Vocality-Sensitive Melody Extraction from Popular Songs

Yu-Ren Chien and Hsin-Min Wang
Institute of Information Science
Academia Sinica, Taiwan
e-mail: yrchien@ntu.edu.tw, whm@iis.sinica.edu.tw

***Abstract*—This paper presents a method for extracting vocal melodies from popular songs. Underlying the extraction procedure is a sinusoidal representation applied to the input song signal. The desired vocal melody is isolated by focusing on specific (amplitude- and frequency-modulated) sinusoids that are identified as vocal, with the identification based on minimum mean square error (MMSE) estimation of the singing voice. The experiment results show that sensitivity to vocality brings a 15% absolute gain in vocal pitch recall, and that the proposed system is effective in indexing a 95-song database in a query-by-singing application.**

I. Introduction

A popular song is typically composed of a solo singing voice and a polyphonic, instrumental accompaniment. The singing voice plays a predominant role in a listener's attention to, and memory of, a song, and unambiguously sets the song apart from other songs; therefore in a song retrieval application, descriptions of the singing voice would naturally make adequate searching criteria. Since it is usually difficult to reproduce the vocal quality in a query, a practical representation of the singing voice would consist of the melody and rhythm only. A system that automatically reduces a song signal into such a representation would be valuable for automatic indexing of a song database. In this context, we present a method for extracting essential singing-voice features from a song signal. Specifically, this work addresses the issue of locating voiced frames in the signal (voicing detection) and estimating the fundamental frequency of the voice for each voiced frame (vocal pitch estimation); that is, for each audio frame in the song signal, we aim to determine 1) whether or not there exists a singing voice, and 2) the pitch sung by the artist if voice is detected at the frame.

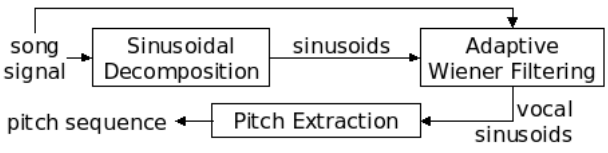

Fig. 1. The vocal melody extraction system.

As represented in Fig. 1, our approach consists of decomposing the input song signal into a set of *amplitude- and frequency-modulated* sinusoids, and subsequently classifying the modulated sinusoids into vocal sinusoids and instrumental sinusoids. We adopt the standard sinusoidal modeling approach [1], [2] to the decomposition. The classification is based on minimum mean square error (MMSE) estimation of the singing voice, which has been inspired by Ozerov et al.'s work [3] on Wiener separation of singing voice. After the voice signal is estimated from the song signal according to separate probabilistic models of singing voice and accompaniment, we compute frame-wise frequency responses that map the song signal to the voice estimate, and apply these responses to each modulated sinusoid separately. The result from this time-varying filtering is supposed to be either slight or significant attenuation in sinusoidal energy, depending on whether a vocal or an instrumental sinusoid is being filtered. To give vocality decisions, we define a *vocality score* for each modulated sinusoid as its associated energy gain under the filtering, which represents the proportion of the sinusoidal energy that is vocal. With such a measurement, a modulated sinusoid is designated as vocal if the vocality score is above a certain threshold; otherwise, it is considered instrumental. Once vocal sinusoids in the input song signal have been identified, both voicing detection and vocal pitch estimation can be carried out as straightforward extensions of sinusoidal modeling.

The contribution of this work is explained in the following two aspects:

- We propose a transformation from a separating filter to an array of timbre recognition modules. Such transformation is achieved by summarizing the frequency response of the filter along various sinusoidal trajectories traced by the input song signal in the time-frequency plane, which gives the vocality scores.
- We demonstrate the practical efficacy of the proposed melody extractor by a retrieval application where 90 sung queries are matched against a 95-song database indexed fully automatically by the extractor.

The remainder of this paper is organized as follows. Section II reviews some related work. In Sections III, IV, and V, we present the sinusoidal decomposition, adaptive Wiener filtering, and pitch extraction methods, respectively. The experiment results are detailed in Section VI. Finally, in Section VII we present our conclusions and suggest some future research directions.

This work was supported in part by Taiwan e-Learning and Digital Archives Program (TELDAP) sponsored by the National Science Council of Taiwan under Grant: NSC98-2631-001-013.

II. Related Work

One of the central precepts of professional music interpretation is the need to make each melodic note in the music stand out from all other concurrent (non-melodic) notes in the resulting performance. This has motivated a body of work [4], [5], [6], [7] where, of all the pitches detected in each audio frame, the one with the largest power is considered melodic. Compared with this family of methods, the development of our approach has been limited by its dependence on the timbral distinction of the melody; however, sensitivity to timbre is required of a melody extraction system in applications where a user's interest in the musical contents is closely tied to such distinction, as in song retrieval, where the user typically does not supply an instrumental melody as a query.

Timbre-aware melody extraction is not new in musical signal processing. Li and Wang [8] built acoustical knowledge into their system, where vocal pitch is detected by checking for beating (a phenomenon usually caused by strong high-order partials in the voice) in a particular frequency range. A purely statistical approach was taken by Ellis and Poliner [9], where timbral selectivity is implicitly learned from a set of vocal-pitch-labeled song spectra prepared for SVM training. Eggink and Brown [10] took a hybrid approach, whereby the audio signal is decomposed by sinusoidal modeling, and modulated sinusoids generated by the solo instrument are identified according to a probabilistic model of such sinusoids, without modeling the accompaniment. The method in [11] is more closely related to our approach, in that monophonic components in the input song signal are classified into vocal components and instrumental components according to both a vocal-component model and an instrumental-component model. In our work, instead of modeling each separate monophonic component, we build models for the entire spectrum (a singing-voice spectrum model and an accompaniment spectrum model), and classify multiple modulated sinusoids according to a single singing-voice spectrum estimated from the input song spectrum.

