Extracting Vocal Melody from Karaoke Music Audio

Yongwei Zhu, Sheng Gao
Institute for Infocomm Research, A-STAR, Singapore
ywzhu@i2r.a-star.edu.sg

Abstract

Extracting the melody from polyphonic musical audio is a nontrivial research problem. This paper presents an approach for vocal melody extraction from dual channel Karaoke music audio. The extracted melody corresponds to the singing voice in the original performance channel, which can then be used for melody-based music retrieval. In the proposed technique, audio signals from both the accompaniment channel and the original performance channel are analyzed. The note partials are firstly extracted from the signal, which is represented in constant-Q transform frequency domain. Then the volume balance between the two channels is estimated based on signal approximation in the sub-bands. Finally the pitch corresponding to the singing voice is identified based on the note partial differences between the two channels. The extracted melody is represented as a sequence of pitch values. This technique assumes that the two channels have similar accompaniment instrument performance except for the singing voices.

Experimental result on 40 Karaoke music audios has shown the performance of the proposed technique. The pitch extraction rate is above 70% and melody retrieval accuracy in an 800-tune-database is 90%.

1. Introduction

Music information retrieval has become an increasingly active research area. Many techniques have been proposed for retrieving symbolic music using the melody, such as query-by-humming. There has been relatively less work done for melody-based retrieval of acoustic music. This is largely due to the difficulty of extracting or identifying the melody in polyphonic music signals. Particularly of interests are the melodies of the singing voices in pop songs, since people prefer to search a song based on the tune of the singer's voices.

There have been a few works on extracting melody from acoustic music. Malik and Goto [1, 2] proposed techniques to extract the predominant pitch from audio signals, whereas the extracted pitch may be over fragmented by the overlapping of multiple instruments. [3] proposed a proposed a Gaussian mixture model for extraction of melody line, by which melody fragments of different instrument with overlapping can be extracted. However, the result has been shown only for a single example.

Karaoke is a very popular form of musical entertainment. Plenty of music video data are available either in VCD or DVD formats. Karaoke music video contains two dedicated audio channels: one is the musical audio with original singer's voices and the other one only the accompaniment without singing voices. Usually the accompaniment is very similar to that in the original performance channel except for the singing voice. From the best of our knowledge, there has been no existing work done on extracting vocal melody from the dual audio channels of Karaoke music videos. We proposed a novel technique for extracting vocal melody, which is represented as a sequence of pitch values. The extracted melody can be used for music retrieval by similarity based sequence matching.

The rest of this paper is organized as follows: section 2 presents the overview of the proposed technique. Section 3 presents note partials extraction. Section 4 presents estimation of volume balance of the two channels. Section 5 presents the singing pitch extraction. Section 6 presents the experimental results and section 7 presents the conclusion.

2. Overview

In this work, we are targeting at a music retrieval system that can accept melody queries. The system consists of a music indexing subsystem and a music querying subsystem, as illustrated in figure 1. The music indexing subsystem extracts the features that characterize the vocal melody in the audio signals. The input channel A stands for the audio signal for the accompaniment, and the input channel B stands for the original performance that is singing plus

0-7803-9332-5/05/$20.00 ©2005 IEEE

accompaniment. Music tempos and beat onsets are detected from channel A [4], so that the melody can be synchronized when doing melody matching. The repeating portions of the music are also detected based on channel A [5]. The detected portions usually correspond to verses or choruses of a song. The melody matching can be done only for the repeating portion. Melody extraction of the singing voice is based on both channel A and channel B. In the music querying subsystem, melody matching is conducted between the extracted melody and a reference melody. This paper focuses on the melody extraction component of the system, while tempo and repeating detection are based on previous techniques. Melody matching is discussed in the experiments on evaluation of the melody extraction technique.

