A HYBRID APPROACH TO SINGING PITCH EXTRACTION BASED ON TREND ESTIMATION AND HIDDEN MARKOV MODELS

Tzu-Chun Yeh[1], *Ming-Ju Wu*[1], *Jyh-Shing Roger Jang*[1], *Wei-Lun Chang*[2], *and I-Bin Liao*[2]

[1]Computer Science Department, National Tsing Hua University, Hsinchu, Taiwan
{kenshin.yeh, brian.wu, jang}@mirlab.org
[2]Telecommunication Laboratories, Chunghwa Telecom, Taoyuan, Taiwan
{overcomer, snet}@cht.com.tw

ABSTRACT

In this paper, we propose a hybrid method for singing pitch extraction from polyphonic audio music. We have observed several kinds of pitch errors made by a previously proposed algorithm based on trend estimation. We also noticed that other pitch tracking methods tend to have other types of pitch error. Then it becomes intuitive to combine the results of several pitch trackers to achieve a better accuracy. In this paper, we adopt 3 methods as a committee to determine the pitch, including the trend-estimation-based method for forward and backward signals, and training-based HMM method. Experimental results demonstrate that the proposed approach outperforms the best algorithm for the task of audio melody extraction in MIREX 2010.

Index Terms— Audio Melody Extraction, Hidden Markov Model, Trend Estimation, Singing Pitch Extraction

1. INTRODUCTION

The task of singing pitch extraction from polyphonic audio music is one of the most important tasks in audio music analysis. There are quite a few applications that rely on singing pitch extraction as a building block, such as singing voice separation, singer identification, cover song identification, melody-based music retrieval, and so on.

A number of approaches to audio melody extractions have been proposed since 2000. Among them, HMM (hidden Markov models) have been used extensively due to their flexibility. Ryynänen et al. [1] proposed an HMM-based method by utilizing both the acoustic and musicological models to form a singing melody transcription system. Li et al. in [2] proposed a method that first filtering and selecting channels from the spectrogram. Peaks in these selected channels are then extracted as the feature to train HMMs, where the transition probabilities are estimated from the pre-labeled training data. Both studies focus on the contextual relationship between neighboring music notes. However, the concurrent pitches produced by music accompany or chord, which may critically influence the overall vocal pitch estimation, were not dealt with in details. To address this problem, Hsu et al. introduced a method based on trend estimation [3], where the singing voices were enhanced first and the concurrent pitches were eliminated by vibrato and tremolo features. After that, pitch range of each frame was identified by finding a coarse path (pitch trend) within high-magnitude T-F (time-frequency) blocks. The use of such a pitch trend can remove the undesirable concurrent pitches produced by accompanied music. However, we observed the pitch tracking over unvoiced regions are quite unpredictable due to the lack of main vocal. This is likely to lead to errors at the beginning and ending of a voice region.

In the proposed system, we combine the three methods (trend-estimation-based method with forward and backward input signals, HMM-based method) to generate 3 pitch contours. Several pitch merging methods are investigated for producing the final best results. Evaluation shows that the proposed method outperforms the best algorithm [3] (in terms of raw pitch accuracy) for the task of audio melody extraction in MIREX 2010.

The rest of this paper is organized as follows. Section 2 introduces related work, especially trend estimation. Section 3 describes the proposed method. Section 4 gives experimental settings and the corresponding results. Section 5 discusses the conclusions and future work.

2. RELATED WORK

In the proposed hybrid method, trend estimation plays an important role in determining singing pitch since it can suppress instrumental partials and enhance vocal partials in a specific pitch range of a given spectrogram. Here, we use the trend estimation algorithm proposed by Hsu et al. [3] [9]. Fig. 1 shows the schematic diagram of trend estimation, which are briefly explained in the next paragraph.

First of all, a harmonic/percussive sound separation (HPSS) method [4] is applied to suppress the energy produced by harmonic instruments of the input mixture. Then the multi-resolution fast Fourier transform (MR-FFT) [5] is used to extract the sinusoidal partial from the input mixture. In particular, MR-FFT can discard unreliable peaks

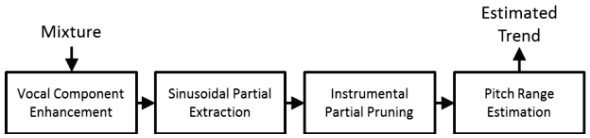

Fig. 1. Schematic diagram of trend estimation [3]

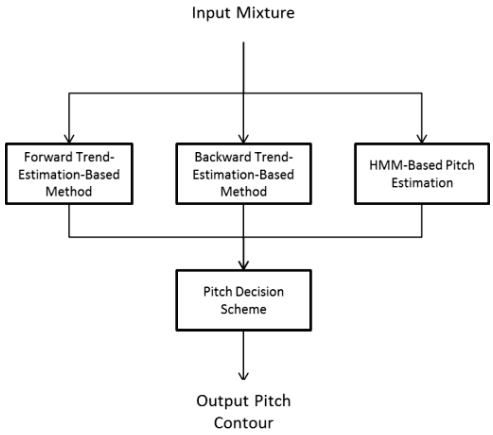

Fig. 2. The basic blocks of the proposed system

not produced by periodic sound. After selecting peaks by their instantaneous frequencies, Hsu's method invokes a grouping method to combine peaks into partials that are likely to be generated by the main vocal and accompanying instruments.