Sinusoidal modeling and grouping, which underlies our signal decomposition and pitch extraction procedures, has recently been pursued by Lagrange *et al.* [12] via a graph-theoretic clustering of spectral peaks. By interpreting the resulting clusters as separated acoustic objects, and pruning simultaneous objects by harmonicity, they achieved singing voice separation from accompanied singing signals.

III. Sinusoidal Decomposition

A musical pitch is realized in an acoustic signal as a short-time quasi-periodic component of the signal. In the frequency domain, the component takes the form of one or more sharp peaks in the short-time spectrum that are often harmonically related. Therefore, to find pitches in the input song signal, the first step of processing would be finding all the short-time spectral peaks in the spectrogram. To approach this task, we adopt the sinusoidal modeling technique in [2], which performs the task in much the same way as how we see continuous curves in a spectrogram of musical sound (see Fig. 2), by organizing all the peaks into a much smaller number

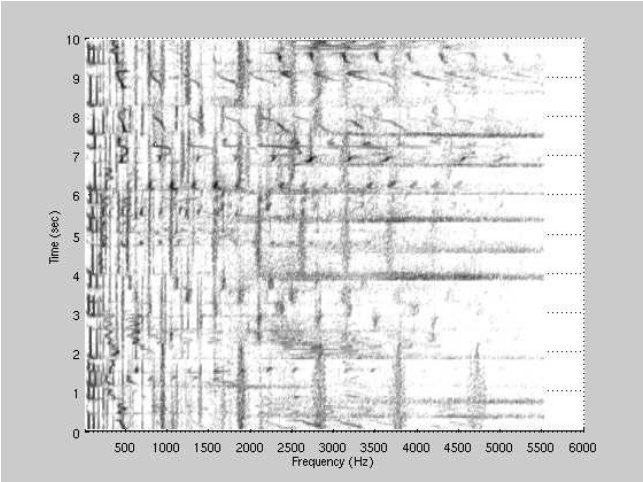

Fig. 2. Spectrogram of a 10-second song clip.

of 2-dimensional trajectories, or what we call *amplitude- and frequency-modulated sinusoids*. The decomposition procedure extracts a list of modulated sinusoids from the song signal, each comprising a sequence of spectral peaks satisfying certain continuity constraints both in frequency and in amplitude. To ensure tractable processing at later stages, any modulated sinusoid that has relatively low average power is excluded from the final list. As an illustration, Fig. 3 shows a sinusoidal

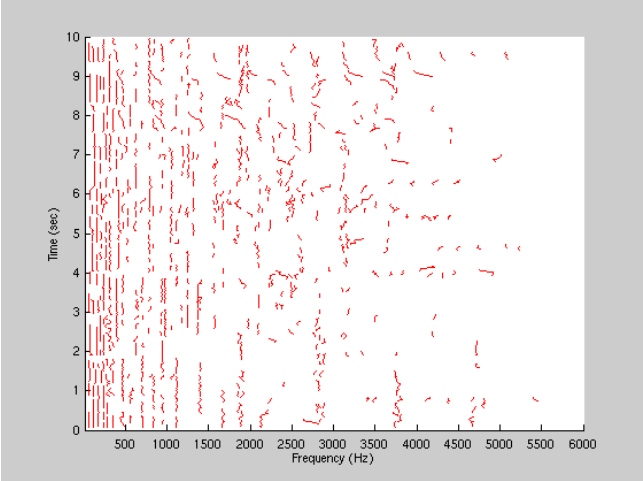

Fig. 3. Modulated sinusoids extracted from a 10-second song clip.

representation of the signal depicted in Fig. 2.

Note that we do not resynthesize separate sinusoids. Instead, each modulated sinusoid in the song signal is represented by its associated sinusoidal parameters, i.e., a frequency sequence, an amplitude sequence, and a frame span. These parameters define the song signal $s(t)$ as follows:

$$s(t) \approx \sum_i a_i(t) \cos\left(2\pi \int_0^t f_i(\tau)\, d\tau\right), \qquad (1)$$

where $a_i(t)$ and $f_i(t)$ are the amplitude and frequency of the ith modulated sinusoid, respectively; the support of $a_i(t)$ is defined by the frame span of the modulated sinusoid; and the amplitude and frequency sequences of the modulated sinusoid sample $a_i(t)$ and $f_i(t)$ within the support, respectively.

IV. Adaptive Wiener Filtering

Standard Wiener filtering separates two stationary source signals using knowledge of their respective power spectral densities (PSDs). In the case of singing voice separation, an extension to the standard technique was proposed in [3] to handle the nonstationary song signal, which can be viewed as adapting the filter to "time-varying" source PSDs estimated from the song signal. The estimation is based on modeling the joint probability of the voice and accompaniment PSDs at a particular instant given a short-time song signal around the instant, which, in our implementation, depends on two Gaussian mixture models (GMMs) built in advance: V-GMM, fitted to a training set of singing voice power spectra [13], and A-GMM, fitted to a training set of accompaniment power spectra. The adaptive Wiener filter is determined by the GMMs as well as by the input song signal.

Given a song signal, estimation of its vocal component is performed in the short-time Fourier transform (STFT) domain, with each short-time voice spectrum inferred solely from the corresponding short-time song spectrum. Given a song power spectrum $\mathbf{s}$, the MMSE estimator of the corresponding voice power spectrum $\mathbf{v}$ is

$$\begin{aligned} & E\left[\mathbf{v}|\mathbf{s}\right] \\ = & \sum_{i,j} P\left(q_v=i,\, q_a=j\,|\mathbf{s}\right) \cdot E\left[\mathbf{v}|\mathbf{s},\, q_v=i,\, q_a=j\right], \end{aligned} \tag{2}$$

where the discrete random variable q_v is the index of the "active" component in V-GMM, and q_a is the same index for A-GMM.