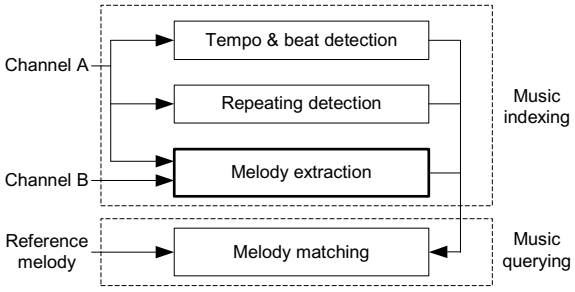

Figure 1. The structure of melody-based Karaoke music retrieval system

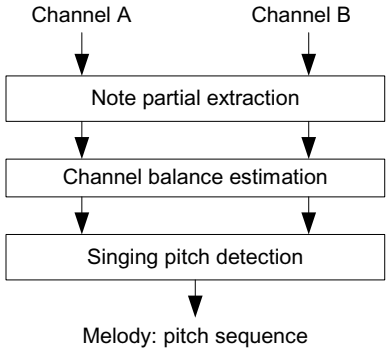

Figure 2. Melody extraction from dual channel Karaoke music

Figure 2 shows the steps in the proposed vocal melody extraction technique. In the first step, the audio signals in the two channels are represented in the frequency domain where each component corresponds to a note partial. The note partials are closely related to music notes. The volume levels of the two channels may not be identical, thus it is necessary to estimate the volume balance between the accompaniment performances in the two channels, which is the second step. Finally, the singing pitch is detected based on the differences of the two channels. The extracted melody is represented as a sequence of pitch values. Details of each of the steps are presented in the remaining sections of this paper.

3. Note partials extraction

The vocal melody of our interests is ideally a sequence of music notes with particular pitch and time duration values. So we have chosen a frequency domain signal representation that is closely related to music notes: the Constant-Q Transform. The constant-Q transform (CQT) [6] converts the audio signal to the frequency domain with constant bandwidth to frequency ratio. Thus the component spacing is linear to pitch intervals defined in music theory, and can be measured in semitones.

A note produced either by an instrument or voice usually has a number of harmonic frequency components, or note partials. The lower frequency partials typically have higher energy than the others. Thus we are interested in extracting the lower frequency partials including the fundamental. These note partials are usually energy peaks in the CQT spectrum.

In our approach for melody extraction, we desire precise note partial extraction and avoid interference between notes adjacent in pitch, e.g. 1 semitone away. This is done by estimating the tuning pitch of the piece of music and extracting those CQT components that are in tune with the tuning pitch. Details of tuning pitch estimation and note partial extraction are presented in an approach for detecting keys from musical signals [7].

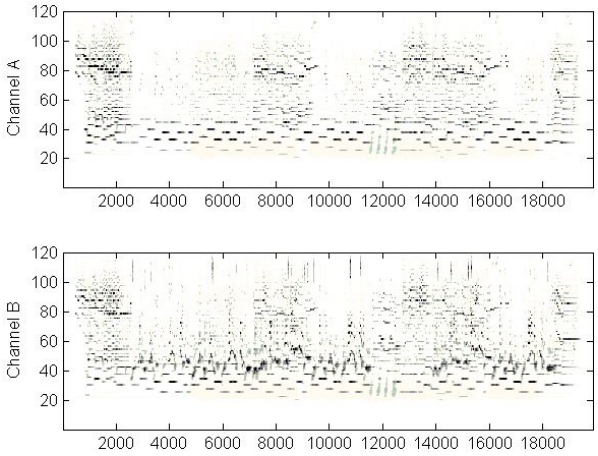

Figure 3. Note partials in CQT spectrogram

The output of note partial extraction of the 2 channels of a music example is shown in figure 3. We considered the pitch range of 10 octaves, starting from 27.5 Hz. There are totally 120 note partials for the 10

octaves, and each of the note partials can be translated to a pitch name, e.g. middle C.

It can be seen that note partials from channel A is quite similar to those from channel B. The difference is mainly caused by the vocal in the channel B. It should be pointed out that the volume level ratio between the two channels is unknown, and we observed that the volume of the instrumental accompaniment in channel B is usually lower than that in channel A. We need to estimate the volume ratio (balance) between the two channels before the pitch of the singing voice can be detected. Volume balance estimation is presented in the next section.

4. Channel volume balance estimation

From our observation, the volumes of the two channels are adjusted to different levels in the production of the Karaoke music recordings. Although the level is fixed for all the time for a particular channel, the volume ratio between the two channels is unknown. We proposed a technique to automatically estimate the volume ratio between the two channels.