After the partials are obtained, Hsu's method performs instrumental partial pruning based on two features: vibrato and tremolo. These two features were firstly proposed by Regnier et al. [6]. Generally speaking, human voices have both strong vibrato and tremolo [7]. Thus, instrumental partials, which may only have either vibrato or tremolo, can be removed according to this observation. Then the magnitudes of T-F (time-frequency) blocks along a specific time are summed to form a map, and the pitch trend is found by using the down-sampled map. The final pitch contour is found as the optimum path by dynamic programming. Detailed coverage of Hsu's method can be found at [3] [7].

3. SYSTEM OVERVIEW

Fig. 2 shows the basic blocks of the proposed system. First of all, we extract the singing pitch from different methods, including two trend-estimation-based methods and one HMM-based method. These pitch contours are then combined to form the final result. We shall describe the functionalities of each block in Fig. 2 in the following subsections.

3.1. Pitch Tracking based on Forward-Backward Trend Estimation

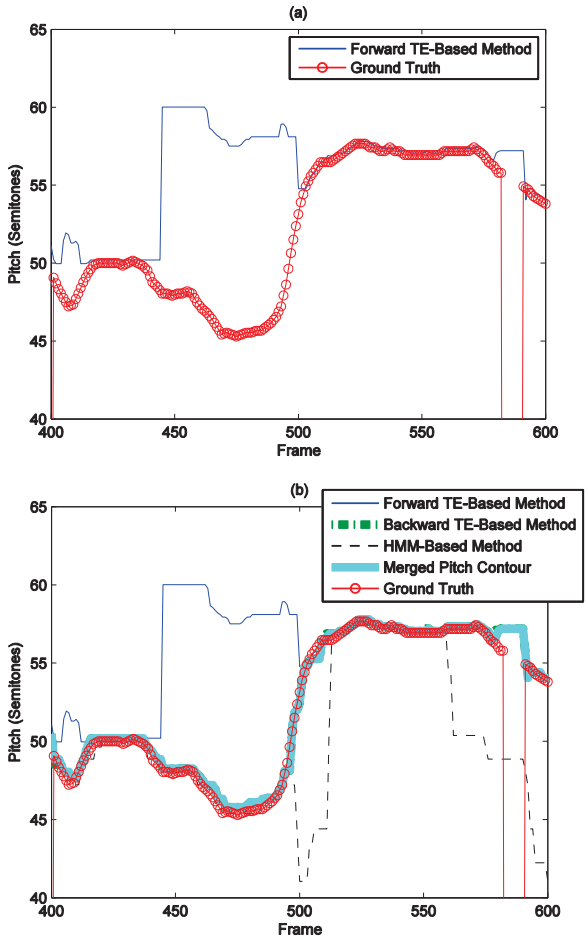

Fig. 3. Illustrations of a typical pitch error and its recovery. (a) The ground-truth pitch contour, and the one generated by the forward trend-estimation-based method. (b) The result of merging pitch contours generated by 3 different methods, which can recover from the pitch error displayed in (a).

After the singing pitch is extracted by Hsu's trend-estimation-based method, we can observe some typical errors. One such example is shown in Fig. 3(a). In particular, we notice that such pitch errors are likely to happen at the beginning of a music phrase. If we reverse the input signals in time axis and send it for pitch tracking, the pitch errors occur elsewhere. As a result, forward and backward (in time) signals are likely to generate complementary results. This observation motivates us to combines pitch contours obtained from forward and backward signals to achieve a better accuracy.

3.2 Pitch Extraction Based On Hidden Markov Models

HMM-based methods have been widely used for audio melody extraction [2] [8]. In this study, we use the maximum 2 values of the NSHS (normalized sub-harmonics summation) map and their locations (frequencies) as the

