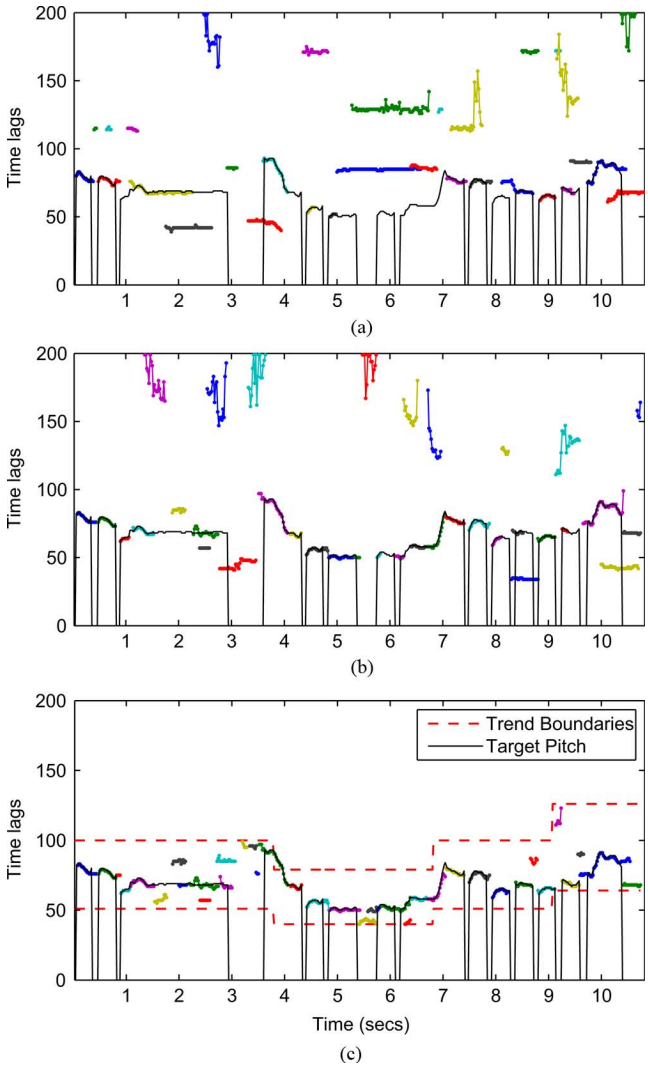

Fig. 3. Pitch estimation result of the tandem algorithm. (a) The result without HPSS and trend estimation. (b) The result with only HPSS. (c) The result with both HPSS and trend estimation. Different pitch contours are indicated by distinct colors.

This creates a sequential grouping problem that each pitch contour has to be assigned either to target voice or background noise. Sequential grouping is a difficult problem and the existing methods are not yet mature [27]. Fortunately, the proposed trend estimation algorithm is able to remove most of the time-overlapped pitch contours because it is unlikely that more than one pitch candidate are dominant in a narrow pitch range [see Fig. 3(c)]. This makes the sequential grouping problem much easier to address.

We group pitch contours as follows. First, we assign pitch points to the target pitch track for the frames that only have one pitch candidate. For the frames that contain more than one pitch candidate, we pick the one supported by most channels as the target pitch. Here, a channel is considered as supporting pitch candidate $\tau_S(m)$ if (10) is met. After that, we fill the gaps where no reliable pitch is estimated by using the most dominant pitch period in the initial estimation and thus generate a continuous pitch track. The corresponding mask to the target pitch track forms the estimated IBM.

VI. Singing Voice Detection

This stage employs a continuous HMM to decode an input mixture into vocal and nonvocal sections. We use the signals after applying HPSS which attenuates the energy from music accompaniment instead of the original mixture. Given the feature vectors $\{X_0, \ldots, X_m, \ldots X_M\}$ of an input mixture, the problem is to find the most likely sequence of vocal/nonvocal states, $\hat{S} = \{s_0, \ldots, s_m, \ldots, s_M\}$:

$$\hat{S} = \arg\max_{S} \left\{ P(s_0) P(X_0|s_0) \prod_{m=1}^{M} \{P(X_m|s_m) P(s_m|s_{m-1})\} \right\} \tag{12}$$

where $P(X|s)$ is the output probability density function (pdf) of binary state s (vocal or nonvocal), and $P(s_m|s_{m-1})$ is the transition probability from state s_{m-1} to s_m. The state output pdfs and features will be specified below in the evaluation section. As shown in [12], HMM generally outperforms static-classifier-based methods, such as GMM or perceptrons, for speech detection.

