Adaptive onset detection based on instrument recognition

Bing Zhu[1], Jiayue Gan[2], Juanjuan Cai[3], Yi Wang[4], Hui Wang[5]
Communication University of China,Beijing, China
zhubing1218@163.com[1], jiayuegan@126.com[2]

***Abstract*—onset detection is the foundation and key to high-level audio processing like music retrieval and transcription. Research shows that the detection algorithm is associated with instrument category, and high accuracy can be achieved in instrument recognition studies. Thus the adaptive detection system based on instrument recognition was proposed in this paper. The system uses HMM classifier to identify input audio falling into four categories, adaptively adopts suitable detection algorithm for each type, and output onset times in the end. The experiment results show that onset evaluation values, such as the F-measure value, have been improved in the system.**

Keywords-HMM; instrument recognition; onset detection;

I. INTRODUCTION

Music features, like pitch, duration, rhythm, are associated with notation system. And onset detection, which detects the beginning point of the attack of the note, is the key to music notation, and has increasingly become an important and difficult task in musical signal processing.

A lot of research found that onset detection had a strong relationship with musical instrument. In 2005, Juan Pablo Bello [1] introduced some detection methods. In the experiment, he classified the instrument into three categories, PNP(pitched non-percussive), PP(pitched percussive), NPP (non-pitched percussive), and evaluate their performance separately. In 2006, Simon Dixon [2] described spectral flux, phase, complex domain phase detection approaches, and compared their performance in the same dataset. In 2007, Ruohua Zhou [3] proposed a novel method based on classification: He firstly determined the onset was 'hard' or 'soft'; Then for the 'hard' type whose energy changed obviously, he used energy detection method, otherwise, for the 'soft' one, choose pitch detection algorithm. In 2008, Olaf Schleusing [4] figured out high-frequency will bring noise interference, so he used sub-band approach and correlation value under 8 KHZ to detect onset. Results showed it can improve the accuracy of PNP instrument detection. In 2010, SHI Xiang-Bin [5] added phase difference information in second order difference of phase, also improve the accuracy of the detection. The same year, Stylianou Y [6] made a experiment on wind, bowed string instrument to compare group delay phase, spectral flux, fused method. In 2012, Carlos Rosao [7] reviewed some processing techniques that can be employed to enhance the performance, like the norm, median filter etc, also gave the result of 6 algorithms on Bello dataset to compare their influence.

Some researchers study music instrument classification. A Eronen [8] used mel-frequency cepstral coefficient with linear prediction coefficient to recognition 16 kinds of orchestral instrument. The accuracy is high up to 77%. In 2004, B Kostek [9] made artificial neural networks (ANNS) as decision machine, extracted frequency envelope distribution (FED) feature to classify 12 types of instrument, average precision reached to 70%. In 2008, JJ Deng [10] mainly doing research on data selecting in training set, not only reduce redundancy but also got 86.9% accuracy rate based on support vector machine(SVM).

The possibility of instrument identification, with the detection difference in different types of instrument of onset method, make us propose a new approach that use appropriate algorithm after instrument is recognized. Firstly, we extract MFCC feature, build models. Then figure out the type of instrument by HMM. Once the type is known, we can choose the best fit detection method. Finally output the onset times. The proposed algorithm and the single detection are evaluated on mix dataset. Clearly the values, such as F-measure etc, are improved.

The paper is organized as follows: In section II, the workflow of the proposed system, HMM instrument classifier, with 5 kinds of onset detection algorithm will be introduced. In section III, the details of database, evaluation method and experiment result are presented. In section IV, the discussion, conclusion, with the future work will be outlined.

II. THE DESIGN AND IMPLEMENTATION OF THE SYSTEM

The system will classify the instrument into 4 types, piano, wind, drum, bowed string, using HMM classifier firstly. Then self-adaptively choose appropriate detection for classified music, outputting exact onset times. The whole workflow is shown as Fig.1.