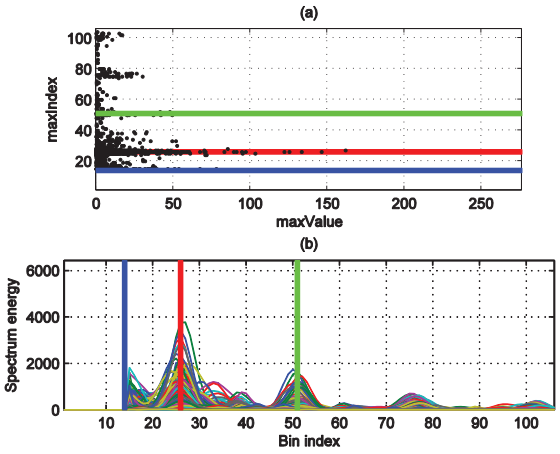

Fig. 4. Illustration of HMM's features of frequency bin 26. (a) Scatter plot of the first two dimensions of the features, including the maximum value of a frame's NSHS and the correspoding index. (b) The corresponding NSHS curves for (a).

features for HMM. Fig. 4(a) illustrates the scatter plot of the first two dimensions of the features, including the maximum value of a frame's NSHS and the corresponding index. Fig. 4(b) shows the corresponding NSHS curves for (a). As can be observed in the Fig. 4, the max values and the corresponding indexes are mostly likely to be around the state bin(indicated as the middle bar in both Fig. 4(a) and Fig. 4(b)). Let the feature vector set be denoted as $V = \{v_0, ..., v_t, ...\}$, our target is to find the most likely state sequence $S = \{S_0, ..., S_t, ...\}$:

$$S = \underset{s}{\operatorname{argmax}} \{p(s_0)\, p(v_0|s_0) \prod_t \{p(v_t|s_t) p(s_t|s_{t-1})\}\} \quad (1)$$

where $p(S_0)$ is the prior probability of s_0, $p(s_t|s_{t-1})$ is the state transition probability from state s_{t-1} to state s_t, and $p(v_t|s_t)$ is the output likelihood of s_t while v_t is the input feature vector. In our implementation, state transition probabilities, and the parameters of output probability density functions are obtained from the training data. We use Gaussian mixture models for the output probability density function of each frequency bin models. The number of Gaussian mixture components is decided to 256 by a 5-fold cross validation on MIR-1k dataset with different number of Gaussian mixture components.

Fig. 3(b) illustrates the resultant 3 pitch contours, including 2 pitch contours obtained from trend-estimation-based method on forward and backward waveforms respectively, and one by HMM-based method. We can see the errors made by these 3 pitch contours are more or less complementary to one another. As a result, we can simply apply a combination scheme to achieve a better accuracy.

3.3 Pitch Combination Scheme

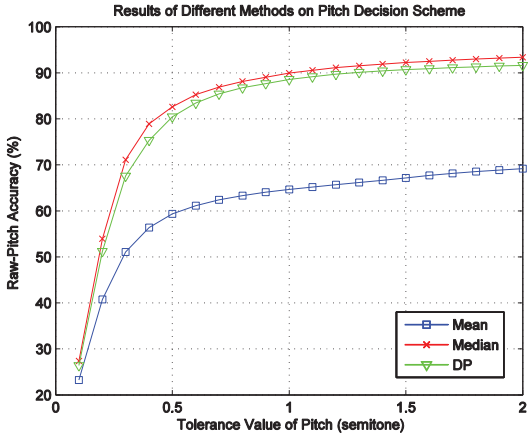

Fig. 5. Evaluation results of different pitch combination methods

After obtaining multiple pitch contours, we need to combine them in an optimal sense. We can arrange these pitch contours into a matrix C, where each row is a pitch contour and each column is possible semitones of a frame. Here we adopt 3 different methods to derive the final pitch contour:

1. Median method: The optimum pitch contour $P_{median}(j)$ at frame j is expressed as:

$$P_{median}(j) = \{\underset{i}{median}(C_{i,j}), j = 1{\sim}M\}, \quad (2)$$

where M is the number of frames.

2. Mean method: The optimum pitch contour $P_{mean}(j)$ at frame j is expressed as:

$$P_{mean}(j) = \{\underset{i}{mean}(C_{i,j}), j = 1{\sim}M\} \quad (3)$$

3. DP-based method: The recurrence function of this method can be formed as follows:

$$D(i,j) = C(i,j) - \min(C_{j-1} - D_{j-1}) \quad (4)$$

$$D(i,1) = 0, \forall i = 1{\sim}N, \quad (5)$$

where N is the number of pitch contour candidates. We can easily obtain the optimal path P_{dp} by the backtracking, which is a common method in dynamic programming.

The evaluation results of the above methods are illustrated in Fig. 5. As can be easily observed, the result of median method is the best among all 3 methods. Thus, the median method will be adopted for the rest of the evaluations presented below.

