A method of subject extension pitch extraction for humming and singing signals

Zhang Jinghui, Yang Shen, Wu Huahua
School of Information Science and Engineering, Wuhan University of Science and Technology

***Abstract*—Pitch extraction is a key task of ahumming query system.The purpose of this paper isto find a method to extract pitch accuratelyfortwo input modes, humming and singing.According to the characteristics of Chinese pronunciation, this paper presentsa new method, namely the Pitch Extraction of Subject Extension.According to the differencesbetween humming and singing signals, this method respectively chooses energy threshold and ratio of energy and entropy to detect endpoint of notes.The candidate method and the shortest distance method are used to determine the pitch periods of the voiced segments, and subject extension method is used to determine the pitch periods of the voiced/unvoiced mixed segments. Finally this algorithm is implemented in a small database, andis compared withother similar algorithms. Experiments show our algorithm is more accurate and robust.**

Keywords—humming query; pitch extraction; endpoint detection; subject extension.

I. INTRODUCTION

Humming query is a typical application of content-based music retrieval and it usually has two kinds of input methods[1]: humming and singing. Humming is a limited input, requiring users to hum a tune by “dada”[2]. However, singing is a more friendly and more casual input mode, users can sing lyrics directly. The pitch detection for the input signal is an important step in thehumming query system.

In recent years, with the development of wavelet transform, the pitch extraction method aiming at humming and singing has made great progress. Upadhyay et al has proposed a method based on the variational mode decomposition and Hilbert transform for the instantaneous pitch frequency extraction[3]. OnurBabacan et al has compared the recent mainstream pitch extraction method[4]. The literature[5]is introduced the most widely used ACF method which is also used in this paper, and the feature of ACF is simple and with higher precision.

The sound of humming or singing are both producedfrom the vibration of vocal cords, but the two kinds of production are not the same. The designed pitch extraction algorithm in the paper can apply to different kinds of input signal by analyzing their different features. The following will be described in detail.

II. THE METHOD OF PITCH EXTRACTION

The ideas of this paper comes from pitch extraction in Chinese speech languages. Based on this, the paper proposes the Subject Extension approaches based on a single note of humming or singing signal. The main steps of the algorithm are shown in Fig.1.

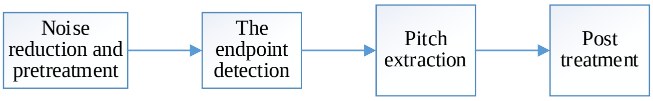

Fig. 1. the main steps of the algorithm.

A. Noise reduction and pretreatment

In the practical environment, the performance of recording is usually affected by the background noise[6].So, the first thing is reducingthe noise of the sampling frequency of 8000 Hz‘s input signal by the spectral subtraction. Then a pre-emphasiz filter is meant to compensate the high frequency components attenuation[7]. A prefilter using elliptic bandpass filter, which the boundary frequency is 60Hz and 3400Hz, is designed to remove the 50Hz power frequency noise and the interference source frequency over half of the sampling frequency[8]. Finally the sampling frequency is added window and frame partition by hamming window, whose length is set 40ms and frame shift is set 10ms[9].

B. Endpoint detection

The composition of humming or singing can be divided into unvoiced and voiced parts, and only the voiced part can be extractedan accurate pitch frequency. The humming signal usually requires to be a plosive pronunciation, so the difference between adjacent time-domain waveform amplitude is very large, it can directly use the energy threshold to find the starting position of each note. But for the singing signal, because the singer is maybe non-professional or the lips radiation effects, the record of singing signal always takes noise and other uncertain influence. The simple use of energy for endpoint is hard to meet the requirements of actual note segmentation. This would requires the Energy entropy ratio which can still keep good performance in low signal-to-noise ratio conditions in detecting the voiced / unvoiced ingredients. The process of detection becomes more complex, andbetween the voiced / unvoiced ingredients also must be subdividedto a transition section. The specific division is shown in Fig.2.

