http://dx.doi.org/10.5755/j01.eee.19.6.4575

Efficient Vocal Melody Extraction from Polyphonic Music Signals

G. Yao[1,2], Y. Zheng[1,2], L. Xiao[1,2], L. Ruan[1,2], Y. Li[1,2]
[1] *State Key Laboratory of Software Development Environment,*
Beijing 100191, China
[2] *School of Computer Science and Engineering, Beihang University,*
Beijing 100191, China
yutianzuijin@cse.buaa.edu.cn

*Abstract***—Melody extraction from polyphonic music is a valuable but difficult problem in music information retrieval. This paper proposes a system for automatic vocal melody extraction from polyphonic music recordings. Our approach is based on the pitch salience and the creation of the pitch contours. In the calculation of pitch salience, we reduce the peaks number of the spectral transform using a two-level filter and shrink the pitch range in accordance with the experiment to improve the efficiency of the system. In the singing voice detection, we adopt a three-step filter using the pitch contour characteristics and their distributions. The quantitative evaluation shows that our system not only keeps the overall accuracy compared with the state-of-the-art approaches submitted to MIREX, but also achieves high algorithm efficiency.**

*Index Terms***—Audio content description, feature extraction, music information retrieval, pitch contour.**

I. Introduction

Vocal melody extraction from polyphonic music is an area of research that has received considerable attention in the past few years. The term melody has different definitions in different context. Nowadays, it mainly points to the pitch sequence of the lead vocal. The pitch sequence is usually manifested as the fundamental frequency (F0) contour of the singing voice in the polyphonic mixture [1]. It is broadly used in many applications such as singing voice separation, music retrieval, and singer identification, especially in Query by Humming [2].

In [3], a comprehensive review of state-of-the-art melody extraction method is provided. The basic processing structure of extraction there comprises three main steps – multi-pitch extraction, melody identification and post processing. This structure is often called the salience-based structure.

Besides the salience-based methods, there are some other designs based on the source/filter model [4], which is sufficiently flexible to capture the variability of the singing voice and the accompaniment in terms of pitch range and timbre. Although this kind of methods can also give a good result, they are hard to be understood and often run slow. Currently, the salience-based architecture is most widely adopted. The salience-based design has a common structure: first, get the spectral representation of the signal. The most popular technique is the short time Fourier transform (STFT). Also, there are a few systems using other methods, such as YIN pitch tracker [5], which often arises in melody extraction from MIDI and monophonic audios. Second, use the spectral representation to compute the F0 candidates. There exist many different strategies to compute the candidates, [6] uses the harmonic summation of the spectral peaks with assigned weights, whereas [1] lets the possible F0 to compete for harmonics based on expectation-maximization (EM) model. [7] takes a radical approach to feed the spectral representation into the support vector machine (SVM) classifier. The classifier will return only one pitch — the appropriate melody. At last, the melody is chosen from the candidate F0 using different methods.

Despite the variety of proposed approaches, vocal melody extraction from polyphonic music remains intractable. The current approach has an overall accuracy of around 70% from Music Information Retrieval Evaluation eXchange (MIREX) [8]. This is still lower compared with the melody extraction from MIDI. The main reason could be attributed to the lack of knowledge of the difference between the vocal and nonvocal melody at the singing voice detection stage. On the other hand, the system which gets a better overall accuracy runs relatively slow due to its high computational complexity.

In this paper, a system with high overall accuracy and low runtime is presented. To reduce the computation time of pitch salience which is the most time-consuming part of the system, the spectral peaks are first dropped using a two-level filter, and then shrink the pitch range of the salience bin. In the singing voice detection stage, a three-step filter using the contour characteristics and their distributions to discriminate the vocal and nonvocal melody is proposed. Besides the distributions in [9], more characteristics and their distributions are adopted. The experiment result shows that our approach not only keeps a high overall accuracy but also decreases the runtime obviously.

Manuscript received January 28, 2013; accepted April 4, 2013.

This research was funded by the Hi-tech Research and Development Program of China (863 Program) under Grant No.2011AA01A205, the National Natural Science Foundation of China under Grant No.61232009, the Doctoral Fund of Ministry of Education of China under Grant No.20101102110018, Beijing Natural Science Foundation under Grant No.4122042, the fund of the State Key Laboratory of Software Development Environment under Grant No.SKLSDE-2012ZX-07 and the Open Research Fund of The Academy of Satellite Application under grant NO.SSTC-YJS-01-03.

