SINGING PITCH EXTRACTION BY VOICE VIBRATO/TREMOLO ESTIMATION AND INSTRUMENT PARTIAL DELETION

Chao-Ling Hsu **Jyh-Shing Roger Jang**

Multimedia Information Retrieval Laboratory
Computer Science Department, National Tsing Hua University
Hsinchu, Taiwan
{leon, jang}@mirlab.org

ABSTRACT

This paper proposes a novel and effective approach to extract the pitches of the singing voice from monaural polyphonic songs. The sinusoidal partials of the musical audio signals are first extracted. The Fourier transform is then applied to extract the vibrato/tremolo information of each partial. Some criteria based on this vibrato/tremolo information are employed to discriminate the vocal partials from the music accompaniment partials. Besides, a singing pitch trend estimation algorithm which is able to find the global singing progressing tunnel is also proposed. The singing pitches can then be extracted more robustly via these two processes. Quantitative evaluation shows that the proposed algorithms significantly improve the raw pitch accuracy of our previous approach and are comparable with other state of the art approaches submitted to MIREX.

1. INTRODUCTION

The pitch curve of the lead vocal is one of the most important elements of a song as it represents the melody. Hence it is broadly used in many applications such as singing voice separation, music retrieval, and auto-tagging of the songs.

Lots of work which focuses on extracting the main melody of songs has been proposed in the literature. Poliner et al. [1] comparatively evaluated different approaches and found that most of the approaches roughly follow the general framework as follows: Firstly, the pitches of different sound sources are estimated at a given time and some of them are then selected as the candidates. The melody identifier then chooses one, if any, of these pitch candidates as a constituent of the melody for each time frame. Finally the output melody line is formed after smoothing the raw pitch line. Since the goal of most of these approaches is to extract the melody line carried by not only the singing voice but also the music instruments, they do not consider the different characteristics between the human singing voice and instruments: formants, vibrato and tremolo. More related work can be found in our previous work [3].

In the present study, we apply the method suggested by Regnier and Peeters [2], which was originally used to detect the presence of singing voice. This method utilizes the vibrato (periodic variation of pitch) and tremolo (periodic variation of intensity) characteristics to discriminate the vocal partials from the music accompaniment partials. We apply this technique to the singing pitch extraction so that the singing pitches can be tracked with less interference of instrument partials.

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 with possible future directions.

2. SYSTEM DESCRIPTION

Fig. 1 shows the overview of the proposed system. The sinusoid partials are first extracted from the musical audio signal. The vibrato and tremolo information is then estimated for each partial. After that, the vocal and instrument partials can be discriminated according to a given threshold, and the instrument partials can be therefore deleted. With the help of instrument partials deletion, the trend of the singing pitches can be estimated more accurately. This trend is referred to as global progressing path and indicates a series of time-frequency regions (T-F regions) where the singing pitches are likely to be present. Since the T-F regions consider relatively larger periods of time and larger ranges of frequencies, they are able to provide robust estimations of the energy distribution of the extracted sinusoidal partials.

On the other hand, the normalized sub-harmonic summation (NSHS) map [3] which is able to enhance the harmonic components of the spectrogram is computed, and the instrument partials which are discriminated with lower thresholds are deleted from NSHS map. After that, the global trend is applied to the instrument-deleted NSHS map.

The energy at each semitone of interest (ESI) [3] is then computed from the trend-confined NSHS map. Finally, the continuous raw pitches of the singing voice are

