SINGING PITCH EXTRACTION AT MIREX 2010

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 extended abstract describes our submission to the MIREX 2010 evaluation task on Audio Melody Extraction. The algorithms are designed for vocal F0 extraction from the music accompaniment.

1. INTRODUCTION

In the submitted algorithm, we apply two methods which make use of the characteristic of human singing voice to discriminate singing partials from those from background music. The first one is Harmonic/Percussive sound separation (HPSS) [1] which originally was used to separate the percussive and harmonic sounds from music. By applying different window size of the short time Furious transform (STFT), it can be adapted to enhance human singing voice or attenuate instrument partials. The second method was 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 the HPSS as pre-processing to enhance the singing voice. The second method is used to discriminate and delete the instrument partials so that we can use the rest of the partials to estimate a singing pitch "trend" which is composed of a series of time-frequency regions that contain the singing pitches.

2. SYSTEM DESCRIPTION

Fig. 1 shows the overview of the submitted algorithm. The HPSS is firstly applied to enhance the singing voice in the input signal. After that, the sinusoid partials are extracted from the musical audio signal. The vibrato and tremolo information is then estimated for each partial. Then 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 HPSS and 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.

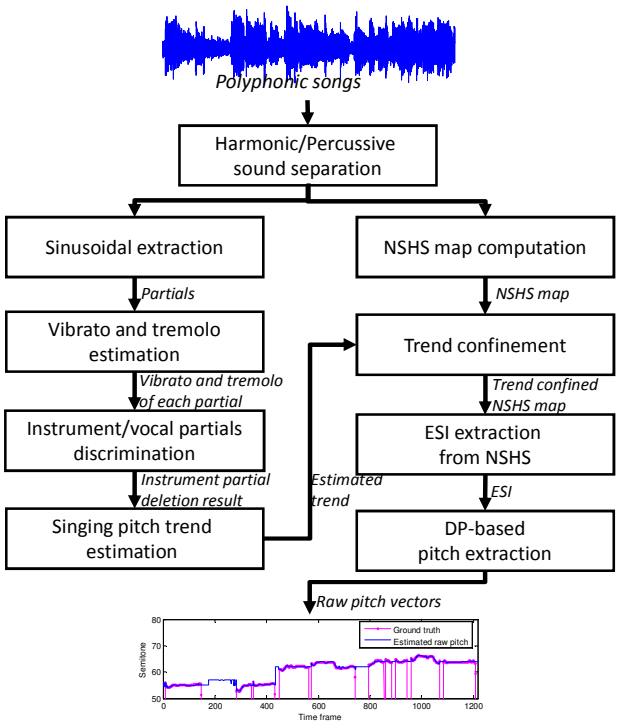

Figure 1. Algorithm 1 overview.

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

The following subsections explain these blocks briefly. For detail description please refer to our ISMIR2010 paper [4].

2.1 Harmonic/Percussive Sound Separation

We applied HPSS proposed by Tachibana et al. [1] to enhance the singing voice in music. They applied two-stage HPSS which make uses of long STFT window first to separate the sounds which have stronger temporal va-

riability (i.e. vocal sound and percussive sound) from music. After that, they use short STFT window to separate vocal and percussive sounds further. Different from them, we only use the first stage of their method as preprocessing which enhances the singing voice and percussive sounds. The reason is that we found that the percussive sounds don't really obstruct us to extract the singing pitches since they don't have harmonic structures. Furthermore, the second step may damage the vocal harmonic structures especially for those singing voice which have strong temporal variability. For more detail about HPSS please refer to [1].

2.2 Sinusoidal Extraction

This block extracts the sinusoidal partials from the musical audio signal by employing the multi-resolution FFT (MR-FFT) proposed by Dressler [6]. 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.

2.3 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. 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. 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}.

2.4 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.

2.5 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.

To deal with this problem, we propose a method to estimate the trend of the singing pitches.

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 frequen-

cy 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.

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:

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

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 $F_T = l$:

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

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.

2.6 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.

2.7 Trend Confinement

In this block, the NSHS map is further confined to the estimated pitch trend (section 2.5). In other words, only the energy along the trend will be retained.

2.8 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.

2.9 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 2.5. The most likely pitch contour can be finally acquired by tracking the ESI computed in the previous block.

2.10 Voiced/Non-voiced Detection

This block employs multilayer perception (MLP) with one hidden layer to classify voiced and non-voiced frames. 39-dimensional MFCCs (12 cepstral coefficients plus log energy, together with their first and second derivatives) were extracted from each frame. The MFCCs were computed from STFT with a half-overlapped 40-ms Hamming window. Cepstral mean subtraction (CMS) was used to reduce channel effects.