In general, onset researchers usually separate the music into NPP, PP, and PNP to evaluate their performance. Piano belongs to PP (pitched percussive) category. We just analyze part of the instruments, so the class is refined. Drum belongs to NPP (non-pitched percussive) category, string belongs to PNP (pitched non percussive) category. Wind is bowed instrument. Detections for them is a more complicated and difficult task. We set them a group alone, which will be more helpful for research and detection. Energy, spectral flux, complex domain phase, wrapping-compensated correlation and conditional independent component analysis (ICA), 5 algorithms will be experimented on 4 types of instrument, to find the most fitting method for each type. Then it will be connected with instrument identification block, which will make our detection efficient and precious.

978-1-4799-2186-7/14/$31.00 ©2014 IEEE

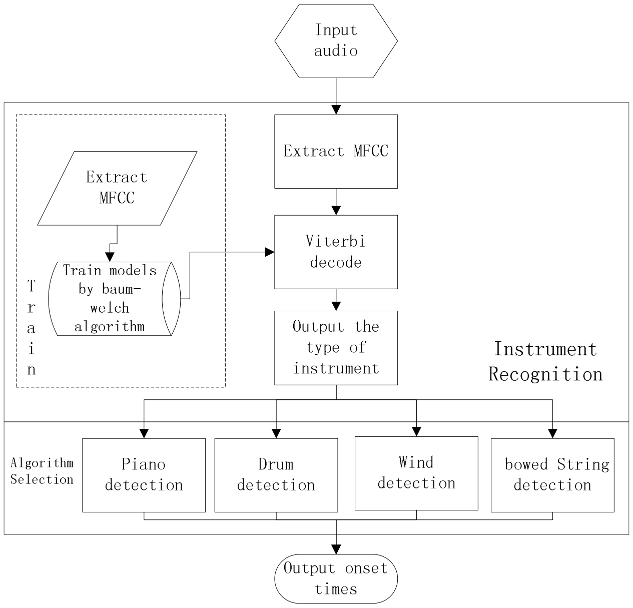

Fig.1. the workflow of detection

A. Instrument recognition

HMM (Hidden Markov Model, HMM) is a statistical analysis model, which can be described by two state set, the hidden state S and observable state O, and 3 probability matrix, including the initial state probability matrix π , hidden state transition probability matrix A and observation state transition probability matrix B. We usually use λ = (A, B, π) triples to represent a simple hidden Markov model [11].

To accurately classify the audio, the feature we extracted seems important. It should be able to adequately represent time domain and frequency domain characteristics and feature, also be robust to changes in the environment and instrument. We use 24-dimensional MFCC coefficients that are extracted cepstral in Mel scale frequency domain as musical properties [12]. Mel scale describes the nonlinear feature in the perception of human auditory. It's an effective feature in instrument recognition. We can calculate MFCC by the following equation (1):

$$Mel(f) = 2595 * \lg(1 + f / 700) \qquad (1)$$

Where the f means frequency, the unit is Hertz. The following figure shows the 24-dimensional triangular filters. MFCC feature extraction steps are as below:

- Hamming window is multiplied to each frame to smooth the edges of the signal, reduce Jibbs effect. Then do Fast Fourier Transform (FFT) to the signal.
- A Mel-scale filterbank is imposed on the FFT of each window to obtain log-energy in every sub-band.
- The spectral coefficients obtained by FFT was filtered with a Mel-scale filterbank (a serious of triangular filter) to obtain log-energy in every sub-band, getting coefficients m1, m2, m3 ....
- A discrete cosine transform is applied to decorrelate the resulting coefficients, obtain the cepstral.