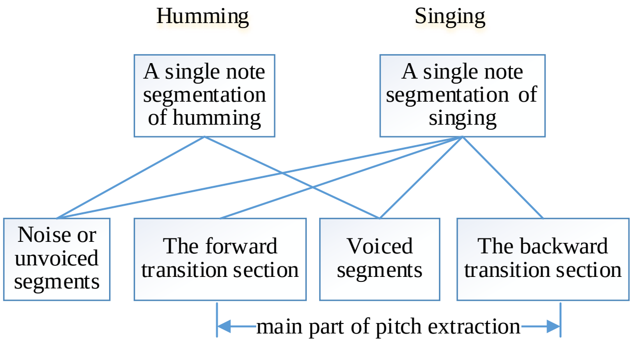
Humming
Singing
A single note segmentation of humming
A single note segmentation of singing
Noise or unvoiced segments
The forward transition section
Voiced segments
The backward transition section
main part of pitch extraction

Fig.2. the specific division for humming and singing

C. *Pitch extraction*

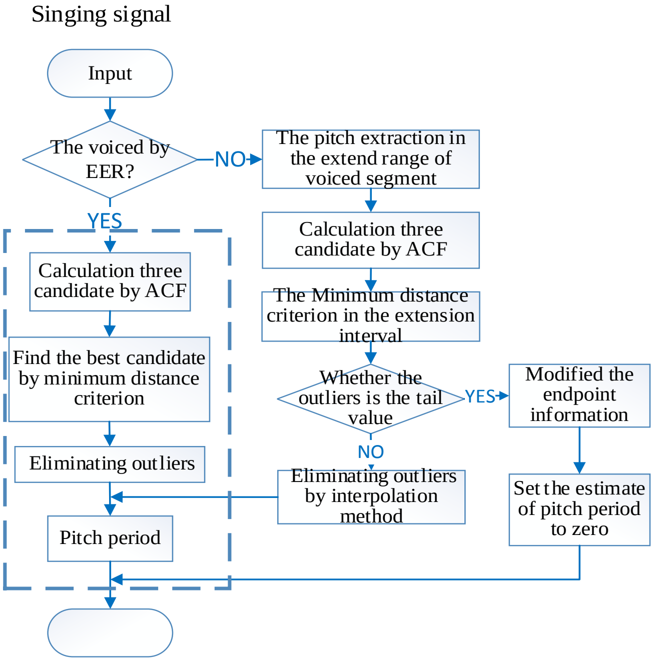
Singing signal
Input
The voiced by EER?
NO
YES
Calculation three candidate by ACF
Find the best candidate by minimum distance criterion
Eliminating outliers
Pitch period
The pitch extraction in the extend range of voiced segment
Calculation three candidate by ACF
The Minimum distance criterion in the extension interval
Whether the outliers is the tail value
YES
NO
Modified the endpoint information
Eliminating outliers by interpolation method
Set the estimate of pitch period to zero

Fig.3. the main process of pitch extraction of singing

For the humming signal, after the endpoint detection, The pitch detection directly using ACF on audio segment, then use the shortest distance criterion and interpolation for elimination outliers. The pitch extraction of singing voiced segment is similar to the humming, but the extension of voiced interval is more complex. The main process of pitch extraction of singing as shown in Fig.3.

1) *The candidates of pitch values*

Because the division of the voiced segment is not very strict and music signals are very complex, there will be outliers even in voiced parts to extract the pitch. When calculated the data of each frame in voiced parts, we would find three peaks and their corresponding positions in the range of p_{min} ~p_{max} (p_{min} andp_{max} represent maximum and minimum estimated period respectively). The values of these three corresponding positions will be treated as pitch candidates and saved in the arrayP_{tk}, P_{tk} retained only the positions and its sequences in accordance with the peak amplitude. The candidate values of pitch period as shown in Fig.4.