A. Evaluating the Conditional Mean

Practical evaluation of the conditional mean expanded in (2) involves some approximations:

- Evaluation of the joint prior probability of q_v and q_a is based on the assumption that they are almost independent, i.e.,

$$\begin{aligned} & P\left(q_v=i,\, q_a=j\,|\mathbf{s}\right) \\ \propto\ & P\left(q_v=i,\, q_a=j\right) \cdot p\left(\mathbf{s}|q_v=i,\, q_a=j\right) \\ \approx\ & P\left(q_v=i\right) \cdot P\left(q_a=j\right) \cdot p\left(\mathbf{s}|q_v=i,\, q_a=j\right), \end{aligned} \tag{3}$$

  where the marginal priors are among the parameters of the GMMs, i.e., the mixing probabilities.
- Evaluation of the likelihood in (3) is based on the assumption that frequency components are almost independent:

$$\begin{aligned} & p\left(\mathbf{s}|q_v=i,\, q_a=j\right) \\ \approx\ & \textstyle\prod_k p\left(s_k|q_v=i,\, q_a=j\right). \end{aligned} \tag{4}$$

  Since power spectral density is additive when two uncorrelated random processes are mixed, and can be shown to equal the expectation of power spectrum [14], here we assume that power spectrum was also additive when the singing voice and the accompaniment were mixed to give the song signal, i.e.,

$$s_k \approx v_k + a_k, \tag{5}$$

  where v_k and a_k are the power spectrum values of the singing voice and the accompaniment, respectively, at frequency bin k. By the construction of diagonal-covariance V-GMM and A-GMM, we have

$$p\left(v_k|q_v=i\right) = \frac{1}{\sqrt{2\pi}\sigma_{vik}} \exp\frac{-(v_k-\mu_{vik})^2}{2\sigma_{vik}^2}, \tag{6}$$

  and

$$p\left(a_k|q_a=j\right) = \frac{1}{\sqrt{2\pi}\sigma_{ajk}} \exp\frac{-(a_k-\mu_{ajk})^2}{2\sigma_{ajk}^2}, \tag{7}$$

  where $\boldsymbol{\mu}_{vi}$ and $\boldsymbol{\sigma}_{vi}$ denote the ith mean and standard-deviation vectors in V-GMM, and $\boldsymbol{\mu}_{aj}$ and $\boldsymbol{\sigma}_{aj}$ denote the jth mean and standard-deviation vectors in A-GMM. Equations (6) and (7), along with (5) and an independence assumption between v_k and a_k, yield an expression for the frequency-specific likelihood in (4):

$$\begin{aligned} & p\left(s_k|q_v=i,\, q_a=j\right) \\ \approx\ & \frac{1}{\sqrt{2\pi\left(\sigma_{vik}^2+\sigma_{ajk}^2\right)}} \exp\frac{-(s_k-\mu_{vik}-\mu_{ajk})^2}{2\left(\sigma_{vik}^2+\sigma_{ajk}^2\right)}, \end{aligned} \tag{8}$$

  which is the density of the sum of two independent Gaussian random variables.
- Suppose that the voice power spectrum is generated by the ith component of V-GMM, and that the accompaniment power spectrum is generated by the jth component of A-GMM. Again, since power spectral density equals the expectation of power spectrum, the PSDs of the voice and accompaniment signals are equal to $\boldsymbol{\mu}_{vi} = E\left[\mathbf{v}|q_v=i\right]$ and $\boldsymbol{\mu}_{aj} = E\left[\mathbf{a}|q_a=j\right]$, respectively. Based on the fact that the song signal is the sum of the voice and accompaniment signals, and the assumption that the latter two signals are orthogonal to each other, we can approximate the MMSE estimation of the voice power spectrum $\mathbf{v}$ by a special case of Wiener filtering [15], as follows:

$$E\left[v_k|\mathbf{s},\, q_v=i,\, q_a=j\right] \approx \left(\frac{\mu_{vik}}{\mu_{vik}+\mu_{ajk}}\right)^2 \cdot s_k \tag{9}$$

B. Timbre Recognition Based on the Voice Estimate

The estimation in (2) amounts to passing the N-point windowed (discrete-time) song signal through the adaptive Wiener filter with frequency (magnitude) response $H(\omega)$ sampled by

$$H\left(2\pi k/N\right) = \sqrt{E\left[v_k|\mathbf{s}\right]/s_k}, \tag{10}$$

where k is the discrete Fourier transform (DFT) frequency index. What we expect of this filter is the effect of attenuating

instrumental components in the song signal, i.e., "vocal-pass filtering."

The goal of the procedure described in this section is to determine whether each modulated sinusoid in the input song signal is vocal or instrumental. This is achieved by taking advantage of instantaneous vocal-pass filters $\{H_n\}_{n=1}^{L}$, where n denotes the frame index, and L denotes the number of frames. Consider a modulated sinusoid composed of P short-time spectral peaks, where the amplitude, digital frequency, and frame index of each peak are denoted by a_i, ω_i, and n_i, $i = 1, \ldots, P$, respectively. Applying $\{H_n\}_{n=1}^{L}$ to this modulated sinusoid yields a new set of amplitudes $\{\bar{a}_i\}_{i=1}^{P}$,

$$\bar{a}_i = H_{n_i}\left(\frac{2\pi}{N} \cdot \left[\frac{\omega_i N}{2\pi}\right]\right) \cdot a_i, \tag{11}$$

where $[\cdot]$ denotes the nearest integer function, which in turn gives a multiplicative change in the sinusoidal energy, which we call the *vocality score* of the modulated sinusoid:

$$V = (\bar{a}_1^2 + \cdots + \bar{a}_P^2)/(a_1^2 + \cdots + a_P^2). \tag{12}$$

An instrumental sinusoid is substantially attenuated by the filters, which implies a small value of V, while a vocal sinusoid has a value of V that is close to unity because, ideally, the filters should not attenuate vocal sinusoids. Consequently, a threshold, `VT`, can be set for the vocality decision, so that if V exceeds the threshold, the modulated sinusoid can be considered to be only slightly attenuated by $\{H_n\}_{n=1}^{L}$, and thus classified as vocal. The modulated sinusoid is deemed instrumental if V does not exceed `VT`. Before fixing the value of `VT` for general vocal melody extraction tasks, a tuning process could be carried out over a validation dataset to choose an optimal value of `VT`. By thresholding vocality scores for all the modulated sinusoids in the input song signal, the adaptive Wiener filtering procedure produces a list of vocal sinusoids, from which the desired pitch information can be extracted. As an illustration, Fig. 4 shows a high-vocality subset of the modulated sinusoids plotted in Fig. 3.