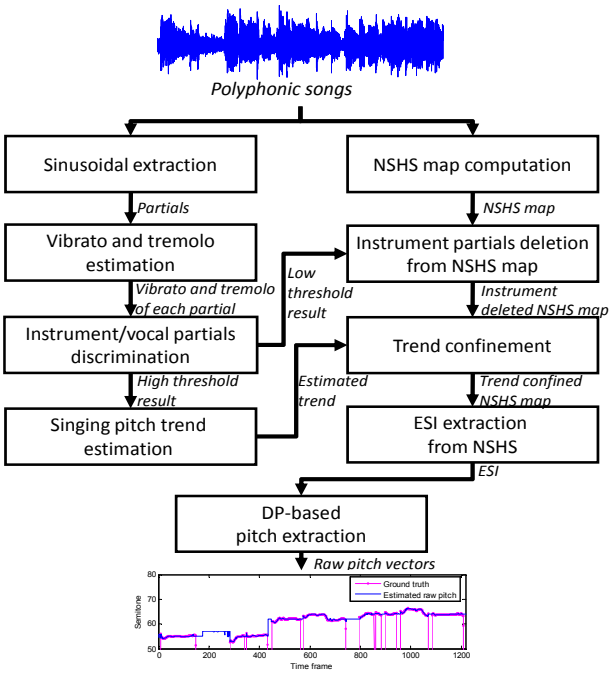

Figure 1. System overview

estimated by tracking the ESI values using the dynamic programming (DP) based pitch extraction.

An example is shown in the evaluation section (3.2). The following subsections explain these blocks in detail.

2.1 Sinusoidal Extraction

This block extracts the sinusoidal partials from the musical audio signal by employing the multi-resolution FFT (MR-FFT) proposed by Dressler [4]. It is capable of covering the fast signal changes and maintaining an adequate discrimination of concurrent sounds at the same time. Both of these properties are extremely well justified for the proposed approach.

The extracted partials with short duration are excluded in this stage because they are more likely to be produced by some percussive instruments or unstable sounds.

2.2 Vibrato and Tremolo Estimation

After extracting the sinusoidal partials, the vibrato and tremolo information of each partial are estimated by this block by applying the method suggested by Regnier and Peeters [2].

Vibrato refers to the periodic variation of pitch (or frequency modulation, FM) and tremolo refers to the periodic variation of intensity (or amplitude modulation, AM). Due to the mechanical aspects of the voice production system, human voice contains both types of the modulations at the same time, but only a few musical instruments can produce them simultaneously [5]. In general, wind and brass instruments produce AM dominant sounds, while string instruments produce the FM dominant sounds.

Two features are computed to describe vibrato and tremolo: frequencies (the rate of vibrato or tremolo) and amplitudes (the extent of vibrato or tremolo). For human singing voice, the average rate is around 6Hz [6]. Hence we determine the relative extent values around 6Hz by using the Fourier transform for both vibrato and tremolo.

More specifically, to compute a relative extent value 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:

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

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

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

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

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

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

2.3 Instrument/Vocal Partials Discrimination

The instrument and vocal partials are discriminated according to the given thresholds of the relative extent of vibrato and tremolo. The instrument partials can then be deleted if both the relative extents are lower than specified values. By selecting the thresholds, we can adjust the trade-off between instrument partials deletion rate and vocal partials deletion error rate. The higher thresholds are, the more instrument partials are deleted, but the more deletion errors of the vocal partials are. Usually a lower threshold is applied for instrument partials deletion from NSHS map, while a higher threshold is applied for the singing pitch trend estimation. The reasons will be explained in the following subsections.

2.4 Singing Pitch Trend Estimation

One of the major error types of singing pitch extraction is the doubling and halving errors where the harmonics or sub-harmonics of the fundamental frequency are erroneously recognized as the singing pitches. Here we refer the harmonic partials to those partials whose frequencies are multiples of the F0 partials. And we use "vocal partials" to indicate the union of the disjoint sets of "vocal F0 partials" and "vocal harmonic partials". Although the error can be handled by considering the time and frequency smoothness of the pitch contours, most of the approaches only consider the local smoothness during a short period of time. However, there are many 'gaps' between successive vocal partials such as the non-vocal pe-

riod between two segments of lyrics where instrument partials may be predominant in these gaps. These instrument partials often act like 'bridges' which may mislead the pitch tracking algorithm to connect two vocal partials erroneously.

