A TREND ESTIMATION ALGORITHM FOR SINGING PITCH DETECTION IN MUSICAL RECORDINGS

Chao-Ling Hsu[1], *DeLiang Wang*[2], *and Jyh-Shing Roger Jang*[1]

[1]Computer Science Department, National Tsing Hua University, Taiwan
{leon, jang}@mirlab.org
[2]Department of Computer Science and Engineering & Centre for Cognitive Science,
The Ohio State University, Columbus, USA
dwang@cse.ohio-state.edu

ABSTRACT

Detecting pitch values for singing voice in the presence of music accompaniment is challenging but useful for many applications. 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 reducing a large number of wrong pitch candidates either produced by musical instruments or the overtones of the singing voice. The proposed algorithm can be applied to improve the performance of singing pitch detection. Quantitative evaluations show that proposed trend estimation improves an existing algorithm significantly. The results from the MIREX 2010 competition show that our system achieves the best overall raw-pitch accuracy for vocal songs.

Index Terms— Singing pitch detection, pitch range estimation, music.

1. INTRODUCTION

Pitch determination for singing voice is a fundamental problem as the pitch contour of singing voice represents the melody of a song [1]. A robust singing pitch detection algorithm is useful for many applications, such as singing voice separation, music retrieval, singer identification, lyric recognition, and auto-tagging. Designing such an algorithm is challenging due to interfering pitch of music accompaniments.

Numerous methods have been proposed to extract the melody of a song, as evidenced by many submissions to the audio melody extraction competition in MIREX 2005~2010[1]. Most of the submissions have a similar framework. First, they identify pitch candidates by estimating pitch values of different sound sources for every time frame. After that, they select one of the pitch candidates as the target pitch. Finally the output pitch track is formed by connecting the target pitches with some post processing techniques (e.g., smoothing).

Another approach is to perform pitch detection after separating/enhancing the singing voice of the input mixture (e.g., [2][3]). The performance of the separation-based methods heavily relies on the result of voice separation. These methods detect singing pitches in some plausible pitch ranges which are chosen heuristically. The pitch ranges are usually large to cover most of the possible pitches of singing voice such as from 80 Hz to 800 Hz. However, it is unlikely that pitch changes in such a wide range in a short period of time. Furthermore, the upper pitch boundary of singing can be as high as 1400 Hz for soprano singers. It is difficult to give an appropriate pitch range if no prior knowledge is given for an input song.

To address the above problems, we propose a trend estimation algorithm extended from our previous work [4]. The advantages of adopting trend estimation are: 1) Appropriate pitch ranges of singing voice are estimated from the input instead of using a fixed range; 2) the trend is estimated dynamically so that it is able to fit the pitch trajectory of the corresponding singing voice; 3) it can be easily applied to improving singing pitch detection of existing systems by using a tighter pitch range for a given time frame.

The rest of this paper is organized as follows. Section 2 describes the proposed system in detail. The experimental results are presented in section 3, and section 4 concludes this work.

2. TREND ESTIMATION

This section describes our trend estimation. Fig. 1 shows a schematic diagram of the proposed algorithm. First, singing voice is enhanced by considering temporal and spectral smoothness. A sinusoidal partial extraction algorithm is then applied to extract the frequency slopes of the harmonic sounds. After that, we extract features from each partial and use a classifier to detect and prune instrumental partials. Finally, we locate vocal fundamental frequencies (F0s) in a series of time-frequency (T-F) blocks according to the remaining partials after pruning. These T-F blocks give pitch ranges along time and are much narrower than that of the entire voice pitch range. Comparing to our previous algorithm in [4], we add the vocal component enhancement as pre-processing. We also improve sinusoidal partial extraction so that the extracted partials are more accurate. In addition, a classifier is trained to perform instrumental partial detection instead of choosing parameter values heuristically. Lastly, we now consider the neighbors of selected T-F blocks to improve pitch range estimation.