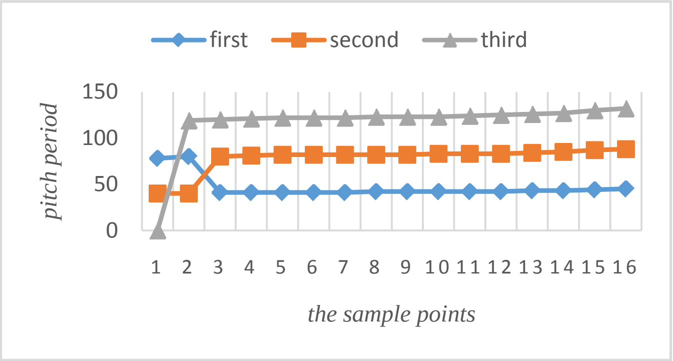

Fig.4. candidate values of pitch period

After obtained the candidates, we will set a reasonable confidence interval. We set the position of the maximum ACF peak in the ithframe as $Kal = P_{tk}(1,i)$, calculate its means $meanx1$ and standard deviations $segma1$, then the confidence interval is $[meanx1 - thegma1, meanx1 + thegma1]$.In most case, the standard deviations is very small, we don't need to deal with them further, but sometimes the standard deviations will be very large, this indicates that the Kal values are oscillation. The value of Kal in confidence interval is assigned to a parameterPam, and others are set to zero.

2) *The shortest distance criterion of pitch extraction*

In thePam, the nonzero partis marked as$span_non$, and we need to make the second confidence interval division for this part by calculating the mean $meanx2$ and standard tions$thegma2$of$Pam(pan_non)$ again. Now the second confidence interval is within$[meanx2 - thegma2, meanx2 + thegma2]$, then $Pamtmp$ is copied fromPam. If values in Pam are out in the secondconfidence interval, the $Pamtmp$ values are set to 0. Thus some samples of large deviationscould be abandoned.

For the points in $Pamtmp$ whose value is zero, we will search the best value between the front and back voiced segments by the shortest distance criterion. We assume that the ith frame in$Pamtmp$is nonzero, and the $i+1th$(or$i-1th$) frame is zero. The pitch candidate value of the $i+1th$ column is$Ptk(:, i+1)$. In the$Ptk(:, i+1)$, we search an element which distance from $Pamtmp(i)$ is minimized. Assumed that the position is mark asml, and we need to take $Ptk(ml, i+1)$ for further judgment: Requested the pitch difference between two adjacent frames is no more than $c1$ sampling periods ($c1$ is a threshold, ranging between 10 and 15). Formula (1) shows this constraint for judgment conditions.

$$\left|Ptk\left(ml,\ i+1\right) - Pamtmp(i)\right| \le c1 \qquad (1)$$

This method can correct some pitch periods. However, because the voiced segment division isn't very strict, it's possible to produce the transition interval of singing, or the voiced segments is inherently unstable. If the formula (1) isn't satisfied, the value of $Pamtmp(i+1)$ is find out only by meant of stacking or interpolation.

D. The pitch extraction in the extend range of voiced segment

1) Calculation of the extended range and length

Extension interval belongs to the singing section, but doesn't belong to the voiced segment. It's the overlapping part of the voiced and unvoiced. We has been obtained the information about endpoint and voiced segments by endpoint detection, but in the division of the unvoiced segments (voicing aliasing) there still exists unstable pitch information. When we have obtained the information about voiced segments, we begin to (before and after) extend the transition interval in the music signal, and the extending length is determined by the singing section and the voiced segments together.In the relationship between the singing section and the voiced segments we define as follows: in a singing section of song, the first voiced segments are both forward and backward extended, the rest of the voiced segments only extends posteriorly.

2) Autocorrelation function calculation

For each frame of extended interval we still extract three pitch candidates among them by the ACF. And the pitch values corresponding to the magnitude of the ACF peak were sorted in descending order and saved inP_{tk}. But the difference is that the periodic in the extended interval of music signals is not as strong as the voiced's, So the value of pitch extraction by ACF is very uncertain.The resulting in an extended interval value often meets the requirements off and on, and there are still mess data.

3) Minimum distance criterion

Because the periodicity performance in the extension interval becomes worse and the region of variable composition becomes more complicated, the most value of pitch extraction in the extended interval can't reflect the actual pitch value. When we are doing the pitch detection, the first thing is to find a suitable pitch period by the minimum distance criterion. If the value is not existed, it's required that the threshold between the different of two frames is no more than$c1$, or the different of more than two frame is no more than$c2$.

$$|P(i) - P(i+1)| \le c1 \qquad (2)$$

$$|P(i) - P(i+j)| \le c2, j = \pm 2, \pm 3, \ldots \qquad (3)$$

Sometimes because it's impossible to seek the adjacent frame that are less than $c1$ in the process of pitch extraction between the transition intervals, it only can look for by a frame or frames.