The rest of this paper is organized as follows. Section II describes the proposed system in detail. The experimental results are presented in section III, and section IV concludes this work with possible future improvements.

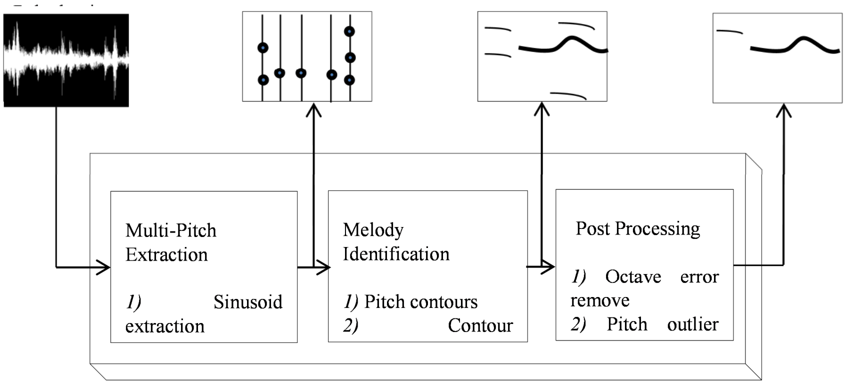


Fig. 1. System overview.

II. System Description

Figure 1 shows the overview of our system. The first stage is the front end of the vocal melody extraction, called multi-pitch extraction. The sinusoid extraction takes the spectral transform of the polyphonic music signal to reveal the sinusoidal peaks. The peaks are first filtered, and then used to compute a representation of pitch salience over time. The peaks of pitch salience form the F0 candidates for the main melody.

In melody identification stage, the main job is to find the vocal melody. To this end, a set of pitch contours are created, which are formed by connecting consecutive pitch candidates with similar frequencies. To reduce the generation of the non-melody contours, the salience peaks will be filtered at first. Using these contours, a lot of contour characteristics will be defined, which can be used to discriminate whether the contour belongs to the melody. After that, vocal melody is chosen out of all contours in a three-step singing voice detection stage with the help of contour characteristics.

In the final stage, mainly called post-processing, the octave error and pitch outlier of the contours are disposed using the melody pitch mean proposed by Salamon [9]. At last, the main melody is selected from the remaining contours.

A. Multi-pitch extraction

1) Sinusoid extraction

Given a frame of music signal, the STFT is defined as

$$X_l(k) = \sum_{n=0}^{M-1} w(n) * x(n+lH) e^{-j\frac{2\pi}{N}kn},$$
$$l=0,1,\ldots \text{and } k=0,1,\ldots,N-1 \qquad (1)$$

where *n* is the wave data of polyphonic music; *w(n)* is the window function; *N* is the number of STFT points; *H* is the time advance of frame (i.e. hop size); *M* is the frame size; *l*is the frame number.

The window used in our system is Hann window, which has a length 2048, the same as the music frame, approximately 46.4 ms for music with 44.1 kHz sample rate (*fs*), and a hop size of 10 ms. FFT length is 8192 with a 4 times zero padding. The long FFT length cannot give more information about the spectrum, but an enhanced frequency resolution. For data sampled at 44.1 kHz, the resolution is limited to *fs*/*N* =5.38Hz.

Some melody extraction systems use a multi-resolution transform instead of the STFT which has a fixed time-frequency resolution [10], [11]. In [12], it was shown that "the multi-resolution FFT did not provide any statistically significant improvement to spectral peak frequency accuracy and only a marginal improvement to the final melody F0 accuracy". So we just opt for the STFT in our system.

2) Spectral peaks filter

After spectrum transform, the signal is transformed from temporal domain to spectral domain. The spectral peaks, originated from vocal or instrumental accompaniment signals, or the noise, are used to calculate the pitch salience. Generally, the number of original peaks is large because of the accompaniment signals. When it comes to a vocal frame, there exist some peaks with salient magnitude. It stands a good chance that they are the candidate pitches. But there is the possibility they are the pitches of instrumental signals. On the other hand, the number is much larger for a silent frame and at the same time the peaks magnitude is smaller compared to vocal frame, as the peaks are all originated from noise with low energy. The spectral transform of vocal and silent frame is depicted in Fig. 2.