The following Fig.2 shows the 256-point FFT amplitude-frequency characteristics of 24-dimensional triangular filters used in our system(because of the Symmetry, only half of the points will be shown):

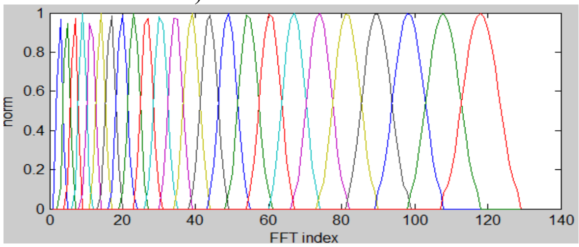

Fig.2. Mel-filter

We extract MFCC as observation sequence O, taking the Baum-Welch algorithm and Viterbi algorithm as the core algorithm for music classification [13]. HMM model is mainly to solve the training, decoding, evaluating issues. In the training stage, we use Baum-Welch algorithm which will gradient decreasingly to find the best parameters based on the observation sequence. In the decoding stage, we use Viterbi algorithm to find the best hidden state transition in observation sequence, the state combination whose output probability is the maximum, the algorithm can effectively reduce the complexity and time, eliminate noise interference. In the evaluating stage, we use forward algorithm to calcu-late the output observation sequence probability for getting the best HMM Known models. The specific process is as following Fig.3:

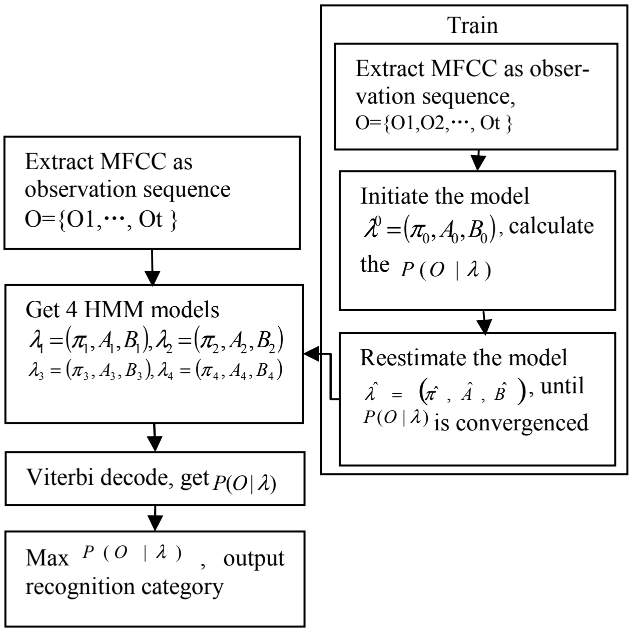
Train
Extract MFCC as observation sequence, O={O1,O2,⋯, Ot }
Initiate the model $\lambda^0=(\pi_0,A_0,B_0)$, calculate the $P(O|\lambda)$
Reestimate the model $\hat{\lambda} = (\hat{\pi}, \hat{A}, \hat{B})$, until $P(O|\lambda)$ is convergenced
Extract MFCC as observation sequence O={O1,⋯, Ot }
Get 4 HMM models
$\lambda_1=(\pi_1,A_1,B_1),\lambda_2=(\pi_2,A_2,B_2)$
$\lambda_3=(\pi_3,A_3,B_3),\lambda_4=(\pi_4,A_4,B_4)$
Viterbi decode, get $P(O|\lambda)$
Max $P(O|\lambda)$, output recognition category

Fig.3. the workflow of recognition

B. Algorithm selection

Onset detection is usually divided into three steps, namely, signal preprocessing, generating detection function and threshold selecting. Generate detection function is the key step to onset detection. Depending on the function, we have energy, phase, spectral flux, wrapping-compensation correl-ation, ICA, etc methods [1] [11].

Preprocessing is mainly to reduce noise interference and enhance the detection performance. Wavelet noise reduction, band decomposition, or filtering are the common methods to select onset. Threshold detection is classified into fixed threshold and dynamic threshold. Fixed threshold approach is fast and convenient; dynamic threshold can adapt to the changes of detection function. We use the formulate (2) below to generate dynamic threshold of the detection function(d) of nth point:

$$\widetilde{\delta} = \delta + \lambda \text{mean}\left(\left|\text{d(n - M)}\right|, \cdots \left|\text{d(n + M)}\right| \right) \quad (2)$$

Where δ， λ is positive common number and values can fine-tune. Usually λ is set as 1, and M is the points to calculate mean.