To deal with this problem, we propose a method to estimate the trend of the singing pitches. Firstly, higher thresholds are applied to delete more instrument partials. This might also delete some vocal partials, but it will not affect the pitch trend estimation as long as we still have enough vocal partials. Secondly, the harmonic partials are deleted based on the assumption that the lowest-frequency partial within a frame is the vocal F0 partial. Moreover, these deleted harmonic partials are accumulated into their vocal F0 partials. This process is repeated until we have only several low-frequency partials representing potential vocal F0 partials. As a result, most of the harmonic partials are deleted and the energy of the vocal F0 partials is strengthened. The energy of the remaining partials is then max-picked for each frame and summed up within a time-frequency region (T-F region). More precisely, given a spectrogram $x[t,f]$ computed from the previous MR-FFT, the strength $s_{T,F}$ of the T-F region is defined as:

$$s_{T,F} = \sum_{t=0}^{M_{time}-1} \max_{f\in[0,M_{freq}-1]} x[t+TL_{time}, f+FL_{freq}],$$
$$T = 0,1,...n-1 \text{ and } F = 0,1,...m-1$$

where

t	is the index of the time frame.
f	is the index of the frequency bin.
n	is the number of T-F regions in the time axis
m	is the number of T-F regions in the frequency axis
T , F	are the indices of the T-F region in time and frequency axes respectively.
L_{time} , L_{freq}	are the time and frequency advance of the T-F region (hop-size) respectively.
M_{time} , M_{freq}	are the number of the time frames and the number of the frequency bins of a T-F region respectively.

The size of the T-F region should be large enough so that the global trend of the singing pitches can be acquired. On the other hand, the T-F region should also be small enough so that the harmonics of the singing pitches can be separated in different frequency bands and the pitch changes can be captured in different time periods. Note that although M_{freq} is fixed for all T-F regions, the frequency ranges are different for the T-F regions in different frequency bands. This is because the frequency bins in the result of sinusoidal extraction via MR-FFT are spaced by 0.25 semitone. In other words, the lower frequency T-F region has smaller frequency range since the frequency differences between low fundamental frequency partials and their harmonics are relatively smaller than that of high fundamental frequency partials.

Because the singing pitch trend should be smooth, the problem is defined as the finding of an optimal path $[F_0,\cdots,F_i,\cdots,F_{n-1}]$ that maximizes the score function:

$$score(F,\theta) = \sum_{T=0}^{n-1} s_{T,F_T} - \theta \times \sum_{T=1}^{n-1} |F_T - F_{T-1}|,$$

where s_{T,F_T} is the strength of the T-F region at the time index T and frequency index F_T. The first term in the score function is the sum of strength of the T-F region along the path, while the second term controls the smoothness of the path with the use of a penalty coefficient θ. If θ is larger, the computed path is smoother.

The dynamic programming technique is employed to find the maximum of the score function, where the optimum-valued function $D(T,l)$ is defined as the maximum score starting from time index 1 to T, with $F_T = l$:

$$D(T,l) = s_{T,l} + \max_{k\in[0,m-1]} \{D(t-1,k) - \theta \times |k-l|\},$$

where $t = [1,n-1]$, and $l = [0,m-1]$. The initial condition is $D(0,l) = s_{0,l}$, and the optimum score is equal to $\max_{l\in[0,m-1]} D(n-1,l)$. At last, this optimal path is applied to the instrument-deleted NSHS map described in section 2.6.

2.5 NSHS Computation

Instead of simply extracting the singing pitches by tracking the remaining vocal partials, the NSHS proposed by our previous work [3] is used since the non-peak values of the spectrum are also useful for the later DP-based pitch extraction algorithm. The NSHS is able to enhance the partials of harmonic sound sources, especially the singing voice. It is modified from the sub-harmonic summation [7] by adding a normalizing term. The reason of the modification is based on the observation that most of the energy in a song locates at the low frequency bins, and the energy of the harmonic structures of the singing voice decays slower than that of instruments [8]. It is therefore that, when more harmonic components are considered, energy of the vocal sounds is further strengthened.