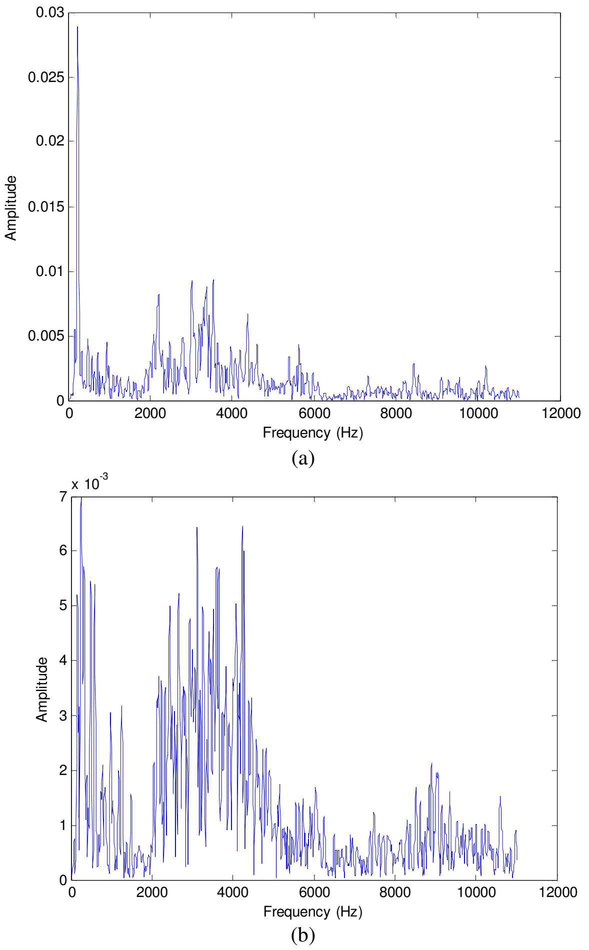
The noisy peaks have a negative effect on the correctness of the system, so a peak filter step is executed before salience computation. As a precursor of voice detection, an excellent filter will drop the generation of nonvocal melody contours obviously.

Spectral peaks are often disposed using the highest spectral peaks. Peaks with a magnitude more than 80 dB below the highest spectral peak in a frame are not considered [12]. Noisy peaks may be deleted, but the instrumental and harmonic peaks still exist. This will influence the speed of salience computation dramatically.

The aforementioned method is substituted with a two-level filter one. First, peaks below a threshold factor ρ of the highest peak are filtered out. Second, increase the value of ρ if the left peaks number is still large:

$$\begin{cases} \frac{leftpeak}{\alpha_1}\rho, & \text{if } \alpha_1 < leftpeak < \alpha_2, \\ 0, & \text{if } \; leftpeak \geq \alpha_2. \end{cases} \qquad (2)$$

If the left peaks number is larger than α_2, we just change ρ to zero. This means the frame has no voice and just is a silent frame since there are no salient peaks.



(a)

(b)

Fig. 2. Spectral transform of vocal and silent frames: a) – vocal frame, b) – silent frame.

After the two-level filter, the number of left peaks is approximately stable to α_1. Through the modification of α_1, the granularity of left peaks will be modified easily.

The range of human fundamental frequency is from 50 Hz to 1.1 kHz; the formant frequency can be expanded to as high as 10 kHz. But the energy is really small when the harmony reaches to five multiples. And the peaks above fifth harmony have little influence to the ultimate result. The peaks which are smaller than 5 kHz are used through experiment for effectiveness.

To select the best parameters for the final result, we use the grid search to find the optimal parameters for ρ, α_1, α_2(0.2, 16, and 64 respectively).

3) Salience function computation

After filtering the spectral peaks, the candidate pitch is often among the left peaks. But there will be wrong situations sometimes: the pitch is filtered because of low energy which often results from masking effect. This could be averted through the computation of pitch salience.