VII. Evaluation

In this section, we evaluate our proposed algorithm in four parts. The first part evaluates the performance of singing voice detection. The performance of trend estimation is evaluated in the second part. The third and the fourth parts evaluate pitch estimation and singing voice separation, respectively.

We use MIR-1K, a publicly available dataset introduced in our previous work [9], to evaluate our proposed system. Although there are several publicly available datasets that are commonly used in speech separation, as far as we know, MIR-1K is the only one designed for singing voice separation. It contains 1000 song clips recorded at 16-kHz sampling rate with 16-bit resolution. The duration of each clip ranges from 4 to 13 seconds, and the total length of the dataset is 133 minutes. These clips were extracted from 110 karaoke songs which contain a mixed track and a music accompaniment track. These songs were selected from 5000 Chinese pop songs and sung by 8 females and 11 males. Most of the singers are amateurs with no professional training. The music accompaniment and the singing voice were recorded at the left and right channels, respectively. The ground truth of the target pitch is estimated by using clean singing voice with manual adjustment. All songs are mixed at -5, 0, and 5 dB for evaluation.

A. Evaluation of Singing Voice Detection

1) Dataset Description: We use all 1000 clips of MIR-1K for training and evaluating the HMM for singing voice detection. The dataset is divided into two subsets of similar sizes (487 versus 513, recorded by disjoint subjects) for two-fold cross validation. Singing voices and music accompaniments are mixed at 0-dB SNR to generate the training samples. Here the singing voice is the signal and the music accompaniment the noise.

2) Acoustic Features: 39-dimensional MFCCs (12 cepstral coefficients plus log energy, together with their first and second derivatives) are extracted from each frame. The MFCCs are computed from STFT with a half-overlapped 40-ms Hamming window. Cepstral mean subtraction (CMS) is used to reduce channel effects.

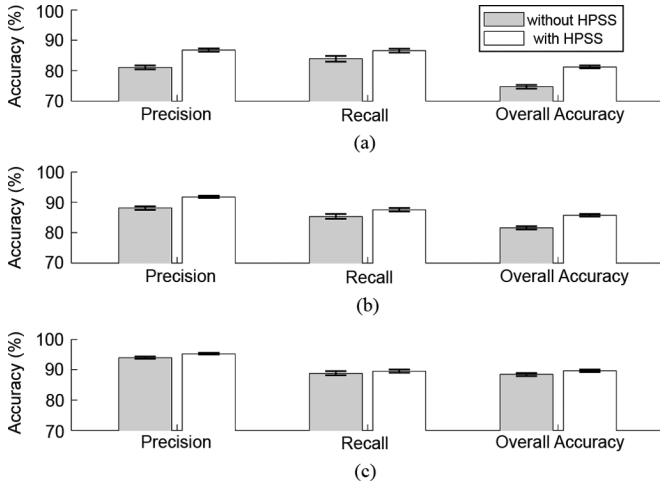

Fig. 4. Performance of singing voice detection. The precision, recall, and overall accuracy at -5, 0, and 5 dB SNR are shown in (a) (b), and (c) respectively.

3) Performance Measure: The performance of singing voice detection is represented by precision, recall, and overall accuracy. The recall for vocal frame detection is the percentage of the frames that are correctly classified as vocal over all the vocal frames; the precision is the percentage of the frames that are correctly classified as vocal over the frames that are classified as vocal. The overall accuracy is the percentage of all the frames correctly classified.

4) Experimental Settings and Results: Two 32-components GMMs are trained for vocal frames and nonvocal frames, respectively. Both GMMs have diagonal covariance matrices. Parameters of the GMMs are initialized via a K-means clustering algorithm [14] and are iteratively adjusted via an expectation–maximization algorithm with 20 iterations. Each of the GMMs is considered as a state in a fully connected HMM, where the transition probabilities are obtained through frame counts of the labeled dataset. For a given new mixture, the Viterbi algorithm is used to decode the mixtures into vocal and nonvocal segments.

Fig. 4 shows the performance of singing voice detection. We evaluate the algorithm for the signals mixed at -5, 0 dB, and 5 dB. The gray bars and white bars show the average performance without and with HPSS as preprocessing, respectively. For each mean, the error bars show the 95% confidence intervals. As can be seen, the results with HPSS perform uniformly better, especially for lower SNR levels. The results show that singing voice detection benefits significantly from HPSS which attenuates the energy of music accompaniment. We also tried other features such as perceptual linear prediction, but found no performance gain.