3. RESULTS

The results of raw-pitch accuracy for all the vocal songs in different datasets are shown in Table 1. The first column shows the names of the submitted algorithms (including the results of 2009) and the top row shows the names of different datasets. The performance of our algorithm is under the name "HJ1". Note that since our algorithm aims to extract the pitches from the vocal songs, 8 out of 20 clips in ADC2004 dataset and 9 out of 24 clips in MIREX05 dataset are non-vocal songs and are not considered in the result.

The proposed algorithm achieves the best raw-pitch accuracy, 82.72%. It is worth noting that the authors of [1] who proposed a method to extract the singing pitches by using HPSS also submitted an algorithm called "TOOS1". Because we also applied the HPSS as the pre-processing,

	ADC2004	MIREX05	INDIAN08	MIREX09 0dB	MIREX09 -5dB	MIREX09 +5dB	Average
HJ1 (proposed)	0.8461	**0.8155**	0.86	**0.8315**	0.7254	0.8849	**0.827233**
TOOS1 (2010)	0.6402	0.7576	0.824	0.8259	**0.7549**	0.8541	0.776117
JJY2 (2010)	0.8026	0.7506	**0.8855**	0.8129	0.6465	0.8888	0.797817
JJY1 (2010)	0.7957	0.7915	0.8845	0.822	0.6611	**0.8959**	0.80845
SG1 (2010)	0.7324	0.6757	0.8566	0.8005	0.6194	0.8774	0.760333
cl1 (2009)	0.856252	0.70807	0.508092	0.59138	0.453909	0.702637	0.636723
cl2 (2009)	0.856252	0.70807	0.508092	0.59138	0.453909	0.702637	0.636723
dr1 (2009)	0.869605	0.761145	0.880057	0.698804	0.537796	0.808947	0.759392
dr2 (2009)	0.832614	0.709258	0.865807	0.66549	0.505318	0.772989	0.725246
hjc1 (2009)	0.631101	0.626594	0.675624	0.726577	0.48658	0.848561	0.66584
hjc2 (2009)	0.465192	0.541294	0.608391	0.516871	0.214502	0.783801	0.521675
jjy (2009)	0.819596	0.762696	0.682963	0.759354	0.585304	0.843853	0.742294
kd (2009)	0.859698	0.774622	0.8782	0.804565	0.624877	0.891898	0.805643
mw (2009)	0.831351	0.757398	0.859869	0.672905	0.530621	0.770268	0.737069
pc (2009)	**0.869624**	0.717068	0.818281	0.508895	0.373777	0.636794	0.654073
rr (2009)	0.81446	0.759506	0.86161	0.686242	0.546785	0.778827	0.741238
toos (2009)	0.597683	0.734258	0.797606	0.822943	0.748896	0.848473	0.75831

Table 1. Raw-pitch accuracy for vocal songs in different datasets (2009 - 2010)

we would like to emphasize the comparison between their algorithm and ours. According to the results, our algorithm performed uniformly better for different datasets except for the case of "MIREX09 -5dB". It shows that the proposed singing pitch trend estimation improves the robustness of the pitch extraction algorithm.

4. CONCLUSIONS

This extended abstract describes our submission to the MIREX 2010 evaluation task on Audio Melody Extraction. Our algorithm has a comparatively better robustness for different datasets and outperforms others in average.

5. REFERENCES

[1] 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.

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

[3] Chao-Ling Hsu, Liang-Yu Chen, Jyh-Shing Roger Jang, and Hsing-Ji Li, "Singing pitch extraction from monaural polyphonic songs By contextual audio modeling and singing harmonic enhancement", *International Society for Music Information Retrieval*, Kobe, Japan, Oct. 2009.

[4] Chao-Ling Hsu and Jyh-Shing Roger Jang, "Singing pitch extraction by voice vibrato/ttremolo estimation and instrument partial deletion", *International Society for Music Information Retrieval*, Utrecht, Netherlands, Aug. 2010.

[5] H. Fujihara, M. Goto, J. Ogata, K. Komatani, T. Ogata, and H. G. Okuno, "Automatic Synchronization between Lyrics and Music CD Recordings Based on Viterbi Alignment of Segregated Vocal Signals," *ISM*, pp. 257–264, 2006.

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

[7] D. J. Hermes, "Measurement of Pitch by Subharmonic Summation," *Journal of Acoustic Society of America*, 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.