The salience computation in our system is similar to [9], where the salience of a given frequency is computed as the sum of the weighted energies found at harmonics of that frequency. Unlike [9], the pitch range is reduced from 90 Hz to 1.44 kHz and the bin number is reduced to 480 from 600 simultaneously. The pitch range includes four octaves. Given a frequency f in Hz, its corresponding bin B(f) is changed to

$$B(f) = \left\lfloor \frac{1200 \cdot log_2\left(\frac{f}{90}\right)}{10} + 1 \right\rfloor. \tag{3}$$

Because of the two-level filter of spectral peaks, the salience function is redefined as

$$S(b) = \sum_{h=1}^{N_h} \sum_{i=1}^{I} g(b,h,f_i) \cdot a_i \, , \tag{4}$$

where a_i is the amplitude of spectral peaks; f_i is the frequency of spectral peaks; N_h is the harmonics number considered; $g(b,h,f_i)$ is the weighed function to a given frequency.

The reason for promoting the basement of lower pitch is that there is little possibility most of our singing voice F0 can reach to a very low level. At the same time, a lot of melody contours which belong to instruments will be filtered. Some instruments tend to have a low fundamental frequency. The shrinking of the pitch range is advantageous to reduce the execution time of the system dramatically since the salience function computation is the most time-consuming part of the system.

B. Melody identification

In the context of melody identification, the problem is to decide which candidate pitches belong to the melody, and to detect whether the melody is active or silent at each frame.

At this stage, some systems simply decide a single best melody pitch at every frame and do not attempt to form them into higher note-type structures [1], [13]. However, some systems track the pitch candidates and then group the candidates to contours—time and pitch continuous sequences of salience peaks [9], [10], [11], [14]. Recently, more and more systems use the latter method as it can give more accurate result from MIREX held in 2011. The reason could be that more information could be extracted from contours which will be exploited to select the correct melody pitch. So we also adopt the latter approach.

1) Pitch tracking

Before the tracking process is carried out, nonsalient pitch candidates are filtered out to minimize the creation of contours belong to instrument or noise. This procedure is also a two-level filter [9]: first, filter out the peaks just like in spectral peaks filter; drop the peaks whose salience is below a threshold factor τ_+ of the salience of the highest peaks. Second, calculate the salience mean μ_s and standard deviation σ_s of all the left peaks in all frames. Then filter out the peaks whose salience is below $\mu_s - \tau_\sigma \cdot \sigma_s$. τ_+ andτ_σ are both experimentally set to 0.9.

After filtering the pitch salience, a set of pitch contours are created using heuristics in terms of the auditory scene analysis [15]. The key of generating contours is based on the following regularity — Gradualness of change:

- A single sound tends to change its properties smoothly and slowly;
- A sequence of sounds from the same source tends to change its properties slowly.

Based on the regularity, the contours are generated by adding the similar pitch to a contour from adjacent frames. An example is illustrated in Fig. 3. The polyphonic signal is with duration 12 s. There exist so many contours because of the instrument and harmonics. The melody could only be identified hazily.

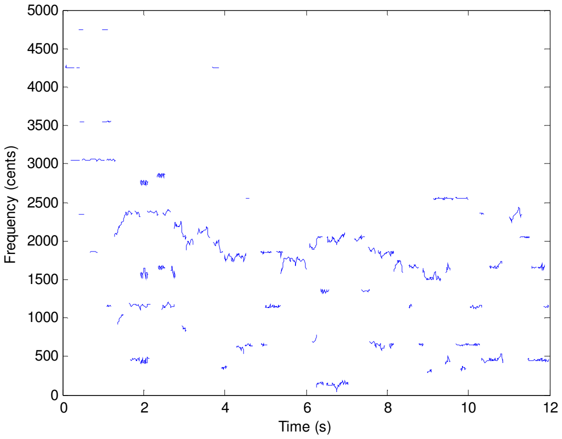


Fig. 3. All the melody contours, including vocal and nonvocal contours.