2.6 Instrument partials deletion and trend confinement

In these two blocks, the instrument partials detected with the lower thresholds in the previous block are first removed from the NSHS map by setting their magnitude to zero (within the range of neighboring local minima). For extracting singing pitches, the thresholds are set to be lower in order to delete the instrument partials without deleting too many vocal partials. After that, the instrument deleted NSHS map can be further confined to the estimated pitch trend (section 2.4). In other words, only the energy along the trend will be retained.

2.7 ESI Extraction from NSHS

The ESI computed from the trend-confined NSHS map in the time frame t can be obtained as follows [3]:

$$v_t(n) = \max_{p_n - \frac{p_n - p_{n-1}}{2} \le p < p_n + \frac{p_{n+1} - p_n}{2}} \left(A_t(f)\right),$$

where $A_t(*)$ is the NSHS map calculated in the previous stage, $n = 0,1,..,N-1$, N is the total number of semitones that are taken into account, and p_n is the frequency of the n-th semitone in the selected pitch range.

Note that we also need to record the maximal frequency within each frequency range of ESI in order to reconstruct the most likely pitch contours.

2.8 DP-based Pitch Extraction

The DP-based pitch tracking algorithm is previously proposed in [3]. It is very similar to the algorithm described in section 2.4. The most likely pitch contour can be finally acquired by tracking the ESI computed in the previous block. Note that we do not perform vocal/non-vocal detection since it is not the focus of this study. In addition, the vocal/non-vocal detection can be implemented by various methods such as [2][3].

3. EVALUATION

Two datasets were used to evaluate the proposed approach. The first one, MIR-1K, is a publicly available dataset proposed in our previous work [9]. It contains 1000 song clips recorded at 16 kHz sample rate with 16-bit resolution. The duration of each clip ranges from 4 to 13 seconds, and the total length of the dataset is 133 minutes. These clips were extracted from 110 karaoke songs which contain a mixed track and a music accompaniment track. These songs were selected (from 5000 Chinese pop songs) and sung, consisting of 8 females and 11 males. Most of the singers are amateurs with no professional training. The music accompaniment and the singing voice were recorded at the left and right channels respectively. The ground truth of the pitch values of the singing voices were first estimated from the pure singing voice and then manually corrected. All songs are mixed at 0 dB SNR, indicating that the energy of the music accompaniment is equal to the singing voice. Note that the SNRs for commercial pop songs are usually larger than zero, indicating that our experiments were set to deal with more adversary scenarios than the general cases. The second dataset, ADC2004, is one of the testing dataset for audio melody extraction task in MIREX. It contains 20 song clips and the average length of the clips is around 20 seconds. Only the 12 vocal songs of ADC2004 are used for testing in this study. Although the size of ADC2004 is much smaller than that of MIR-1K, it is convenient for comparing the performance of different algorithms which were submitted to MIREX.

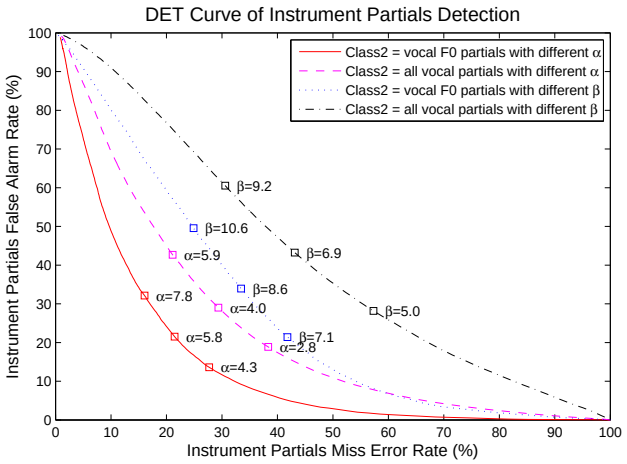

Figure 2. The DET curves of instrument partials false alarm rate versus instrument partials miss error rate by using different values of α and β as the thresholds alone, respectively. (Here we assume class 1 is instrument partials, and class 2 is either vocal F0 partials or all vocal partials.)