B. Evaluation of Trend Estimation

1) Dataset Description: All 1000 song clips in MIR-1K are used for evaluating the performance of trend estimation.

2) Performance Measure: The estimated trend is treated as correct if the ground truth pitch is located within the lower

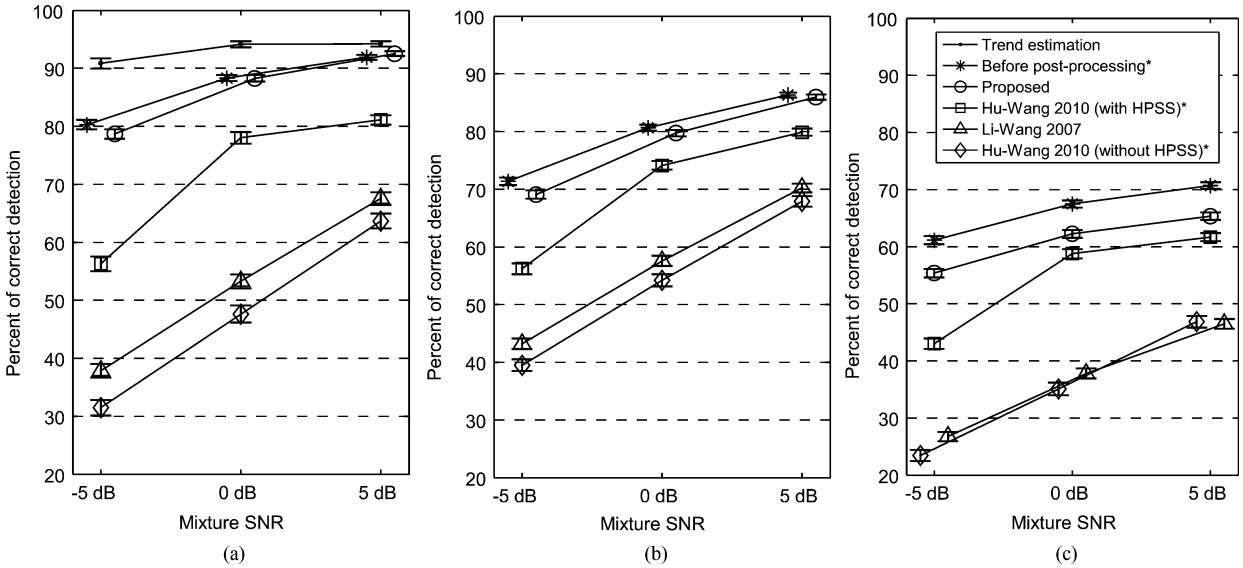

Fig. 5. Results of trend estimation and pitch extraction for different algorithms. (a) The results for voiced parts of the recordings. (b) The results for whole recordings with vocal detection. (c) The results for whole recordings without vocal detection. $*$ after a method indicates the ceiling performance.

bound and upper bound of the trend. The correct rate is calculated for vocal frames.

3) Experimental Settings and Results: The duration and pitch range for each T-F block is set to be 1.5 second and 8 semitones, respectively, with 50% overlap for both time and frequency. The bandwidth of the trend is equal to one octave (12 semitones) after boundary extension.

Fig. 5 shows the performance of trend estimation and singing pitch detection at different SNR levels. Means together with 95% confidence intervals are shown. As we can see in Fig. 5(a), our trend estimation performs very well. Accuracy of the trend estimation rises slightly when the SNR increases. It is expected since the energy of vocal sounds is more prominent at higher SNRs. Nevertheless, the trend estimation achieves at least 90% accuracy on average and is robust at different SNRs.

C. Evaluation of Singing Pitch Extraction

1) Dataset Description: The dataset is divided into two subsets in the same way as described in Section VII-A for two-fold cross validation. Singing voices and music accompaniments are mixed at 0-dB SNR to generate the samples for training the MLP mentioned in Section IV-B.

2) Performance Measure: The estimated pitch is treated as correct if the difference to the ground truth pitch is less than 5% in Hz. In addition, we calculated correct detection rate for voiced parts and whole recordings with and without singing voice detection.