2) *Pitch contour characterization*

After creating the contours, the remaining problem is to choose the correct contours which belong to the vocal melody. Actually, this is the most difficult part of the vocal melody extraction. Using the pitch contours, a serious of characteristics will be proposed. All the characteristics are defined based on the pitch, length and salience. In addition to such characteristics computed directly using pitch and salience, there exist two other complicated features: vibrato and tremolo which are calculated using STFT on contours. In our system, the characteristics computed for each contour are comprised of the characteristics in [9] and tremolo. Tremolo which is similar to vibrato refers to the periodic variation of intensity, or amplitude modulation [16]. The characteristics are quite intuitive and easy to compute except vibrato and tremolo. But sometimes the characteristics alone cannot give us more information to the insight of the truth. The distributions of the characteristics are calculated to get more information with respect to the difference between vocal and accompaniment melodies [9]. Although most distributions have no discriminations at most times, the contour salience mean and standard deviation distributions reveal great discrimination ability.

3) *Singing voice detection*

As an independent field, singing voice detection usually extracts a set of audio features from the audio signal and then uses them to classify frames using a threshold method or a statistical classifier [16].

As for vocal melody extraction using pitch contour characteristics, the problem of singing voice detection can be simplified to distinguish the vocal contours from all the contours. Methods originally used to detect the presence of singing voice can be migrated here. Hsu [11] applies the method by utilizing the vibrato and tremolo. Salamon uses the distributions of characteristics to filter out the nonvocal melodies. We propose a three-step filter to filter out these melodies.

Before going the singing voice detection, the contours with short duration (less than 60 ms) are excluded in this stage because they are more likely to be produced by some percussive instruments or unstable sounds.

From the contour characteristics, it will be seen the nonvocal contours tend to have a smoother trajectory since they have a smaller pitch standard deviation. Using this feature, the contours can be filtered out with a low pitch standard deviation (σ<20) and have a long length (l>10). This procedure as the first step of singing voice detection will improve the final accuracy through deleting more nonvocal contours.

We achieve two different aforementioned strategies for singing voice detection kernel to compare their effectiveness and correctness. Neither of the two strategies will filter out all the nonvocal contours. There is a trade-off in selecting the parameters to try best to save more vocal contours and less nonvocal contours.

As the third filter of singing voice detection, contour pitch mean ($\bar{C}_P$) and pitch standard deviation ($\sigma_{\bar{C}_P}$) of all the left contours are used to drop the contours which have a low pitch mean since the fundamental frequency of the instrumental contour is often low. If the pitch mean of a contour is lower than $\bar{C}_P - \nu * \sigma_{\bar{C}_P}$, the contour is dropped. Parameter ν is experimentally set to 1.2. The contours left after the singing voice detection is depicted in Fig. 4.

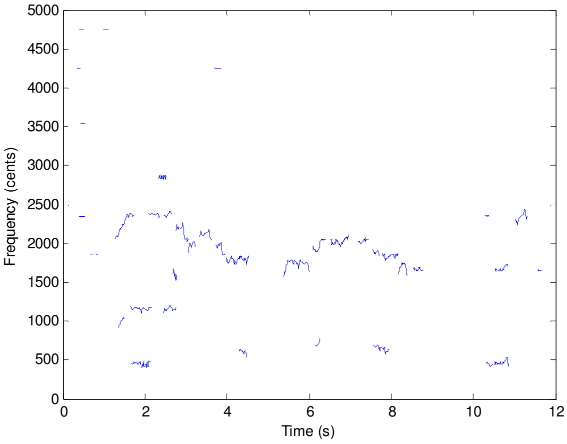


Fig. 4. Remaining contours after singing voice detection

The contour number is much less compared with Fig. 3 since the singing voice detection step drops most nonvocal contours, and the melody is revealed more clearly. Of course, there are some octave and exceptional contours left, too. These contours will be excluded in the next stage.

C. *Post processing*

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, commonly referred to as octave error. Various approaches have been proposed for the minimization of octave errors: [14] eliminates one of the harmonics if its salience is less than 40 % of the most salient pitch contour if they differ by one octave, 20 % if they differ by two octaves, and so forth. The harmonic contours are just deleted in light of the assumption that the lowest frequency contour within a frame is the vocal F0 partial in [11]. Salamon iteratively calculates a “melody pitch mean” to solve this problem. Unfortunately, none works well in all conditions. The contours with weaker energy or the higher frequency may be the real melody sometimes. Among all ways, the way calculating a “melody pitch mean” works the best since it reflects the melody trend. So we adopt this way in our system to delete harmonics and pitch outlier.

At last, the melody is selected from the remaining contours.