4) Post treatment

The value of pitch extraction in the extended interval by the shortest distance from the pitch detection, in the corresponding position of the three formant, can't meet the conditions of the numerical.So we can set the estimate of pitch period $Pkint(i)$ is zero, and the further processing has to wait until the end of calculation in the extended interval.There are three kinds of situations for $Pkint(i)$ in the extended interval: the head, middle and tail. For the head and the middle part of the extended interval, the post treatment are used in the interpolation method. And for the tail of the extended interval. It's considered that the inaccurate endpoint detection cased such results and unvoiced segment is divided into the voiced segment incorrect. So we will modified the starting point in the music signal.

III. EXPERIMENTAL RESULTS AND ANALYSIS

A. *Experimental results*

The experimental data was recorded in a quiet laboratory. 15 boys and girls were invited to be recorded a song through two methods of humming and singing. There are four record versions of this song: humming by a male, singing by a male, humming by a female, and singing by a female. In this paper the volunteers hummed or sang the lyrics “我一直都在流浪” from the song "Cruel Moonlight". The effect is shown in Fig.5 and Fig.6.

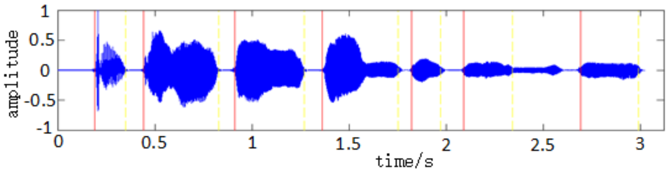

Fig.5. the endpoints of a humming signal by a female

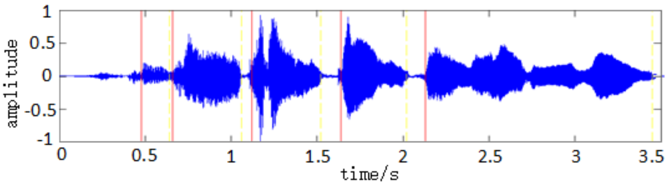

Fig.6. the endpoints of a singing signal by a female

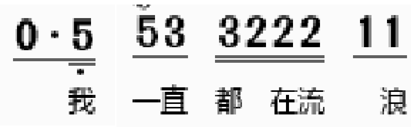

Fig.7: the musical notation of lyrics "我一直都在流浪"

The solid line shows the start of the voiced segments, and the dotted line indicating the end of the voiced segments. From Fig.5 and Fig.6, it can be seen that the results of humming signal are much better than the singing one. We have the singing section further to be divided into the voiced segments. The result is shown in Fig.8.

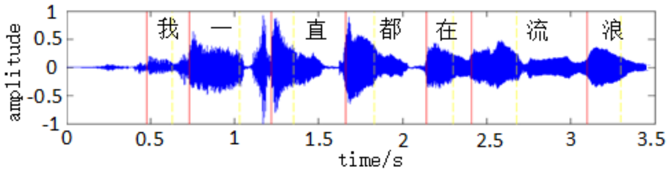

Fig.8. the voiced part segmentation for the singing signal

Bycomparing Fig.7 withFig.8, it can be seen that Pitch detection results of this algorithm is able to detect the pitch of the voiced segment very well.

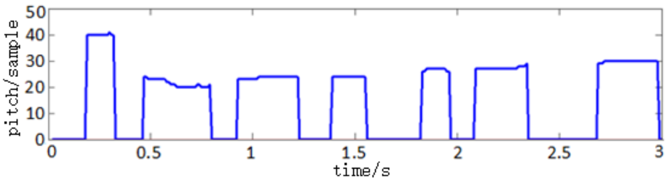

Fig.9. the pitch extraction result of the singing signals

The vertical coordinate of Fig.9 and Fig.10 is the sampling points, the sampling interval is 1/8ms. From Figure 9, it can be seen that the algorithm is able to extract the pitch very well from humming or singing signal. The result of Fig.10 (b) is come from the pitch extraction by this paper’s algorithm based on the Fig.10(a). It’s fortunate that the result of the Subject Extension of Pitch Extraction don’t appear the errors of most half frequency and double.