[1] The extended abstracts of the submissions can be found at http://www.music-ir.org/mirex/

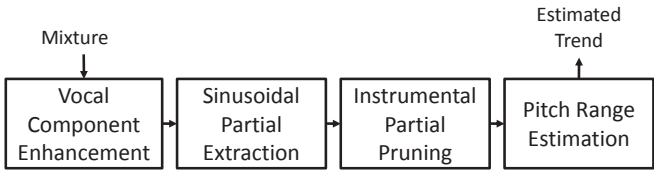

Fig. 1. Schematic diagram of trend estimation.

2.1. Vocal Component Enhancement

Our trend estimation starts from a vocal enhancement algorithm proposed by Tachibana et al. [3]. They used harmonic/percussive sound separation (HPSS) to enhance singing voice in two stages. In the first stage, they attenuate the energy of harmonic instruments (e.g., guitar and flute). The energy of percussive instruments (e.g. drum and cymbal) is further attenuated in the second stage.

In this study, we only apply the first stage of 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 the pitch of singing voice. In addition, the second stage usually corrupts the singing voice in the spectrogram and degrades the performance of our partial extraction algorithm.

2.2. Sinusoidal Partial Extraction

This stage extracts sinusoidal partials from a mixture. First, we apply the multi-resolution fast Fourier transform (MR-FFT) proposed by Dressler [5]. It removes the unreliable peaks that do not originate from periodic sounds by considering the local characteristics of phase spectrum, or more precisely, the instantaneous frequencies of neighboring frequency bins.

Reliable peaks are then used to form sinusoidal partials. 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, the one with the largest magnitude is selected.

A grouping algorithm is applied after peak selection. The goal of the algorithm is to group peaks so that each peak corresponds to a partial. It consists of three steps: initial grouping, re-grouping, and refining.

1) Initial grouping: It starts by selecting any ungrouped peak in the spectrogram as the first peak in the group and recursively groups other peaks neighboring to the group until every peak belongs to a group. To choose a peak from the time-overlapping peaks in a group, we apply a pitch dynamic prediction algorithm which finds a least-squares straight line fitting to the three most recent peaks and predicts the next peak value by extrapolation [6]. The closest peak is selected if the difference between it and the previous peak is not larger than a semitone. If no peak exists, the group is divided into two.
2) Re-grouping: The objective of this step is to connect the partials that are likely to be originated from the same one. We first compute the Euclidean distance between each pair of groups. Two groups are merged into one if their distance is less than 4.5 grid points on the T-F plane and have a time overlap not longer than 2 frames.
3) Refining: Because the re-grouping step introduces time-overlapping peaks or gaps for a group, this step selects one peak from overlapping peaks and fills the gaps for each partial. For selecting time-overlapping peaks, the one with the largest magnitude is retained. For filling the gaps, values of frequency and magnitude are interpolated according to the adjacent peaks.

2.3. Instrumental Partial Pruning

This stage considers the natural differences between vocal partials and instrumental partials: vibrato and tremolo. Vibrato refers to the periodic variation of pitch (or frequency modulation), and tremolo refers to the periodic variation of intensity (or amplitude modulation). These two features were proposed by Regnier et al. [7] and were previously used to detect the presence of singing voice. The idea of using these two features comes from the fact that human voice naturally contains strong vibrato and tremolo at the same time while most of the musical instruments contain only one of them [8].

Two attributes are computed to describe vibrato and tremolo: the rate and the extent of vibrato or tremolo. For human singing voice, the average rate is around 6Hz for both vibrato and tremolo. Hence we determine the relative extent around 6Hz by using the Fourier transform for both vibrato and tremolo.

More specifically, to compute the relative extent of vibrato for a partial $p_k(t)$ existing from time t_i to t_j, the Fourier transform of its frequency values $f_{p_k}(t)$ is given by:

$$F_{p_k}(f) = \sum_{t=t_i}^{t_j} (f_{p_k}(t) - \mu_{f_{p_k}}) e^{-2i\pi f \frac{t}{L}}, \quad (1)$$

where $\mu_{f_{p_k}}$ is the average frequency of $p_k(t)$ and $L = t_j - t_i$. The relative extent is given by:

$$\Delta f_{relp_k}(f) = \frac{F_{p_k}(f)}{L\mu_{f_{p_k}}}. \quad (2)$$

Lastly, the relative extent around 6Hz is computed as follows:

$$\Delta f_{p_k} = \max_{f \in [4\text{Hz}, 8\text{Hz}]} \Delta f_{relp_k}(f). \quad (3)$$

The relative extent for tremolo can be computed in the same way except that amplitude a_{p_k} is used instead of f_{p_k}.