3.1 Evaluation for Instrument Partials Detection

The frame size and hop size used in the sinusoidal extraction by MR-FFT are 64 ms and 8 ms respectively. The frequency bins in MR-FFT are spaced by 0.25 semitone from 80Hz to 1280Hz, resulting a total of 192 bins. The partials whose durations are less than 56 ms are removed since they are more likely to be generated by percussive instruments or unstable sounds. With regard to the relative vibrato and tremolo extent estimation, the parameters are set to be the same as those suggested by [2].

Figure 2 shows the DET (detection error tradeoff) curves of instrument partials false alarm rate versus instrument partials miss error rate by using different relative vibrato extent (α) and relative tremolo extent (β) as the thresholds alone, respectively. A higher instrument partials false alarm rate indicates more vocal partials are erroneously recognized as instrument partials. On the other hand, a higher instrument partials miss error rate indicates more instrument partials are recognized as vocal partials. Here we assume class 1 is instrument partials, and class 2 is either vocal F0 partials or all vocal partials. The solid line and dotted line show the results of using vocal F0 partials as class 2 with different α and β respectively. The dashed line and dash-dot line show the results of using all vocal partials as class 2 with different α and β respectively. We want to show the results of using vocal F0 partials as class 2 because the goal of this study is to extract the singing pitches carried by these vocal F0 partials. In contrast, the harmonic partials of the singing voice are comparably not as important. All of these partials are extracted from the MIR-1K dataset. Since the MIR-1K has separated tracks of singing voice and accompaniment, the sources of the partials can be distinguished.

From Figure 2, it is obvious that α has better discriminative capability to detect instrument partials than β.

	Vocal F0	Non-vocal F0
Partials remaining in the pitch trend tunnel	82.47 %	19.19 %
Partials remaining in the pitch trend tunnel but deleted by instrument partial deletion	8.07 %	66.18 %
Final partials remaining	75.82%	6.49%
Vocal pitches remaining in the pitch trend tunnel	86.30%	

Table 1. Performance of singing pitch trend estimation

This is because the pop music in MIR-1K has less wind and brass instruments than string instruments. We have found in our preliminary experiment[1] that β has better vocal/instrument discriminative power for wind and brass instruments.

The instrument partials deletion block applied α = 0.1125 and β = 3. The vocal F0 remaining rate is around 94.3% (or equivalently, 5.7% instrument partials false alarm rate) and instrument partial deletion rate is around 60.4% (or equivalently, 39.6% instrument partials miss error rate). On the other hand, singing pitch trend estimation applied α = 0.3 and β = 5.5 as the thresholds. The vocal F0 partials remaining rate is 72.9% and instrument partials deletion rate is 82.8%.

3.2 Evaluation for Singing Pitch Trend Estimation

The parameters for this experiment were set as follows. The sizes along time and frequency axes for each T-F region were 3 seconds and 13.5 semitones, respectively. Their hop sizes were 1.5 seconds and 4 semitones, respectively. The penalty coefficient θ for the dynamic programming step was set to 1 empirically.

Table 1 shows the results of the singing pitch trend estimation. More than 82% of vocal F0 partials remain in the pitch trend tunnel and the singing pitches remaining rate is 86%. On the other hand, only 19.19% of instrument and vocal harmonic partials are retained within the pitch trend tunnel. In addition, 66.18% of the non-vocal F0 partials left in the pitch trend tunnel are deleted by the NSHS computation stage, and 8.07% of the remaining vocal F0 partials are deleted erroneously at the same time. Finally, 75.82% of vocal F0 partials remain while only 6.49% of non-vocal F0 partials are kept in both deletion procedures.