3) Experimental Results: The results of singing pitch extraction are shown in Fig. 5, where the voiced only results in Fig. 5(a) show the performance of different pitch extraction algorithms when voiced singing is present (according to the ground truth). Fig. 5(b) and (c) gives the overall results with and without singing voice detection, where an overall result is the percentage of all the frames (voiced or unvoiced) with correct pitch detection; in other words, both voiced/unvoiced classification errors and pitch estimation errors (for correctly classified voiced frames) are counted in the overall results. In each case, the performance of the tandem algorithm before and after postprocessing is shown. In addition, we show the performance of the original tandem algorithm [11] with and without applying HPSS; note that the plausible pitch range in [11] has been widened to 80–800 Hz to account for singing voice. The results of the singing pitch detection algorithm proposed by Li and Wang [13] are also given. An asterisk after a method indicates the ceiling performance of the algorithm where we treat a result as correct if one of the two estimated pitch candidates is correct. The algorithms without an asterisk only produce one pitch value for a frame. Note that in the figure we slightly shift some curves to avoid overlapping.

As can be seen in Fig. 5(a), the results of the proposed algorithm with sequential grouping as postprocessing are very close to the ceiling performance before postprocessing. This confirms that our postprocessing step deals with the sequential grouping problem successfully with the help of trend estimation that eliminates most of overlapped pitch contours.

Comparing with previous methods, our proposed algorithm performs substantially better. It is worth noting that applying HPSS to the original tandem algorithm improves its ceiling performance significantly. This shows that vocal enhancement using HPSS is clearly helpful for singing pitch extraction. However, our proposed algorithm still outperforms theirs in this case and shows the effectiveness of our trend estimation. Comparing Fig. 5(b) and (c), the performance is significantly improved for all algorithms when singing voice detection is applied and this shows the importance of singing voice detection.

We also submitted a preliminary version of the pitch extraction algorithm to the Music Information Retrieval Evaluation eXchange (MIREX) 2010 audio melody extraction competition. Each submission has been tested on six datasets which are hidden to the participants. Our algorithm achieved the best average raw-pitch accuracy for vocal songs comparing to the other 16 submissions in the last two years. Details can be found at http://nema.lis.illinois.edu/nema_out/mirex2010/results/ame/. These results indicate the effectiveness of our pitch detection

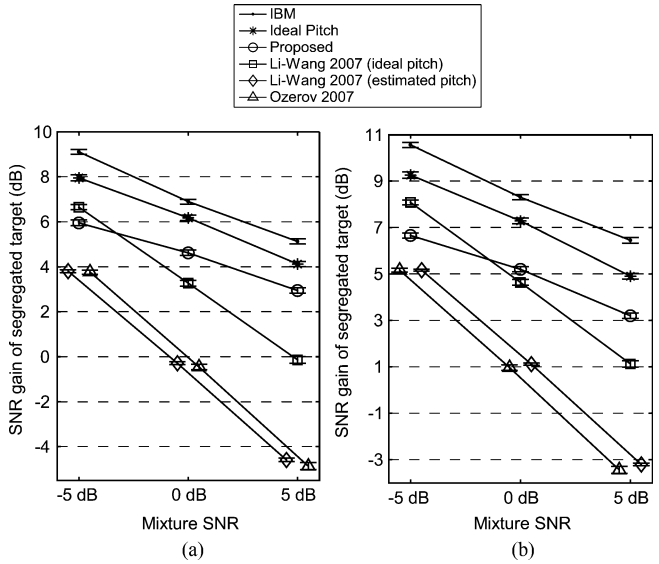

Fig. 6. Results of singing voice separation. (a) SNR gains for voiced parts of the recordings. (b) SNR gains for whole recordings with vocal detection.

in corpora in addition to MIR-1K. Since we have improved both trend estimation and pitch extraction, we expect that better performance will be achieved by the proposed algorithm.

D. Evaluation of Singing Voice Separation

1) Performance Measure: To compare the waveforms directly, we measure the SNR of the separated singing voice in decibels [10]:

$$SNR = 10\log_{10}\frac{\sum_n v^2(n)}{\sum_n \left[v(n)-\hat{v}(n)\right]^2} \tag{13}$$

where $v(n)$ is the target signal which is generated by applying an all-one mask to clean singing voice and $\hat{v}(n)$ is the separated singing voice.