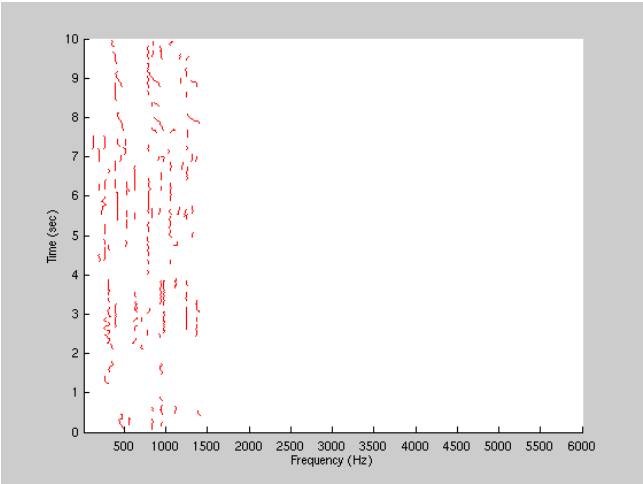

Fig. 4. Vocal sinusoids extracted from a 10-second song clip. Here in adaptive Wiener filtering, the signal is represented by power spectra bandlimited to 1.4 kHz, and modulated sinusoids distributed above that frequency are unconditionally excluded from this collection of vocal sinusoids.

V. Pitch Extraction

The procedure described in this section determines for the input song signal a vocal pitch sequence $\{p_n\}_{n=1}^{L}$, where L is the number of frames in the signal. If frame n is voiced, p_n should be assigned the fundamental frequency of the singing voice at the frame; otherwise, p_n should indicate the unvoicedness with a non-positive constant `NV`. The rest of this section examines the procedure in the order of processing.

A. Grouping Modulated Sinusoids

The pitch extraction procedure starts by grouping the vocal sinusoids into one or more harmonic series [16]. (Such sinusoids are identified by the procedure defined in Section IV.) This is achieved by enumerating seed sinusoids and collecting, for each seed, "contemporary" sinusoids that are harmonically related to it. (A "contemporary list" is generated for each seed sinusoid from a frame-number index of all the vocal sinusoids.) Each seed sinusoid may be cast in at most four harmonic roles: the fundamental, the second partial, the third, and the fourth. It is cast only in the roles that respectively imply average fundamental frequencies within the vocal range. With the seed in a particular harmonic role, only contemporary sinusoids that satisfy the following conditions are included as member sinusoids of the harmonic series associated with the seed-role combination:

- Each member sinusoid must overlap markedly in time with the seed, i.e.,

$$|\mathcal{M} \cap \mathcal{S}| > 0.5|\mathcal{M}|, \tag{13}$$

  where $\mathcal{M}$ and $\mathcal{S}$ denote the sets of frames traversed by the member and the seed, respectively;
- Each member sinusoid must stick to one of the following harmonic roles throughout the overlap: the fundamental, the second partial, ..., and the eighth partial. In other words, the frequency of the member must keep a fixed and appropriate ratio to that of the seed. For example, if the seed plays the role of the second partial and the frequency of a candidate sinusoid remains 1.5 times that of the seed throughout their common duration, then the candidate is acceptable as the third partial in the harmonic series; and
- To avoid octave errors, we require that the member sinusoids as a whole not exhibit significant concentration of their trajectories on the even partials. Such concentration implies mistaking the fundamental of the true pitch for the second partial of the perfect octave below the true pitch. The degree of the concentration is evaluated by dividing the average trajectory length of the even partials by that of the odd partials. Any missing partial is assigned a zero length and counted toward the average. Similarly, the members must also come with a small quotient of dividing the average length of the third and sixth partials by that of the other partials, in order not to erroneously

report the perfect twelfth under the true pitch.

At this stage, we have as many sinusoidal groups as there are seed-role combinations enumerated, but many of them may be redundant, in the sense that one group may be subset of another group. To resolve the redundancy, the grouping procedure eliminates any group that is subset of another group, which is accelerated by sorting the groups by energy in advance. As an illustration, Fig. 5 shows the result of grouping

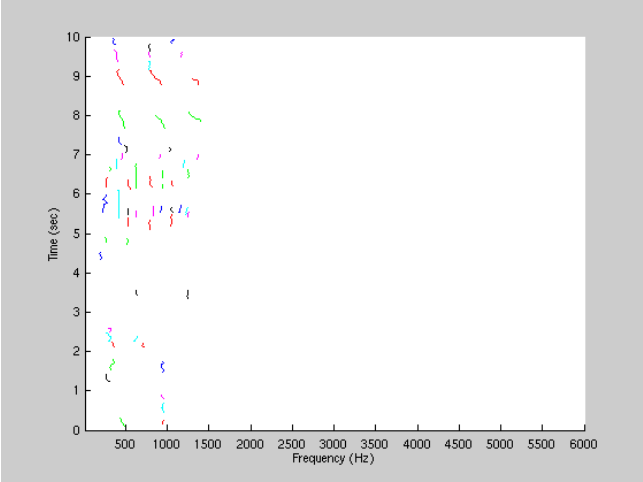

Fig. 5. Vocal harmonic series extracted from a 10-second song clip. Modulated sinusoids within the same harmonic series are shown in the same color.

the vocal sinusoids shown in Fig. 4.