Figure 3 shows the stage-wise results in singing pitch extraction. Figure 3(a) shows all the partials after sinusoidal extraction. Figure 3(b) and 3(c) applies different thresholds on 3(a) to delete instrument partials for different purposes. Because 3(b) applies lower thresholds than those of 3(c), more instrument partials are removed in 3(c). The harmonic partials in Figure 3(c) are then further deleted in 3(d). Figure 3(f) is obtained by subtracting the

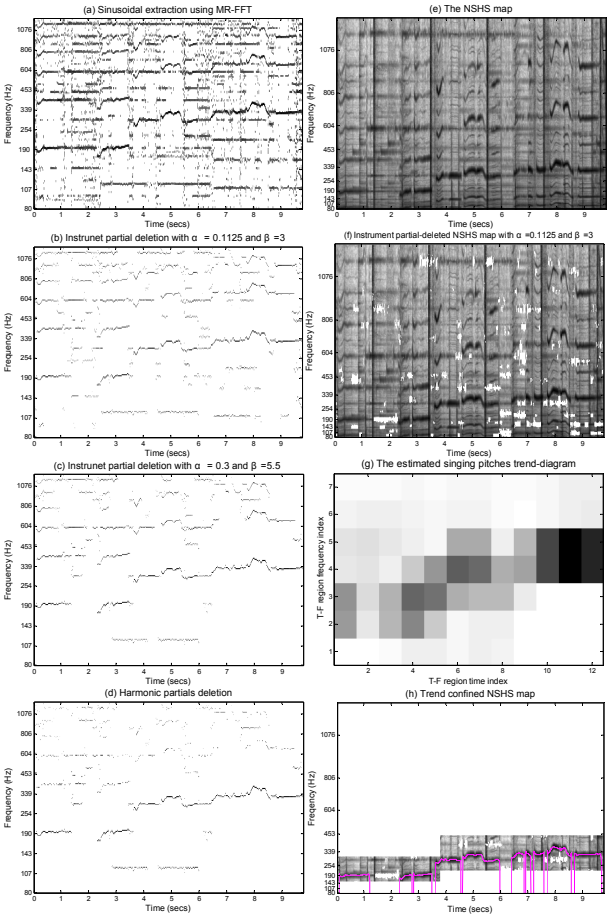

Figure 3. Stage-wise results of singing pitch extraction for the clip 'Ani_4_05.wav' in MIR-1K. (a) Results after sinusoidal extraction using MR-FFT. (b) The remaining partials after instrument partial deletion thresholds of α = 0.1125 and β = 3. (c) The remaining partials after instrument partial deletion after threshold of α = 0.3 and β = 5.5. (d) The result after harmonic partials deletion. (e) The NSHS map. (f) Instrument partial-deleted NSHS map with threshold of α = 0.1125 and β = 3. (g) The estimated singing pitches trend-diagram. (h) Trend confined NSHS map, where the solid line represents the ground truth of the singing pitches.

detected instrument partials in Figure 3(b) from the NSHS map in 3(e). Figure 3(g) illustrates the T-F regions computed from Figure 3(d), with color depth indicating the strength each T-F region. Finally, Figure 3(h) is the NSHS map (Figure 3(f)) confined by the pitch trend tunnel. As can be seen in this example, the identified pitch trend tunnel is capable of covering the vocal F0 partials (represented by solid lines) while most of the instrument partials are deleted.

3.3 Evaluation for Singing Pitch Extraction

Figure 4 shows the results of singing pitch extraction. The raw pitch accuracy is computed over the frames which were labeled as voiced in the ground truth. An estimated singing pitch is considered as correct if the deviation from the ground truth is small than 1/4 tone (or 1/2

[1] The experiment was also performed on the University of Iowa Musical Instrument Samples which is available at http://theremin.music.uiowa.edu/