2) Experimental Results: Fig. 6 shows SNR gains of the separated singing voice for different methods. Fig. 6(a) and (b) shows the results for voiced parts and for whole recordings with singing voice detection, respectively. The IBM results show the ceiling performance of a binary mask based algorithm. The results with ideal pitch show the performance when ground truth pitch is given to the tandem algorithm. The performance of the proposed algorithm is shown by circle line. The results of a previous CASA-based singing voice separation method [13] and a model-based method proposed by Ozerov *et al.* [16] are also shown for comparison. The latter results were generated by the authors of [16].

As can be seen, the proposed algorithm outperforms the previous systems significantly. Even with provided target pitch, the Li–Wang system does not perform as well as the proposed algorithm except for -5 dB SNR. This is because classification-based mask estimation performs much better than the rule-based one used in their system. By providing ideal pitch to both the proposed and the Li-Wang system, we can see that the separation results of the current system are significantly better and much closer to the IBM results.

VIII. Conclusion

We have proposed an extended tandem algorithm that estimates target pitch and separates singing voice from music accompaniment iteratively. By coupling with the proposed trend estimation, the tandem algorithm is improved significantly for both pitch estimation and voice separation in musical recordings. Systematic evaluation shows that the proposed algorithm performs significantly better than a previous CASA and a model-based system. Together with our previous system for unvoiced singing voice separation [9], we have a CASA system to separate both voiced and unvoiced portions of singing voice from music accompaniment.

In this study, we use the MIR-1K corpus which contains song mixtures synthetically created so that the ground truth is available for evaluation purposes [9]. Whether the observed performance from MIR-1K extends to recorded mixtures with professional singers needs to be investigated in future research. Another interesting issue is the tradeoff between the pitch range of a trend and the utility of trend estimation in pitch estimation. A broader range increases the chance of enclosing the true pitch point, leading to more accurate trend estimation; on the other hand, it is less useful as a constraint for pitch estimation. In this paper, trend estimation is performed prior to pitch estimation. Future work needs to analyze this tradeoff to see if there is an optimal trend range for pitch detection.

References

[1] T. Abe and M. Honda, "Sinusoidal model based on instantaneous frequency attractors," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1292–1300, Jul. 2006.

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

[3] G. J. Brown and M. Cooke, "Computational auditory scene analysis," *Comput. Speech Lang.*, vol. 8, pp. 297–336, 1994.

[4] A. de Cheveigne, "Multiple F0 Estimation," in *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, D. L. Wang and G. J. Brown, Eds. Hoboken, NJ: Wiley and IEEE Press, 2006, pp. 45–79.

[5] K. Dressler, "Sinusoidal extraction using an efficient implementation of a multi-resolution FFT," in *Proc. Int. Conf. Digit. Audio Effects*, 2006, pp. 247–252.

[6] J.-L. Durrieu, G. Richard, and B. David, "Singer melody extraction in polyphonic signals using source separation methods," in *Proc. IEEE ICASSP*, 2008, pp. 169–172.

[7] H. Fujihara, M. Goto, J. Ogata, K. Komatani, T. Ogata, and H. G. Okuno, "Automatic synchronization between lyrics and music CD recordings based on Viterbi alignment of segregated vocal signals," in *Proc. IEEE Int. Symp. Multimedia*, San Diego, CA, 2006, pp. 257–264.

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

[9] C.-L. Hsu and J.-S. R. Jang, "On the improvement of singing voice separation for monaural recordings using the MIR-1K dataset," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 310–319, Feb. 2010.

[10] G. Hu and D. L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. Neural Netw.*, vol. 15, no. 5, pp. 1135–1150, Sep. 2004.

[11] G. Hu and D. L. Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 8, pp. 2067–2079, Nov. 2010.

[12] J. Li and C.-H. Lee, "On designing and evaluating speech event detectors," in *Proc. Interspeech*, Lisbon, Portugal, Sep. 2005.

[13] Y. Li and D. L. Wang, "Separation of singing voice from music accompaniment for monaural recordings," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1475–1487, May 2007.