B. Polyphony Resolution

To ensure that the vocal sinusoids represent a valid solo voice, any spurious polyphony assumed by two sinusoidal groups is resolved by pruning the lower-power group. The polyphony can result by accident from a temporal overlap between a group of vocal sinusoids and a group of instrumental sinusoids with high vocality scores. If not fixed, the overlap would contradict the monophony of the solo voice, leaving ambiguities in pitch estimation.

Specifically, for any two temporally overlapping groups, we compute the average power of each over their common frames, which gives two power values, say, P_1 and P_2. While keeping the group with the higher power value, the procedure prunes the other group, by masking its overlap segment and any non-overlap and "weak" segment. The lower-power group is considered weak over a non-overlap range of frames if its average power over the frames is under $\sqrt{P_1 P_2}$, the geometric average of the two power values computed for the overlap across the two groups. There are three cases to the masking:

- If the lower-power group is weak both before and after the overlap, i.e., its instantaneous power does not vary a lot over its duration, the procedure masks the group in its entirety, assuming the group to be instrumental;
- If the lower-power group is strong either only before or only after the overlap, the group is simply cropped to the only strong segment. Here it is assumed that the segment is vocal, the rest of the group is instrumental, and they were joined by spectral collisions in sinusoidal modeling; and
- If the lower-power group is strong both before and after the overlap, we truncate the group to whichever of the two strong segments is longer.

For example, in the top pane of Fig. 6, both the green group

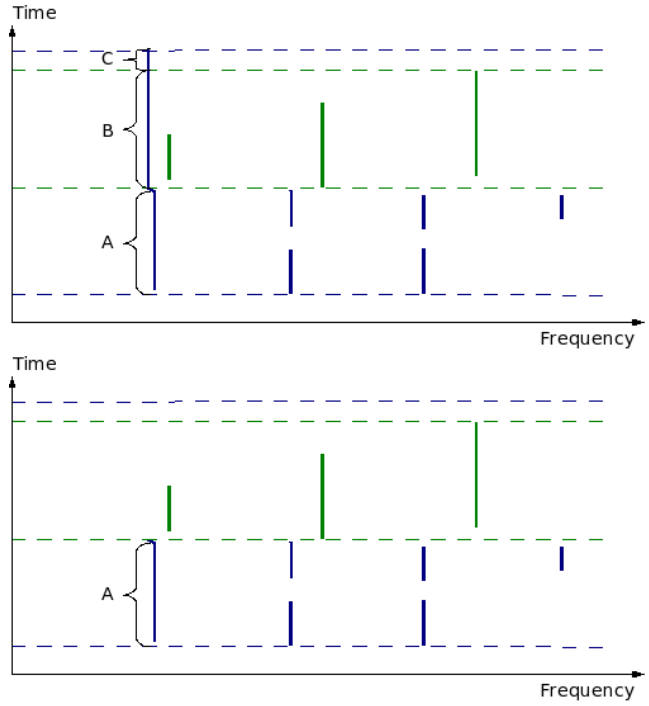

Fig. 6. An example of polyphony resolution. **Top:** Two groups of modulated sinusoids that imply polyphony of the sound. Segment B is polyphonic. **Bottom:** Two groups of modulated sinusoids that represent a monophonic sound.

and the blue group traverse the time segment B. Suppose that within segment B, the average power of the green group is higher than that of the blue group, and that the blue group is strong only at segment A. The result of resolving this polyphony is represented in the bottom pane of Fig. 6, where we can see that segments B and C of the blue group have been removed.

C. Voicing Detection and Vocal Pitch Estimation

Finally, to detect voicing for frame n, the procedure checks the frame spans of all the grouped vocal sinusoids. If none of the vocal sinusoids has a frame span covering frame n, then p_n is set to `NV`; otherwise, a pitch value is computed from the relevant sinusoid(s) and assigned to p_n.

Note that, to compute the vocal pitch for a voiced frame, each vocal sinusoid that sounds at the frame implies a separate fundamental frequency estimate for the frame. This estimate can be computed by dividing the sinusoidal frequency by the harmonic role of the modulated sinusoid, so that more than one fundamental frequency estimate may be available for each voiced frame. The procedure computes the pitch by

averaging as many fundamental estimates as there are active vocal sinusoids at the frame. This method for pitch estimation works, even when the fundamental sinusoid is missing due to errors in earlier processing. See Fig. 7 for the pitch sequence

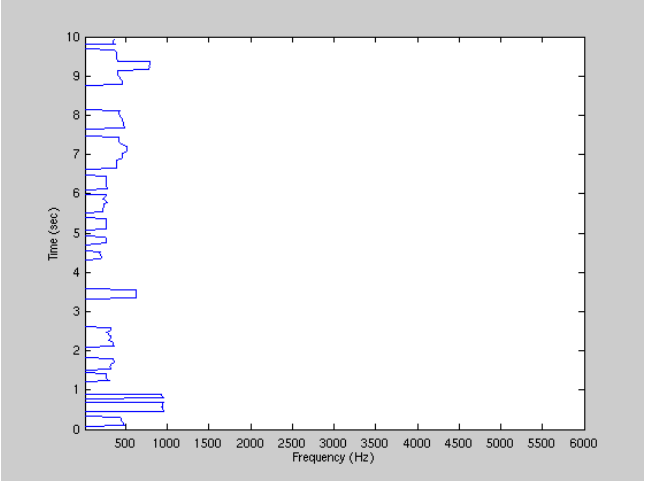

Fig. 7. Vocal melody extracted from a 10-second song clip.

estimated from the grouped sinusoids depicted in Fig. 5.

VI. Experiments

We start the description of our experiments by giving implementation details in Section VI-A. The experiments on melody extraction are documented in Section VI-B, followed by additional experiments dedicated to the application to song retrieval in Section VI-C.