In most times, there lefts only one pitch at every frame, so just use this one as the ultimate pitch. If there are more than one pitches left, the melody is selected from the pitches with bigger salience sum. If no contour is present the frame is regarded as unvoiced.

The ultimate melody is provided in Fig. 5.

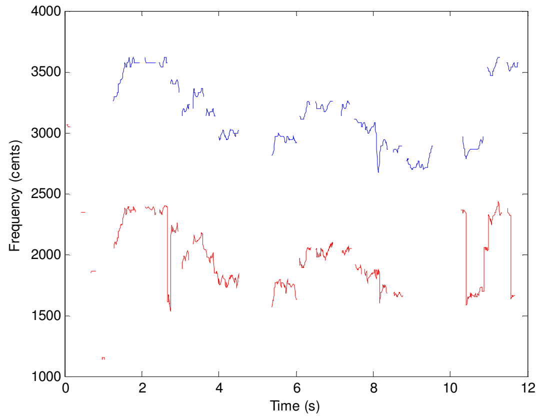


Fig. 5. Final melody extracted by our system (red) and the ground truth (blue, shifted up one octave for clarity)

The red melody contour is the final melody estimated by our system, the blue melody contour is the ground truth (shifted up one octave for clarity). It can be seen that the extracted melody is very similar to the ground truth. But there also exists an apparent flaw: the vocal melody is excluded between 9-10 s.

III. EVALUATION

In this section, we present an experimental evaluation for our vocal melody extraction. First, the difference between the approach based on the vibrato/tremolo and the one based on the contour characteristics distributions is evaluated. Then, next is the effectiveness of every step of the three-step singing voice detection. At last, the distribution of the overall accuracy is analyzed.

A. *Evaluation Set*

There are many datasets used in MIREX. The number of songs in three datasets is small (i.e. ADC2004, MIREX05, and MIREX08), one is large (i.e. MIREX09). From [17], [18], ADC04, MIREX05 and MIREX08 collections are unstable because the performance variability is due to song difficulty differences rather than algorithm itself. As such, results from these collections alone are expected to be unstable, and therefore evaluations that rely solely on one of these collections are not very reliable. So it's reasonable to discard these datasets, and use MIR-1K which is a publicly available dataset proposed in [17]. It has 1000 song clips with a duration ranging from 3 to 12 seconds, and the total length is up to 133 minutes. There are 19 singers, 8 females and 11 males, most of them are amateurs with no professional training.

B. *Evaluation metrics*

The algorithms in MIREX are evaluated in terms of five metrics, including Voicing Recall Rate (VRR), Voicing False Alarm Rate (VFAR), Raw Pitch Accuracy (RPA), Raw Chroma Accuracy (RCA), and Overall Accuracy (OA). The detail is depicted in [3]. Since RCA measures the capacity of the algorithm to remove octave melodies, and is no use to the computation of the overall accuracy. It's harmless to neglect this metric.

C. *Evaluation for singing voice detection kernel*

We evaluate two different strategies for singing voice detection: vibrato/tremolo and characteristics distribution. The difference of overall accuracy between them is huge, and the result is shown in table 1. It is obvious to see that the VFAR is much higher using vibrato/tremolo than that using characteristics distribution although the VRR is high. This proves that the only use of vibrato/tremolo is not enough, more complicated discrimination method should be used, just like in [11]. What's more, taking the algorithm complexity into consideration, the latter is more advantageous than the former.

TABLE I. THE RESULT OF DIFFERENT STRATEGIES TO SINGING VOICE DETECTION

Strategies	**VRR**	**VFAR**	**RPA**	**OA**
Vibrato/Tremolo	0.92	0.61	0.70	0.61
characteristics distribution	0.83	0.21	0.72	0.74

D. *Evaluation for three-step singing voice detection*

We verify the impact of every filter on the overall accuracy through getting rid of each of them. Figure 6 shows the overall accuracy results of singing voice detection. The blue bar (first) shows the result of only using contour characteristic distributions to detect the singing voice. The red bar (second) shows the result when the first step is added. The green bar (third) shows the result when the third step is added. The purple bar (forth) shows the result which is achieved by using all the three steps. It is clear that the proposed three-step singing voice detection can improve the overall accuracy considerably.