Different from previous work, we consider partial detection as a classification problem. Given feature vectors $X = \{X(p_k)\}_{p_k}$ where $X(p_k) = (\Delta f_{p_k}, \Delta a_{p_k})$, two Gaussian mixture models (GMMs) Γ_V and Γ_I are trained for vocal partials and instrumental partials. The vocal/instrumental decision for the

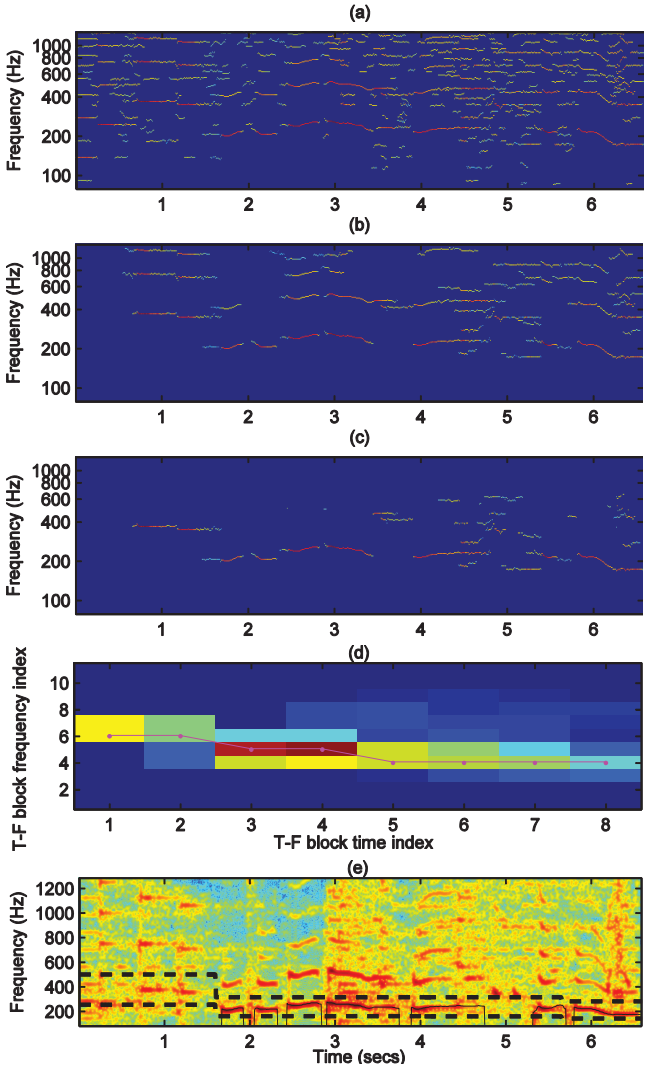

Fig. 2. An example of trend estimation. (a) Extracted partials. (b) The result of instrumental partial pruning. (c) The result after deleting the harmonic partials. (d) Magnitude-downsampled diagram. A color indicates the energy strength of a T-F block. The solid line indicates the optimal path found by DP. (e) The spectrogram of the song mixture. Dashed lines show the extended boundaries of the estimated trend and solid lines show the ground truth pitch.

partial p_k can be made by comparing the log-likelihood ratio with a threshold ψ :

$$\log p(X(p_k)|\Gamma_V) - \log p(X(p_k)|\Gamma_I) \underset{instrumental}{\overset{vocal}{\gtrless}} \psi \,. \qquad (4)$$

Fig. 2(a) and Fig. 2(b) show examples of partial extraction and instrumental partial pruning from a song mixture, respectively. As we can see, lots of the instrumental partials are pruned while most of the vocal F0 partials are retained.