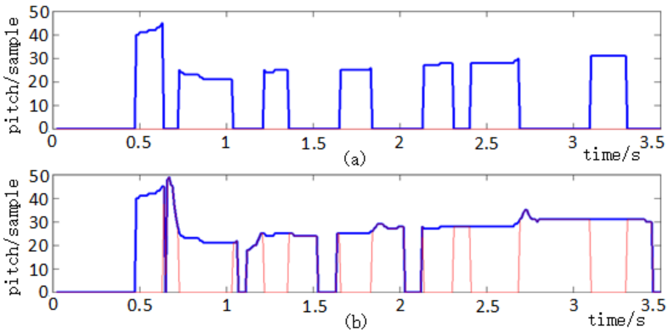

Fig.10. the pitch detection results ofthe singing signal. (a) The voiced fragment pitch periods (b) the pitch periods after the subject extension

B. *Compared with several method*

In order to test the performance of the proposed algorithm in this paper, it will be compared with three methods:① three-level cross-correlation method; ②the combination method of ACF and AMDF[10]; ③linear prediction and cepstrum method. Fig.11 and Fig.12 show the comparison result.It can be seen that for humming and singing signals, all of the method①,②,③ have some errors of multiple or half frequency. The smoothing filter can remove most of outliers, but there are still some unreasonable pitch periods. It’s obvious that the algorithm proposed in this paper have the best performance.

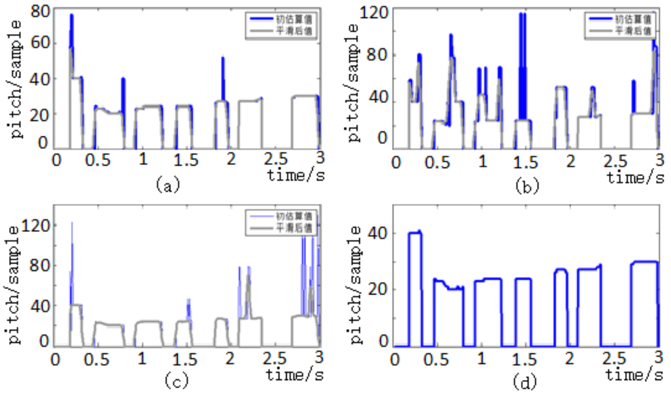

Fig.11. results of four pitch extraction algorithms for the humming signal by a female. Among them (a) method① (b) method② (c) method③ (d) the proposed method

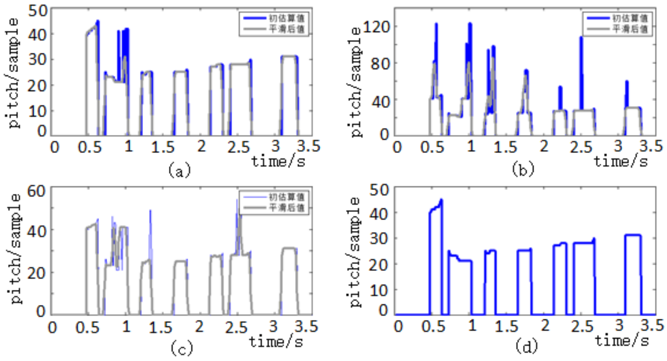

Fig.12. results of four pitch extraction algorithms for the singing signal by a female. Among them (a) method① (b) method② (c) method③ (d) proposed method

In order to verify the robustness of the proposed algorithm, we respectively add Gaussian white noises of SNR=-5dB, 5dB, 10dB and 15dB to original humming and singing signals. Fig.13 and 14 show the experiment results. For the humming signal Fig.13 show that the proposed method extracts pitch periods accurately under the conditions of SNR=5dB, 10dB and 15dB, and there is no obvious outliers. However all of the other methods are appeared outliers in some extent. It also can be seen that the detect effects of this four methods are decreased quickly, but the proposed algorithm still can basically ensure the accuracy of pitch detection.