A. Implementation Details

In order to train a male-artist V-GMM, a female-artist V-GMM, and the A-GMM, we collected 58 recordings from 29 male and 29 female artists. The recordings are in the karaoke format, which consists of two audio channels saved in an ordinary stereo-audio file, where the two-channel signal is made up of a stereo accompaniment overlaid with a right-channel singing voice. Such a format allows a recording to be played back as a stereo song (except that the voice has been completely panned to the right), or as a monaural accompaniment by muting the right channel. To train the V-GMMs, we extracted from the 58 recordings approximately 68 minutes of clean singing voice signals in the STFT domain by subtracting the accompaniment from the song, i.e., by setting to zero particular right-channel frequency components that have a magnitude comparable to (within a threshold multiple of) that of their left-channel counterparts. To verify the quality of the resulting 58 voice signals, one of the authors listened to all of them, hearing almost no instrumental sound, although he did exclude from the training set a number of other karaoke files for which such clean singing voice could not be obtained in this way. Also from the 58 recordings, approximately 68 minutes of purely instrumental accompaniment was taken to train the A-GMM. With a frame rate of 21.5 Hz, approximately 100,000 power spectra were extracted from the raw data for each GMM. All GMMs were defined by multivariate Gaussians with diagonal covariance matrices.

In our implementation of the proposed method, all audio data are sampled at 11,025 Hz. In the voice estimation procedure defined in Section IV, to compute the short-time power spectra, STFTs were performed using 93-ms non-overlapping Hamming windows and 1,024-point FFT. To experiment with various resolutions and bandlimits of the power spectra, we convert a 513-bin power spectrum with a 5.5 kHz bandlimit, into an n_b-bin spectrum with a bandlimit of f_c kHz, by summing the power over every $\frac{512 f_c}{5.5(n_b-1)}$ bins in the original spectrum, to give the power in each of the n_b bins in the new spectrum. In addition, sinusoidal modeling was implemented according to [2] with 93-ms Hann windows overlapping by 75% (without multi-resolution analysis), where the high overlap ratio is essential to adequate tracking of spectral peaks.

B. Melody Extraction

To verify the effectiveness of the proposed method, we tested our software implementation of the method on 10 70-second audio excerpts, taken from commercially released recordings of popular Chinese songs, and covering voices of five male and five female artists. As in collecting training data, the test recordings are in the karaoke format, which makes it possible to obtain unaccompanied voice excerpts for the ground truth, as well as accompanied song excerpts for the test, as described in Section VI-A. A ground-truth melody was computed from each voice excerpt by extracting a pitch sequence from all modulated sinusoids as in Section V, for experimental convenience, although a better approach would be using another monophonic pitch estimator. Furthermore, in this test we evaluated our method by the recall of vocal pitch, which is defined by the percentage of truly voiced frames whose estimated pitch falls within one quartertone of the ground-truth pitch.

To see the effect of parameter tuning on the recall, we conducted multiple experiments by varying 1) the number of frequency bins, `NB`, for representing each power spectrum, 2) the bandlimit of each power spectrum, `FC`, 3) the number of components in each V-GMM, `KV`, 4) the number of components in the A-GMM, `KA`, and 5) the vocality threshold `VT`. As shown in Fig. 8, a bandlimit of 5.5 kHz comes with a low frequency resolution and significantly degrades the recall; otherwise, the recall appears quite robust to the number of bins. In addition, a closer look at Fig. 8 reveals that the effect of the bandlimit as decoupled from that of the frequency resolution, while in part representing the completeness of information, seems to be dominated by the curse of dimensionality, in that the 2.8-kHz, 65-bin recall is lower than the 1.4-kHz, 33-bin recall, and the 5.5-kHz, 129-bin recall is even lower. The recalls obtained by varying the number of components in each of the male-artist and the female-artist voice GMMs are listed in Table I, where an increase in the number of V-GMM components is consistently observed to raise the recall. However, it seems unnecessary to set the number of A-GMM

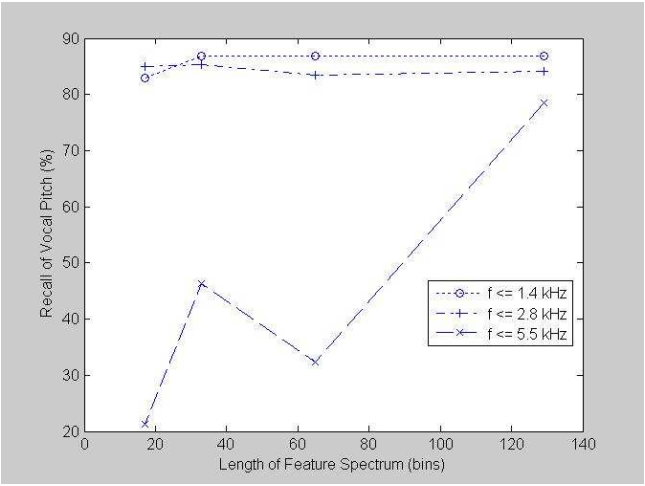

Fig. 8. Vocal pitch recalls with various lengths and bandlimits of the feature spectrum. KV $= 32$; KA $= 1024$; VT $= 0.25$.

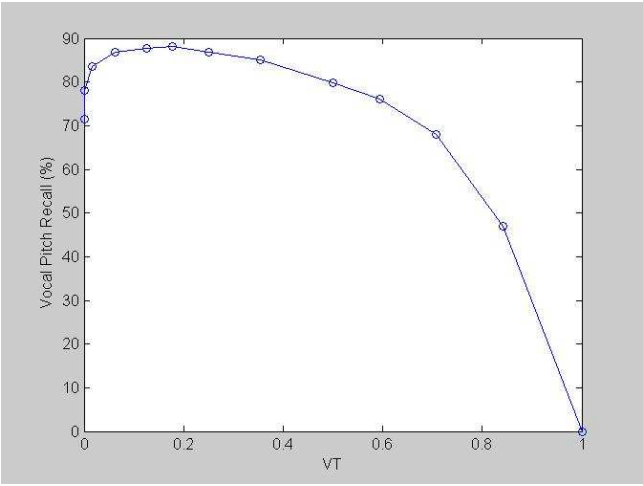

Fig. 9. Vocal pitch recalls with various vocality thresholds. NB $= 129$; FC $=$ 1.4 kHz; KV $= 32$; KA $= 1024$.