2.4. Pitch Range Estimation

The last stage of the trend estimation is to find a sequence of relatively tight pitch ranges where the F0s of the singing voice are present.

Harmonic partials are first deleted based on the observation that the vocal F0 partial can only be the lowest-frequency partial within a frame. This process is repeated until we have only several partials representing possible vocal F0 partials. Fig. 2(c) shows an example of deleting the harmonic partials from that of Fig. 2(b). The rest of the partials represent possible vocal F0 partials of the song mixture.

We then downsample the magnitudes of the partials 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 represents a time duration. The entire plane is divided into a fixed set of T-F blocks with 50% overlap in time and in frequency. Finally, we find an optimal path consisting of a sequence of T-F blocks that contain the largest downsampled magnitudes by using dynamic programming (DP). The above procedure is similar to that in [4]. Fig. 2(d) 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.

In this study, we also consider the neighboring T-F blocks along the optimal path because they partly overlap in frequency. If the strength of the higher frequency neighbor of a T-F block is much larger than that of its lower frequency neighbor, the upper boundary is extended since most energy is concentrated in the higher frequency part, and vice versa. By using this technique, we can employ smaller bandwidth T-F blocks to find the optimal path and then extend it. This makes the trend estimation algorithm 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(e) shows the spectrogram of the song mixture along with extended boundaries of the estimated trend (dashed lines) and ground truth pitch (solid line). As can be seen, the estimated trend locates the singing F0s successfully.

3. EVALUATION

We use MIR-1K, a publicly available dataset proposed in our previous work [9], to evaluate our trend estimation algorithm. It contains 1000 song clips. The duration of each clip ranges from 4 to 13 seconds, and the total length of the dataset is 133 minutes. Music accompaniment and singing voice were recorded at left and right channels respectively. The ground truth of the pitch values of a singing voice was first estimated from the pure singing voice and then manually adjusted. All songs are mixed at -5 dB, 0 dB, and 5 dB signal to noise ratio (SNR) for evaluation. Note that the signal and noise here refer to the singing voice and music accompaniment respectively.

3.1. Evaluation of Instrumental Partial Pruning

Two 16-components GMMs are trained for vocal partials and instrumental partials, respectively. All GMMs have diagonal covariance matrices. Parameters of the GMMs are initialized via a K-means clustering algorithm and are iteratively adjusted via an expectation-maximization algorithm with 20 iterations.

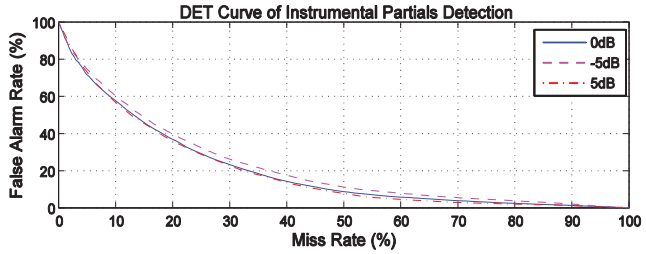

Fig. 3. Detection tradeoff curves for instrumental partial detection.

Fig. 3 shows the DET curves of instrumental partial detection by controlling the threshold ψ in Eq. (4). The goal of this stage is to prune as many instrumental partials as possible without deleting too many vocal partials so that the trend of singing pitch can be estimated robustly. We pick ψ when the false alarm rate in training set is 15%. In this case, at least 57% of instrumental partials are pruned at different SNRs while 15% of vocal partials are pruned erroneously. As can be seen from the figure, the three curves are close to each other. This shows that instrumental partials detection and pruning is robust to different amounts of background interference.

3.2. Evaluation of Trend Estimation and Singing Pitch Detection

The parameters for this experiment are set as follows. The sizes along time and frequency axes for each T-F block are 3 seconds and 8 semitones, respectively, with 50% overlap. The bandwidth of a trend after boundary extension is one octave (12 semitones).