[14] S. P. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 2, pp. 129–137, Mar. 1982.
[15] N. Ono, K. Miyamoto, J. L. Roux, H. Kameoka, and S. Sagayama, "Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram," in *Proc. EUSIPCO*, 2008.
[16] A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval, "Adaptation of bayesian models for single channel source separation and its application to voice/music separation in popular songs," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 5, pp. 1564–1578, Jul. 2007.
[17] R. D. Patterson, J. Holdsworth, I. Nimmo-Smith, and P. Rice, "An efficient auditory filterbank based on the gammatone function," MRC Appl. Psychol. Unit, 1988, Rep. 2341.
[18] L. R. Rabiner, "A tutorial on hidden Markov models and selected application in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.
[19] B. Raj, P. Smaragdis, M. V. Shashanka, and R. Singh, "Separating a foreground singer from background music," in *Proc. Int. Symp. Frontiers Res. Speech Music (FRSM)*, Mysore, India, 2007.
[20] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," in *Parallel Distributed Processing*, D. E. Rumelhart and J. L. McClell, Eds. Cambridge, MA: MIT Press, 1986, pp. 318–362.
[21] H. Tachibana, T. Ono, N. Ono, and S. Sagayama, "Melody line estimation in homophonic music audio signals based on temporal-variability of melody source," in *IEEE ICASSP*, 2010, pp. 425–428.
[22] S. Vembu and S. Baumann, "Separation of vocals from polyphonic audio recordings," in *Proc. ISMIR*, 2005, pp. 337–344.
[23] T. Virtanen, A. Mesaros, and M. Ryynänen, "Combining pitch-based inference and non-negative spectrogram factorization in separating vocals from polyphonic music," in *Proc. SAPA*, Brisbane, Australia, 2008.
[24] 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*, Geneva, Switzerland, 2003.
[25] D. L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, P. Divenyi, Ed. Norwell, MA: Kluwer, 2005, pp. 181–197.
[26] D. L. Wang and G. J. Brown, "Separation of speech from interfering sounds based on oscillatory correlation," *IEEE Trans. Neural Netw.*, vol. 10, no. 3, pp. 684–697, May 1999.
[27] , D. Wang and G. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Hoboken, NJ: Wiley and IEEE Press, 2006.
[28] Y. Wang, M.-Y. Kan, T. L. Nwe, A. Shenoy, and J. Yin, "LyricAlly: Automatic synchronization of acoustic musical signals and textual lyrics," in *Proc. 12th Annu. ACM Int. Conf. Multimedia*, New York, 2004, pp. 212–219.
[29] T. Zhang, "Perception of singing," in *Psychology of Music*, D. Deutsch, Ed., 2nd ed. New York: Academic, 1999, pp. 171–214.
[30] T. Zhang, "System and method for automatic singer identification," in *Proc. IEEE ICME*, 2003, pp. 33–36.

Chao-Ling Hsu received the B.S. degree in computer science and information engineering from Chung-Hua University, Hsinchu, Taiwan, in 2003, and the Ph.D. degree in computer science from National Tsing Hua University, Hsinchu, Taiwan, in 2011.

In 2010, he was a Visiting Scholar in the Department of Computer Science and Engineering, Ohio State University (OSU), Columbus. He is currently a Senior Engineer at Mediatek, Inc., Hsinchu, Taiwan, developing audio algorithms and related products. His research interests include melody recognition, music signal processing, computational auditory scene analysis, and speech enhancement.

DeLiang Wang, (F'04) photograph and biography not available at the time of publication.

Jyh-Shing Roger Jang received the B.S. degree in electrical engineering from National Taiwan University, Taipei, Taiwan, in 1984, and the Ph.D. degree in the Electrical Engineering and Computer Science Department, University of California at Berkeley, in 1992.

He was with the MathWorks, Inc., from 1993 to 1995, and coauthored the Fuzzy Logic Toolbox. Since 1995, he has been with the Department of Computer Science, National Tsing Hua University, Taiwan. He has published several books, including *Neuro-Fuzzy and Soft Computing* (Prentice Hall, 1997), *MATLAB Programming* (2004, in Chinese), and *JavaScript Programming and Applications* (2007, in Chinese). He has also maintained two online tutorials on "Audio Signal Processing and Recognition" and "Data Clustering and Pattern Recognition." His research interests include speech recognition/assessment, music analysis and retrieval, face detection/recognition, pattern recognition, neural networks, and fuzzy logic.

Ke Hu (S'09) received the B.E. and M.E. degrees in automation from University of Science and Technology of China, Hefei, in 2003 and 2006, respectively, and the M.S. degree in computer science and engineering from The Ohio State University, Columbus, in 2010, where he is currently pursuing the Ph.D. degree.

His research interests include monaural speech separation, statistical modeling, and machine learning.