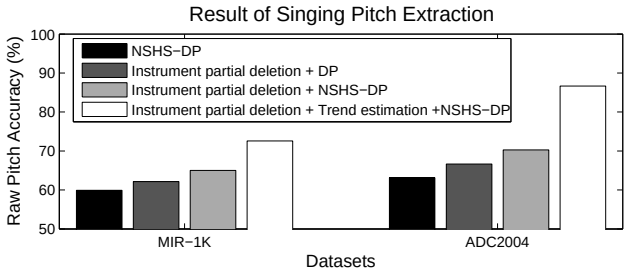

Figure 4. The results of singing pitch extraction.

semitone). The black bars show the performance of the

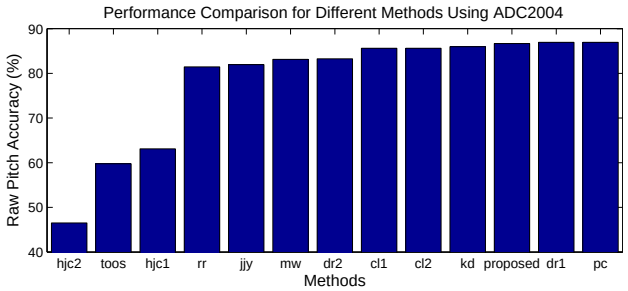

Figure 5. Performance comparison.

previous NSHS-DP method [3] (ranked 5-th out of 12 in MIREX2009). The dark gray bars show the result of combining the proposed instrument partial deletion and dynamic programming without using the NSHS. The light gray bars are the same as the dark gray bar except that the NSHS map is applied. The light gray bars perform better than the ones without using the NSHS map, which confirms the argument that the non-peak values of the spectrum are also useful. Lastly the white bars show the performance of the proposed approach where instrument partial deletion, singing pitch trend estimation, and NSHS are applied.

It is clear that the proposed instrument partial deletion and singing pitch trend estimation facilitate extracting singing pitches since its performance improves significantly over the rest of the compared methods in both datasets. The raw pitch accuracy of proposed approach achieves 72.57% and 86.67% for MIR-1K and ADC2004, respectively, with the same setting of the parameters described in previous subsections. Comparing to the MIREX 2009 results shown in Figure 5, the performance of the proposed approach is comparable to the state of the art approaches.

4. CONCLUSIONS AND FUTURE WORK

In this paper, we propose a novel approach for singing pitch extraction by deleting instrument partials. It is surprising that the vocal and instrument partials can be discriminated by only two simple features, and the performance is also encouraging. Besides, a singing pitch trend estimation algorithm is proposed to enhance the pitch extraction accuracy.

Since only the features suggested in [2] were used in this study, other characteristics of voice vibrato and tremolo could be use as new features for improving the performance. Moreover, it is worth noting that the proposed instrument partial deletion and singing trend estimation techniques are general for pitch extraction, in the sense that they can be applied to any other spectrum-based methods to delete the unlikely pitch candidates. Our immediate future work is to explore the use of the proposed techniques on top of existing methods to confirm their feasibility in further improving the performance.

5. ACKNOWLEDGEMENT

This work was conducted under the ”Digital Life Sensing and Recognition Application Technologies Project” of the Institute for Information Industry which is subsidized by the Ministry of Economy Affairs of the Republic of China.

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

[3] C. L. Hsu, L. Y. Chen, J. S. Jang, and H. J. Li, “Singing pitch extraction from monaural polyphonic songs by contextual audio modeling and singing harmonic enhancement”, *ISMIR*, pp. 201-206, 2009.

[4] K. Dressler, “Sinusoidal extraction using an efficient implementation of a multi-resolution FFT,” *DAFx*, pp. 247–252, 2006

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

[6] E. Prame, “Measurements of the vibrato rate of ten singers,” *JASA,* vol. 96, pp. 1979, 1994.

[7] D. J. Hermes, “Measurement of Pitch by Subharmonic Summation,” *JASA*, vol.83, pp. 257-264, 1988.

[8] Y. Li and D. L. Wang, “Detecting pitch of singing voice in polyphonic audio,” *IEEE ICASSP*, pp. 17–20, 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*, volume 18, pp. 310-319, 2010.