The volume balance estimation is based on the best approximation of a note partial sequence in channel A to the same note partial sequence in channel B, which is presented as follows.

The note partial in channel A is denoted as A(*p,t*), where *p* is the pitch and *t* is the time window index. Similarly the note partial in channel B is denoted as B(*p,t*). The approximation error of A to B for pitch *p* is denoted as E(*p,r*), where *r* is the volume ratio.

$$E(p,r)=\sum_{t=1}^{T}\left|r\times A(p,t)-B(p,t)\right| \quad (1)$$

where T is the total time length of the signal. The best ratio *R* between B to A at pitch *p* is then denoted as

$$R(p)=\arg\min_{r\in[0,2]} E(p,r) \quad (2)$$

and the corresponding approximation noise to signal ratio is computed as

$$NSR(p,R(p))=\frac{\sum_{t=1}^{T}\left|R(p)\times A(p,t)-B(p,t)\right|}{\sum_{t=1}^{T}\left|A(p,t)\right|} \quad (3)$$

The final volume balance estimation result is then based on the note sequence of pitch $\hat{P}$, which has minimal noise to signal ratio (eq.4). And the corresponding volume ratio $\hat{R}$ is the final result (eq.5).

$$\hat{P}=\arg\min_{p\in[1,120]} NSR(p,R(p)) \quad (4)$$

$$\hat{R}=R(\hat{p}) \quad (5)$$

Figure 4 illustrates the volume balance estimation result for the example music. The top subfigure shows the sorted minimal noise to signal ratio for all the 120 pitches under consideration. The middle subfigure shows the corresponding volume ratio between channel B to A. The bottom subfigure shows the corresponding (unsorted) pitch that has the minimal noise to signal ratio.

It can be seen that the best approximation has noise-to-signal ratio below 0.05 (top). It can also be observed that the pitches with small noise-to-signal ratio are around 30 (bottom). These pitches are typically in the pitch range of low pitch instruments, and lower than the pitch of human singing voice. For this example the final volume ratio is 0.49 (middle).

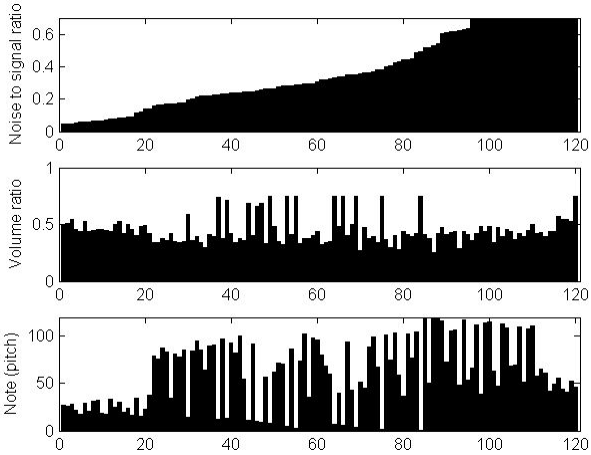

Figure 4. Channel volume ratio estimation

5. Singing pitch extraction

The spectral difference between the two channels is obtained by subtracting the CQT spectrogram of channel B with the CQT spectrogram of channel A multiplied by the estimated volume ratio.

$$D(p,t)=B(p,t)-A(p,t)\times\hat{R} \quad (6)$$

The spectral difference $D(p,t)$ characterizes mainly the pitch content of the singing voice. The pitch of the singing voice at any particular time is detected based on the peaks in $D(p,t)$. In the proposed approach, a peak $\hat{p}$ in $D(p,t)$ is claimed as singing pitch, if the following conditions are met.

(1) the peak $\hat{p}$ has highest energy in the spectrum for that particular time *t*; (2) the component just 1 octave above the peak, $\hat{p}$ +12 is a local peak in the spectrum; (3) the claimed pitch has a temporal continuity, such as accumulated length larger than a time duration (e.g. ¼ beat).

These conditions are based on the observations that the fundamental frequency of a pitch usually has highest energy in the spectrum. And the second harmonic usually also has a high energy value, and is a local peak. The correct pitch should stay stable for a period of time, and the peaks generated by noise usually are isolated. Figure 5(a) illustrates the pitch extracted for a portion of the music signal based on the spectral difference. The corresponding note graph (ground truth obtained from MIDI file) is illustrated in 5(b). The resemblance between the extracted pitch sequences with the ground truth is quite obvious.