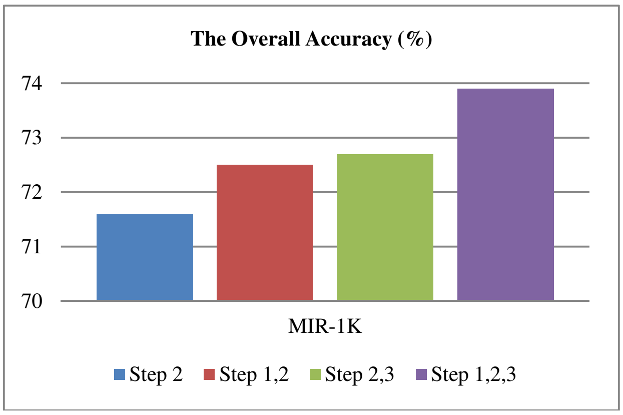


Fig. 6. The overall accuracy of MIR-1K with different filter strategies.

The dataset mainly used in MIREX is MIREX09 dataset, which is constructed using the similar way like in the construction of MIR-1K dataset. So the results using the MIREX09 could be compared with our system on some level. The best result using MIREX09 mixed at 0 dB SNR in 2012 is 69 %, which is obviously smaller than the results (78 %) in 2011. It means this field needs more research to get a better result. By contrast, our system gets a 74 % overall accuracy.

Although it's smaller than the best result in 2011, the runtime descends considerably. The spectral peaks filter can drop the peaks number from more than 30 to less than 10.

The shrinking of pitch range in salience computation can also improve the efficiency. What's more, we just neglect some time consuming procedures such as equal loud filter and frequency correction. So our system is about 4 times faster than the one with best overall accuracy.

E. The OA distribution of MIR-1K dataset

Although the overall accuracy of our system is lower than the best approach submitted to MIREX in 2011 (about lower 1 %). Our system runs much faster, and through the overall accuracy distribution, it can be seen our system is actually better than expected. The number whose OA is greater than mean OA, than 70 %, than 60 % is 54 %, 66 % and 87 % respectively, as depicted in Fig. 7. That means most music in the dataset have an overall accuracy greater than 60 %. This result is sufficient for the actual applications, such as Query by Humming.

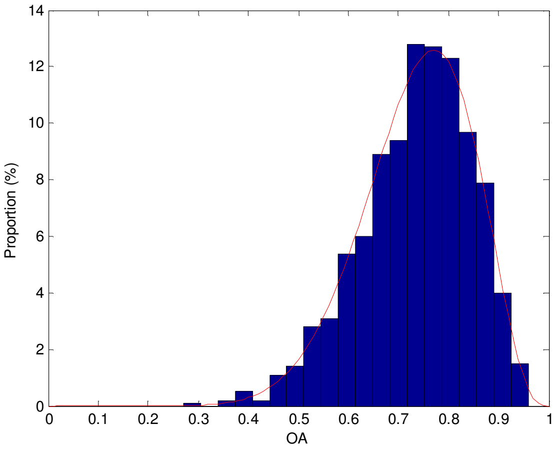


Fig. 7. The overall accuracy distribution of all the music in MIR-1K dataset.

IV. CONCLUSIONS

Melody extraction from polyphonic music is a valuable problem because the melody can be used in many valuable applications, especially in Query by Humming. However, the relative long extraction time and the low accuracy limit its extensions. In this paper, we proposed a system for automatic vocal melody extraction from polyphonic music recordings. The vocal and nonvocal melodies are mainly discriminated using a three-step singing voice detection. Although the shrinking of spectral peaks number and the range of pitch salience, the overall accuracy does not reduce so much and moreover the speed is improved many times. The distribution of the overall accuracy of all the songs in the dataset manifest that our system can be applied into actual applications.

In the future, more work can be done on singing voice detection to further improve the overall accuracy and use the melody extracted to Query by Humming. Actually, the problem of singing voice detection can be fallen into the problem of classification. So the classic methods of classification, e.g. SVM, neural network, could be used to classify the vocal and nonvocal contours, and maybe give a better result.