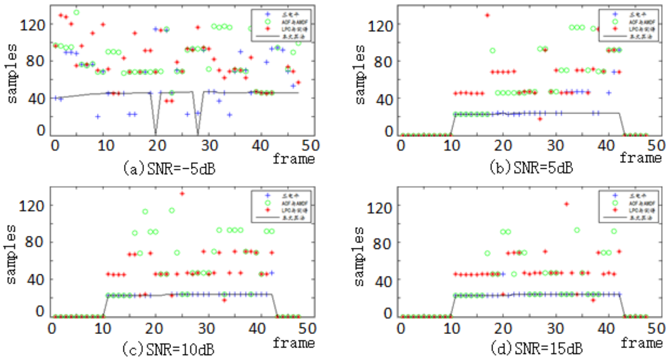

Fig.13. the results of four pitch extraction algorithms under different SNR.

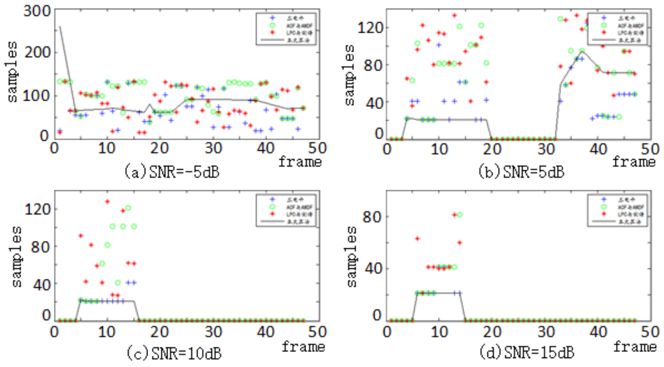

Fig.14. the results of four pitch extraction algorithms under different SNR.

For singing signal Fig.14 show that, under the condition of SNR=10dB and 15dB, the proposed algorithm is substantially unaffected by the noise. When SNR=5dB, it still can ensure the accuracy of pitch extraction, but the voiced interval is extended excess. Because the voiced/unvoiced segments can’t be estimated correctly under the condition of SNR=-5dB, the process of pitch extraction is influenced.

IV. Conclusion

In this paper, we have present a method of subject extension pitch extraction for humming and singing signals. This method respectively choose energy threshold and ratio of energy and entropy to divide into the voiced and unvoiced segment. Then ACF method is used to extract the pitch periods of the voiced segments. The candidate method and the shortest distance method are used to improve the accuracy of pitch detection. Finally this algorithm is compared with other similar algorithms. Experiments showed our algorithm hadhigher accuracy and better anti-noise performance.

V. REFERENCES

[1] Li, Zhou Mingquan, Xia Xiaoliang etc.. Improved pitch detection method and in music retrieval application [J]. Computer engineering and applications, 2011, 47 (6): 127-130.

[2] Chhayani, N.H, Patil, H. development of corpora for person recognition using humming, singing and speech[C] Oriental Cocosda held jointly with 2013 Con-ference on Asian spoken language research and Evaluation. IEEE 2013:1-6.

[3] Upadhyay A, Pachori R B. A new method for determination of instantaneous pitch frequency from speech signals[C]// Signal Processing and Signal Processing Education Workshop. IEEE, 2015.

[4] Babacan O, Drugman T, D'Alessandro N, et al. A comparative study of pitch extraction algorithms on a large variety of singing sounds[C]// Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2014:7815-7819.

[5] Rabiner L. On the use of autocorrelation analysis for pitch detection[J]. Acoustics Speech & Signal Processing IEEE Transactions on, 2012, 25(1):24-33.

[6] Yuan Gang, Liu Zhikun, Tang Xiaoming, et al. Study on the mechanism of speech enhancement algorithms [C].2008 Annual Conference on communication theory and signal processing.2008:402-407.

[7] Xing Weili. Research and implementation of content based audio retrieval technology [D]. Northwestern University, 2004

[8] Li Xuelong. Design and implementation of music retrieval system based on melody matching [D]. Beijing University of Technology, 2010

[9] Zi Lin. Humming music retrieval system research and design based on [D]. University of Electronic Science and technology, 2011

[10] Li Zhijun, Yin Xia. An improved algorithm [J]. audio technology, pitch detection based on AMDF and ACF 2011, 35 (1): 50-52.