Table 1 shows the results of trend estimation and singing pitch detection where the second and third rows show the performance of trend estimation in [4] and our proposed algorithm at different SNR levels respectively. A pitch range is considered as correct if the ground truth of that frame is within the trend. It can be seen that our trend estimation outperforms the previous method, especially at low SNR.

We perform pitch detection by using the dynamic programming algorithm proposed in [4]. The fourth and fifth rows of Table 1 show the performance of pitch detection with and without the proposed trend estimation, respectively. An estimated pitch is considered correct if its difference to the ground truth is less than 0.5 semitone. As we can see, the performance of pitch detection is improved significantly at all three SNR levels.

We also submitted this algorithm to the MIREX 2010 audio melody extraction competition. Fig. 4 shows the results of the last two years. Each submission has been tested on six datasets. The first bar shows the result of our system and it achieves the best raw-pitch accuracy for vocal songs.

4. CONCLUSIONS

This paper proposes a trend estimation algorithm for singing pitch detection in musical recordings, which has been little investigated in previous research. Our algorithm decides the pitch range without the prior knowledge of input songs and substantially improves the performance of pitch detection. This algorithm can be adopted by existing singing pitch detection algorithms to improve their results. The evaluation results show the proposed trend estimation algorithm is robust at different SNRs and perform well on different datasets.

Table 1. Results of trend estimation and singing pitch detection

	-5 dB	0 dB	5dB
Trend Estimation in [4]	77.01%	88.91%	93.94%
Proposed Trend Estimation	88.78%	93.44%	94.42%
Raw-pitch Accuracy (without trend estimation)	53.68%	70.30%	76.94%
Raw-pitch Accuracy (with trend estimation)	69.65%	81.15%	86.79%

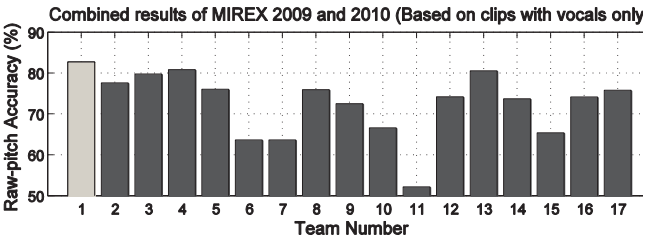

Fig. 4. The combined results of MIREX 2009 and 2010 for vocal songs. The first bar indicates the performance of the proposed algorithm.

Acknowledgements. Part of the work was conducted while C. L. Hsu was visiting OSU and supported in part by an NSC grant (NSC 96-2628-E-007-141-MY3). D. L. Wang was supported in part by an AFOSR grant (FA9550-08-1-0155).

5. REFERENCES

[1] G. E. Poliner, D. P. W. Ellis, A. F. Ehmann, E. Gomez, S. Streich, and B. Ong, "Melody transcription from music audio: approaches and evaluation," *IEEE TASLP*, vol. 15, pp. 1247-1256, 2007.

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

[3] 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.

[4] C. L. Hsu and J. S. Jang, "Singing pitch extraction by voice vibrato/tremolo estimation and instrument partial deletion", *ISMIR*, pp. 525-530, 2010

[5] K. Dressler, "Sinusoidal extraction using an efficient implementation of a multi-resolution FFT," *Proceedings of the International Conference on Digital Audio Effects*, pp. 247–252, 2006

[6] Parsons, T.W., "Separation of speech from interfering speech by means of harmonic selection", *J. Acoust. Soc. Am,* vol. 60, No. 4, pp. 911-918, 1976.

[7] L. Regnier and G. Peeters, "Singing voice detection in music tracks using direct voice vibrato detection," *IEEE ICASSP*, pp. 1685-1688, 2009.

[8] V. Verfaille, C. Guastavino, and P. Depalle, "Perceptual evaluation of vibrato models," *Proceedings of Conference on Interdisciplinary Musicology*, 2005.

[9] C. L. Hsu and J. S. Jang, "On the improvement of singing voice separation for monaural recordings using the MIR-1K dataset," *IEEE TASLP*, vol. 18, pp. 310-319, 2010.