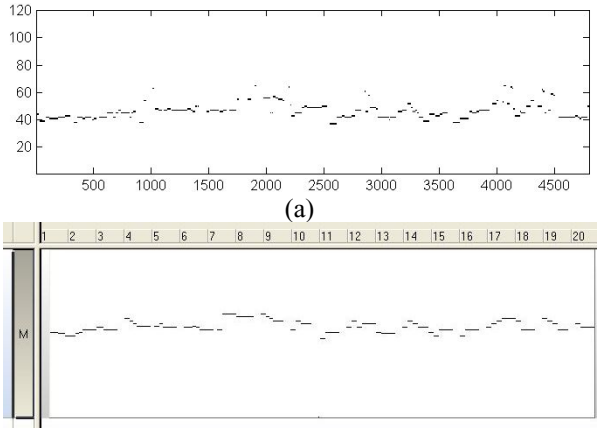

Figure 5. (a) extracted pitch sequence for vocal melody; (b) ground truth MIDI note graph

6. Experiments

We did experiments to evaluate the melody extraction technique. 40 Karaoke music titles are collected from 4 VCDs. MIDI file versions of these popular songs can be easily collected from the Internet.

Retrieval on Karaoke music based MIDI melody (100% precision) cannot reflect the system performance, since the database size (40) is small. Thus we conducted the following 2 experiments on evaluating the proposed method. In the first experiment, we evaluate how many pitch extracted are correct. In the second experiment, we use the extracted melody to query a MIDI melody database with 800 MIDI songs and the retrieval performance can reflect the quality of the extracted vocal melody.

6.1. Pitch detection rate

The pitch detection is evaluated only for the chorus part of each song. And the ground truths are manually extracted from the MIDI song that is obtained from the Internet. The two melodies are aligned in time and pitch manually, and any pitch value in the extracted melody is considered correct if it coincide with the ground truth MIDI note.

For the 40 songs, the mean pitch detection rate is 72.5%, and the variance is 0.0028. We have observed that some pitch errors are due to low level of singing volume and some residual instrumental pitches are picked up. And some errors are octave errors, where the detected pitch is 1 octave higher than the ground truth.

6.2. MIDI file retrieval

We used the chorus portion of the extracted melody to search a MIDI database with 800 songs. The beat positions are utilized in melody matching, so that two melodies under comparison are synchronized time. The similarity between two melodies is measured by the pitch accuracy as mention previously. Pitch shifting (transposition) is not considered in the matching, since the MIDI files are usually authored using the same key with the original music title.

The average retrieval precision is shown in the following table.

Table 1. MIDI files retrieval precisions

Top list size	1	5	10
Precision	0.78	0.88	0.91

7. Conclusion

This paper presents an approach for extracting vocal melody from dual channel Karaoke music audio. The extracted melody is in a form of pitch sequence, and can be used in similarity melody matching for retrieval. Our future works include using video information, such as the lyrics in the video, to assist locating the time and pitch of the singing notes. We would also increase the number of music titles, and investigate query-by-humming style of music retrieval.

8. References

[1] H. Malik, A. Khokhar, and etc, "Predominant pitch contour extraction from audio signals", *Proc. ICME'02.*

[2] M. Goto, "A robust predominant-F0 estimation method for real time detection of melody and bass lines in CD recordings", *Proc. ICASSP 2000.*

[3] M. Marolt, "Gaussian mixture models for extraction of melodic lines from audio recordings", *Proc. ISMIR 2004.*

[4] S. Gao and etc, "A unsupervised learning approach to musical event detection", *Proc. ICME 2004.*

[5] M. Goto, "A chorus selection detecting method for musical audio signals", *Proc. ICASSP 2003.*

[6] J.C. Brown, "Calculation of constant Q spectral transform", *J. Acoust. Soc. Am.*, 89(1):425-434, 1991.

[7] Y. Zhu, M. Kankanhalli, S. Gao, "Music key detection for musical audio", *Proc. MMM 2005.*