TABLE I
VOCAL PITCH RECALLS WITH VARIOUS NUMBERS OF COMPONENTS IN EACH OF THE MALE-ARTIST AND THE FEMALE-ARTIST VOICE GMMS. NB $= 129$; FC $= 1.4$ KHZ; KA $= 1024$; VT $= 0.25$.

#component	32	64	128
Recall (%)	86.83	89.68	91.64

components to a value as large as 1,024, as evidenced in Table II. Finally, the effect of VT is illustrated in Fig. 9, where

TABLE II
VOCAL PITCH RECALLS WITH VARIOUS NUMBERS OF COMPONENTS IN THE ACCOMPANIMENT GMM. NB $= 129$; FC $= 1.4$ KHZ; KV $= 32$; VT $= 0.25$.

#component	1024	512	256	128
Recall (%)	86.83	87.0	87.6	87.31

we can see that the optimum recall 88.03% is achieved by a value of the vocality threshold around 0.1768. In contrast to this optimum, the zero threshold leads to the reduced recall at 71.57% by allowing all modulated sinusoids to enter the stage of pitch extraction regardless of the vocality. In such a degenerative system, high-power instrumental sinusoids could usually be selected for pitch extraction, masking lower-power vocal sinusoids according to the rules set forth in Section V-B for polyphony resolution. As VT increases, the instrumental, or low-vocality, sinusoids tend to be pruned before pitch extraction, which renders the extracted melody more reliable and explains the ascending trend on the left of the optimum in Fig. 9. On the other hand, when the threshold is raised above the vocality scores of some vocal sinusoids, the increase in the threshold demonstrates the negative effect of eliminating more vocal sinusoids, giving a decrease in recall. At the extremum, with the unity threshold, no modulated sinusoid enters the stage of pitch extraction, and all voiced frames are taken as unvoiced, receiving a zero recall.

C. Song Retrieval

To confirm the efficacy of the proposed method in indexing a song database, we additionally carried out a query-by-singing experiment QBS-1 with the song database indexed by the proposed method. This experiment differs from that in [17], mainly in that while segmentation labels are manually specified for each song in [17], they are automatically produced by the proposed method as a sequence of voicing onsets in this work. The labels serve as candidate times in each song at which the query melody may be spotted. In [17], melodies are extracted from songs by subharmonic summation without voicing detection. In addition, we use the methods given in Sections III and V to transcribe query voices.

For this set of experiments, the feature spectrum for GMM training is represented by 129 frequency bins with a 1.4 kHz bandlimit. For each V-GMM, the number of Gaussians mixed is set to 32; for A-GMM, to 1,024. The vocality threshold VT defined in Section IV is set to 0.25.

We made the following modifications to the method proposed in [17] for melody matching:

- Pitch is quantized in 10-cent resolution. The average pitch deviation of each melody from the standard tuning system is computed and subtracted from the melody. To be specific, we encode the deviation of each pitch from the standard tuning system by a point on the unit circle. The angle (in radians) of the point is given by

$$\theta = \frac{\pi}{50}\left(p - \lfloor p \rfloor\right), \tag{14}$$

  where p is the pitch in cents, and $\lfloor p \rfloor$ denotes the highest standard pitch below p. A point at 0 radian represents zero deviation, while a point at π radians represents the maximum, 50 cents. For each melody, taking the two-dimensional centroid of the points accumulated on the

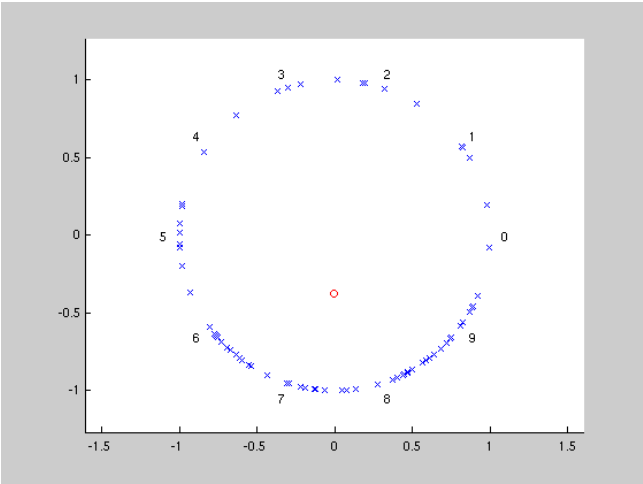

Fig. 10. Illustrating the computation of the average pitch deviation. The blue x-marks represent deviations of pitches from the standard tuning system in a query melody. The 10 labels uniformly spaced around the circle are values of pitch difference in the unit of 10 cents, marking the scale on which the angle of each point encodes a pitch deviation. To facilitate visualization, the points were plotted for un-quantized pitch values. The red circle represents the centroid, from which we find that the person who sang this melody intended a tuning system that is lower than the standard by nearly 25 cents.

circle (see Fig. 10), and converting the angle of the centroid back to the pitch difference (multiplying the angle in $[0, 2\pi]$ by $\frac{50}{\pi}$), gives the average pitch deviation. This alignment is of particular importance for resolving a query melody sung without first listening to a standard pitch;

- Each voicing gap is filled with the pitch immediately preceding the gap, so that melodies can be matched by dynamic time warping (DTW) without discontinuities. For a pitch sequence that begins with silence, the beginning gap is filled with the first voiced pitch value that follows. In case the pitch sequence is silent throughout its duration, we fill it with a constant middle C;
- Pitch is wrapped into chroma to disregard octave errors occurring in melody extraction. Accordingly, instead of computing the average pitch value for each melody [17], the matching procedure computes the average chroma value. Specifically, we map each chroma value onto the unit circle, with 0 radian representing the C pitch class and π radians representing the G-flat. Taking the two-dimensional centroid of the points accumulated on the circle, and converting the angle of the centroid back to the chroma value, gives the average. Furthermore, the distance between two chroma values x and y (cents) is given by

$$\min\{|x-y|, 1200-|x-y|\}, \tag{15}$$

which never exceeds 600 cents; and

- Multiple-level data abstraction (MLDA) is adopted from [18] to speed up the matching.