REFERENCES

[1] M. Goto, "A real-time music-scene-description system: predominant-f0 estimation for detecting melody and bass lines in real-world audio signals", *Speech Communication*, vol. 43, pp. 311–329, 2004. [Online]. Available: http://dx.doi.org/10.1016/j.specom.2004.07.001

[2] R. B. Dannenberg, W. P. Birmingham, B. Pardo, N. Hu, C. Meek, G. Tzanetakis, "A comparative evaluation of search techniques for query-by-humming using the MUSART testbed", *J. of the American Soc. for Inform. Science and Technology*, vol. 58, no. 5, pp. 687–701, 2007. [Online]. Available: http://dx.doi.org/10.1002/asi.20532

[3] G. E. Poliner, D. P. W. Ellis, F. Ehmann, E. G´omez, S. Steich, B. Ong, "Melody transcription from music audio: Approaches and evaluation", *IEEE Trans. on Audio, Speech and Language Process.*, vol. 15, no. 4, pp. 1247–1256, 2007. [Online]. Available: http://dx.doi.org/10.1109/TASL.2006.889797

[4] A. Ozerov, P. Philippe, F. Bimbot, R. Gribonval, "Adaptation of bayesian models for single-channel source separation and its application to voice/music separation in popular songs", *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 15, no. 5, pp. 1564–1578, 2007. [Online]. Available: http://dx.doi.org/10.1109/TASL.2007.899291

[5] E. Vincent, M. Plumbley, "Predominant-F0 estimation using Bayesian harmonic waveform models", *MIREX Melody Extraction Abstracts,* London, U.K., 2005.

[6] A. Klapuri, "Multiple fundamental frequency estimation by summing harmonic amplitudes," in *Proc. of 7th Int. Conf. on Music Inform. Retrieval,* Victoria, Canada, 2006, pp. 216–221.

[7] G. Poliner, D. Ellis, "A classification approach to melody transcription," in *Proc. of Int. Conf. Music Inf. Retrieval.,* London, U.K., 2005, pp. 161–166.

[8] J. S. Downie, "The music information retrieval evaluation exchange 2005–2007: A window into music information retrieval research", *Acoustical Science and Technology*, vol. 29, no. 4, pp. 247–255, 2008. [Online]. Available: http://dx.doi.org/10.1250/ast.29.247

[9] J. Salamon, E. G´omez. "Melody extraction from polyphonic music signals using pitch contour characteristics", *IEEE TASLP*, vol. 20, no. 6, 2012.

[10] K. Dressler, "Sinusoidal extraction using an efficient implementation of a multi-resolution FFT," in *Proc. of 9th Int. Conf. on Digital Audio Effects (DAFx-06),* Montreal, Canada, 2006, pp. 247–252.

[11] C. Hsu, J. R. Jang, "Singing pitch extraction by voice vibrato/tremolo estimation and instrument partial deletion," in *Proc. of 11th Int. Soc. for Music Inform. Retrieval Conf., Utrecht,* The Netherlands, 2010, pp. 525–530.

[12] J. Salamon, E. G´omez, J. Bonada, "Sinusoid extraction and salience function design for predominant melody estimation," in *Proc. of 14th Int. Conf. on Digital Audio Effects (DAFx-11),* Paris, France, 2011, pp. 73–80.

[13] V. Rao, P. Rao, "Vocal melody extraction in the presence of pitched accompaniment in polyphonic music", *IEEE Trans. on Audio Speech and Language Process.*, vol. 18, no. 8, pp. 2145–2154, 2010. [Online]. Available: http://dx.doi.org/10.1109/TASL.2010.2042124

[14] R. P. Paiva, T. Mendes, A. Cardoso, "Melody detection in polyphonic musical signals: Exploiting perceptual rules, note salience, and melodic smoothness," *Computer Music J.*, vol. 30, pp. 80–98, 2006. [Online]. Available: http://dx.doi.org/10.1162/comj.2006.30.4.80

[15] A. S. Bregman, "Auditory Scene Analysis: Hearing in Complex Environments", *Thinking in Sound*, pp. 10–13, 1993.

[16] L. Regnier, G. Peeters, "Singing voice detection in music tracks using direct voice vibrato detection", in *Proc. of the IEEE ICASSP*, 2009, pp. 1685–1688.

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

[18] J. Salamon, J. Urbano, "Current Challenges in the Evaluation of Predominant Melody Extraction Algorithms", in *Proc. of the 13th International Society for Music Information Retrieval Conference (ISMIR 2012)*, 2012.