4. EVALUATION RESULTS

The dataset for our evaluation is MIR-1k [9], which contains 1000 pop-song singing mixture with leading vocals and

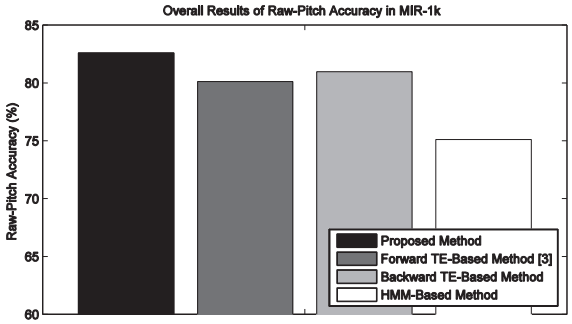

Fig. 6. Overall results on MIR-1k

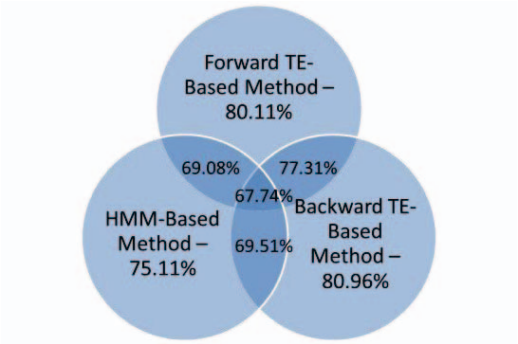

Fig. 7. A Venn diagram of the percentages of the correctly identified pitches by the 3 methods

music accompanies. A 5-fold singer-specific cross validation is then employed to obtain the average raw-pitch accuracy for the proposed method.

Fig. 6 shows the evaluation results of raw-pitch accuracy with a pitch tolerance of 0.5 semitones. The proposed method outperforms Hsu's trend-estimation-based method (the best method in the MIREX 2010 competition [3]) and the HMM-based method. The proposed method achieves an accuracy of 82.60%, while the trend-estimation-based method is 80.11%, indicating an error reduction of 12.53% for MIR-1k. This serves to demonstrate that the error in pitch contours are complementary, and the combination scheme is effective in selecting the correct pitch values.

Fig. 7 illustrates a Venn diagram of the percentage of correct pitches for each method. This figure indicates that the percentage of the union of the correct pitches of these 3 methods is 88.02%, which is also the upper-bound of our proposed method.

5. CONCLUSIONS AND FUTURE WORK

In this study, we have proposed a hybrid method to extract the singing pitch from polyphonic audio music. Since pitch errors are more likely to happen in the beginning and ending of a vocal segment, we adopted 3 methods and explored the best way to combine them. Experimental results demonstrate that the proposed method can effectively deal with the errors to achieve a better recognition rate (in terms of raw-pitch accuracy) than the best algorithm for the audio melody extraction task in MIREX 2010.

From this study, it is obvious that there is no single pitch tracker can take care of all kinds of pitch errors, and the use of a committee of methods can effectively take advantage of each individual method to enhance the accuracy. Therefore there are two tasks as our immediate future work. The first one is to improve individual pitch tracker, such as finding a more effective set of features for HMM-based method. The second one is to apply more systematic way for combination, such as AdaBoost, which has been proven in the field of face detection [10].

6. REFERENCES

[1] M. Ryynänen and A. Klapuri, "Transcription of the Singing Melody in Polyphonic Music," *7th ISMIR*, pp. 222-227, 2006.

[2] Y. Li and D. L. Wang, "Detecting Pitch of Singing Voice in Polyphonic Audio," *IEEE ICASSP*, pp.17–20, 2005.

[3] C. L. Hsu, D. L. Wang, and J. S. Jang, "A Trend Estimation Algorithm for Singing Pitch Detection in Musical Recordings," *IEEE ICASSP*, pp.393-396, 2011

[4] H. Tachibana, T. Ono, N. Ono, and S. Sagayama, "Melody Line Estimation in Homophonic Music Audio Signals Based on Temporal Variability of Melody Source," *IEEE ICASSP*, pp. 425-428, 2010.

[5] K. Dressler, "An Auditory Streaming Approach on Melody Extraction," *Extended abstract for 7th ISMIR*, 2006.

[6] L. Regnier and G. Peeters, "Singing Voice Detection in Music Tracks Using Direct Voice Vibrato Detection," *IEEE ICASSP*, pp. 1685-1688, 2009

[7] C. L. Hsu and J. S. Jang, "Singing Pitch Extraction by Voice Vibrato/Tremolo Estimation and Instrument Partial Deletion", *ISMIR*, pp.525-530, 2010

[8] C. L Hsu, L. Y. Chen, J. S. Jang, and S. J. Li, "Singing Pitch Extraction from Monaural Polyphonic Songs by Contextual Audio Modeling and Singing Harmonic Enhancement," *10th ISMIR*, pp.201-206, 2009

[9] C. L. Hsu and J. S. Jang, "On the Improvement of Singing Voice Separation for Monaural Recordings Using the MIR-1K Dataset," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, pp.310-319, 2010.

[10] P. Viola, M. J. Jones, "Robust Real-Time Face Detection," *International Journal of Computer Vision*, vol. 57(2), pp.137-154, 2004