The test results of QBS-1 are listed in the first row of Table III, where the automatic segmentation is made possible by the proposed method, which generates 28,771 segmentation labels (voicing onsets) over the 95-song database. In the second, third and fourth rows of Table III are results of three control experiments. Replacing the automatic segmentation in QBS-1 with a 771-label manual segmentation gives QBS-2. False voicings in the automatic segmentation only lowered the top-1 accuracy by 2.2%, possibly because the size of database is not realistically large. Experiment QBS-3 is identical to QBS-1 except that 1) the proposed method is replaced by the melody extractor in [17]; and 2) voicing gaps in the 90 queries are bypassed by zero local distances in DTW as in [17], rather than filled with a constant pitch. Substituting the 771-label manual segmentation for the automatic segmentation in QBS-3 gives the experiment QBS-4, where the lower accuracies as compared with QBS-2 verify the benefit of the proposed melody extractor. Also notice that the segmentation automation in QBS-3 gave rise to an 8.9% drop in top-1 accuracy, much deeper than the 2.2% drop experienced in QBS-1, as may be attributed to the lower melodic ambiguity exhibited by the proposed method. Finally, since no advantage of the proposed method is taken in QBS-4, comparing the results between QBS-1 and QBS-4 reveals that the proposed method mainly acts to automate the indexing, or more specifically, the segmentation of the song database, without degrading the retrieval accuracy.

TABLE III
RETRIEVAL EXPERIMENT RESULTS.

Experiment	Song Segmentation	Song Accuracy (%)		
		Top 1	Top 3	Top 10
QBS-1	Automatic	58.9	65.6	75.6
QBS-2	Manual	61.1	70.0	76.7
QBS-3	Automatic	47.8	58.9	71.1
QBS-4	Manual	56.7	64.4	74.4

VII. CONCLUSION

We have presented a system for extracting a vocal melody from the signal of an accompanied singing performance. The extraction exhibits sensitivity to vocality, which originates from a voice estimator through a novel transformation. The feasibility of the method in indexing a song database for a song retrieval application has been verified by experiments on real popular song data and real sung queries.

One limitation of our approach is the assumption that the accompaniment to songs is purely instrumental. In the future, it would be interesting to consider the timbral difference between a solo voice and a choral voice, so that if there are any choral voices in the accompaniment, the solo voice can be distinguished from them. It is also worth further research to reduce the number of segments in each song from that of all the voicing onsets given by the proposed method, in pursuit of shorter search time.

ACKNOWLEDGMENT

The authors would like to thank Dr. Yi-Wen Liu for his help with software development. They are also grateful for

the comments from the anonymous reviewers, which were valuable in enhancing the quality of this paper.

REFERENCES

[1] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 34, no. 4, pp. 744–754, Aug. 1986.

[2] S. N. Levine, *Audio Representations for Data Compression and Compressed Domain Processing*, Ph.D. thesis, Stanford University, 1998.

[3] A. Ozerov, P. Philippe, R. Gribonval, and F. Bimbot, "One microphone singing voice separation using source-adapted models," in *WASPAA*, 2005.

[4] M. Goto and S. Hayamizu, "A real-time music scene description system: Detecting melody and bass lines in audio signals," in *Working Notes of the IJCAI-99 Workshop on Computational Auditory Scene Analysis*, 1999.

[5] M. Goto, "PreFEst: A predominant-F0 estimation method for polyphonic musical audio signals," in *Proc. 1st Annual Music Information Retrieval Evaluation Exchange*, 2005.

[6] M. Marolt, "On finding melodic lines in audio recordings," in *Proc. 7th Int. Conference on Digital Audio Effects*, 2004.

[7] R. P. Paiva, T. Mendes, and A. Cardoso, "On the detection of melody notes in polyphonic audio," in *ISMIR*, 2005.

[8] Y. Li and D. Wang, "Detecting pitch of singing voice in polyphonic audio," in *ICASSP*, 2005.

[9] D. P.W. Ellis and G. E. Poliner, "Classification-based melody transcription," *Mach. Learn.*, 2006.

[10] J. Eggink and G. J. Brown, "Extracting melody lines from complex audio," in *ISMIR*, 2004.

[11] H. Fujihara, T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "F0 estimation method for singing voice in polyphonic audio signal based on statistical vocal model and Viterbi search," in *ICASSP*, 2006.

[12] M. Lagrange, L.G. Martins, J. Murdoch, and G. Tzanetakis, "Normalized cuts for predominant melodic source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 278–290, Feb. 2008.

[13] I. Potamitis and A. Ozerov, "Single channel source separation using static and dynamic features in the power domain," in *EUSIPCO, 16th European Signal Processing Conference*, 2008.

[14] Simon Haykin, *Communication systems*, Wiley, 3rd edition, 1994.

[15] Henry Stark and John W. Woods, *Probability, random processes, and estimation theory for engineers*, Prentice Hall, 2nd edition, 1994.

[16] D. PW Ellis, "A computer implementation of psychoacoustic grouping rules," in *ICPR*, 1994.

[17] Hung-Ming Yu, Wei-Ho Tsai, and Hsin-Min Wang, "A query-by-singing technique for retrieving polyphonic objects of popular music," in *Proc. Asian Information Retrieval Symposium*, 2005.

[18] Hung-Ming Yu, Wei-Ho Tsai, and Hsin-Min Wang, "A music retrieval system based on query-by-singing for karaoke jukebox," in *Proc. Asian Information Retrieval Symposium*, 2006.