Five kinds of detection algorithm evaluated in this paper have advantages and disadvantages, performing well in their intended domains, but they can't be adaptive to all instruments. Energy method is quick and appropriate for detecting humming or percussive music whose changes in energy envelop is relatively sharp. Phase detection's advantage is to detect low-frequency notes, having less demand on the magnitude of the notes, but it is susceptible to noise effects. Spectral flux method can effectively detect increase changes in spectral domain, eliminating the interference of decreasing points. Wrapping-compensation correlation detection adds compensation coefficients into detection while computing the correlation of adjacent frames, validly reducing interference by vibrato, more suitable for pitched non-percussive instruments. Conditional ICA algorithm is an explicit probability model. Firstly the signal is decomposed into independent components. Then calculate the entropy of them, the maximum value which is relatively the most uncertain point is detected as the onset point. However it's time expensive. These methods will be introduced below.

1) Energy

The energy envelope is obtained by a first-order Gaussian filter [14] [15]. After denoising preprocessing, music signal is filtering with Gaussian filter, thus we can get their energy changes envelope, and the peak points of which are corresponding to the onsets. In time domain, music signal and filter are mutual convolution, while in frequency domain, spectrum of both signal are multiplied. After the variance is determined, the first-order Gaussian filter is obtained as follows (3):

$$h'(n) = -(n - \frac{N}{2})e^{-\frac{(n-\frac{N}{2})^2}{2\sigma^2}} \quad (3)$$

Where n is a independent variable, N is the length of filter, N/2 is the mean of Gaussian function, σ^2 is the variance of Gaussian.

2) Spectral flux

Among the wide variety of spectral flux method study, S. Dixon and Duxbury get better results. S. Dixon [2] uses L1-norm distance calculation formula, and Duxbury [1] uses the L2-norm square deviation distance calculation formula. We use the L1-norm distance formula (see below), because the literature [2] shows that it outperforms better than the L2 distance formula (4).

$$SF(k) = \sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1} H(|S(\omega, k)| - |S(\omega, k-1)|) \quad (4)$$

In the above formula, $H(X) = (X + |X|)/2$ is a half-wave rectifier function. If x is negative, the result is 0, which makes the result will only count the increase of the distance, emphasizing on the growth part of the change, that is, onset points instead of offset points.

3) Complex phase

According to Bello[16], when the phase is in stable period, the phase difference values are equal, as shown by the following formula (5):

$$\widetilde{\varphi}_{\text{k}}(\text{n - 1}) - \widetilde{\varphi}_{\text{k}}(\text{n - 2}) = \widetilde{\varphi}_{\text{k}}(\text{n}) - \widetilde{\varphi}_{\text{k}}(\text{n - 1}) \quad (5)$$

In order to analyze amplitude and phase, we define the target signal in polar coordinates as follows (6):

$$\widetilde{S}_k(m) = \widetilde{R}_k(m)e^{j\widetilde{\phi}_k(m)} \quad (6)$$

Using the phase difference and the phase of the front frame, the target phase is obtained as shown in the following formula (7):

$$\widetilde{\phi}_{\text{k}}(\text{n}) = \arg\left(2\widetilde{\varphi}_{\text{k}}(\text{n - 1}) - \widetilde{\varphi}_{\text{k}}(\text{n - 2})\right) \quad (7)$$

The target amplitude is considered equal to the amplitude of the front frame. Shown as follows (8):

$$\widetilde{R}_k(m) = R_k(m-1) \quad (8)$$

By calculating the Euclidean distance of target vector and the current vector in complex domain, we can quantitateively describe the changes in the stability. The specific formula is as follows (9):

$$\begin{aligned} \Gamma_k(m) = \{&[\Re(\widetilde{S}_k(m)) - \Re(S_k(m))]^2 + \cdots \\ &[\Im(S_k(m)) - \Im(S_k(m))]^2\}^{1/2} \end{aligned} \quad (9)$$

When $d_{\phi_k(m)} = 0$, just considerthe magnitude of the difference; when $d_{\phi_k(m)} \neq 0$, also consider additional phase difference.

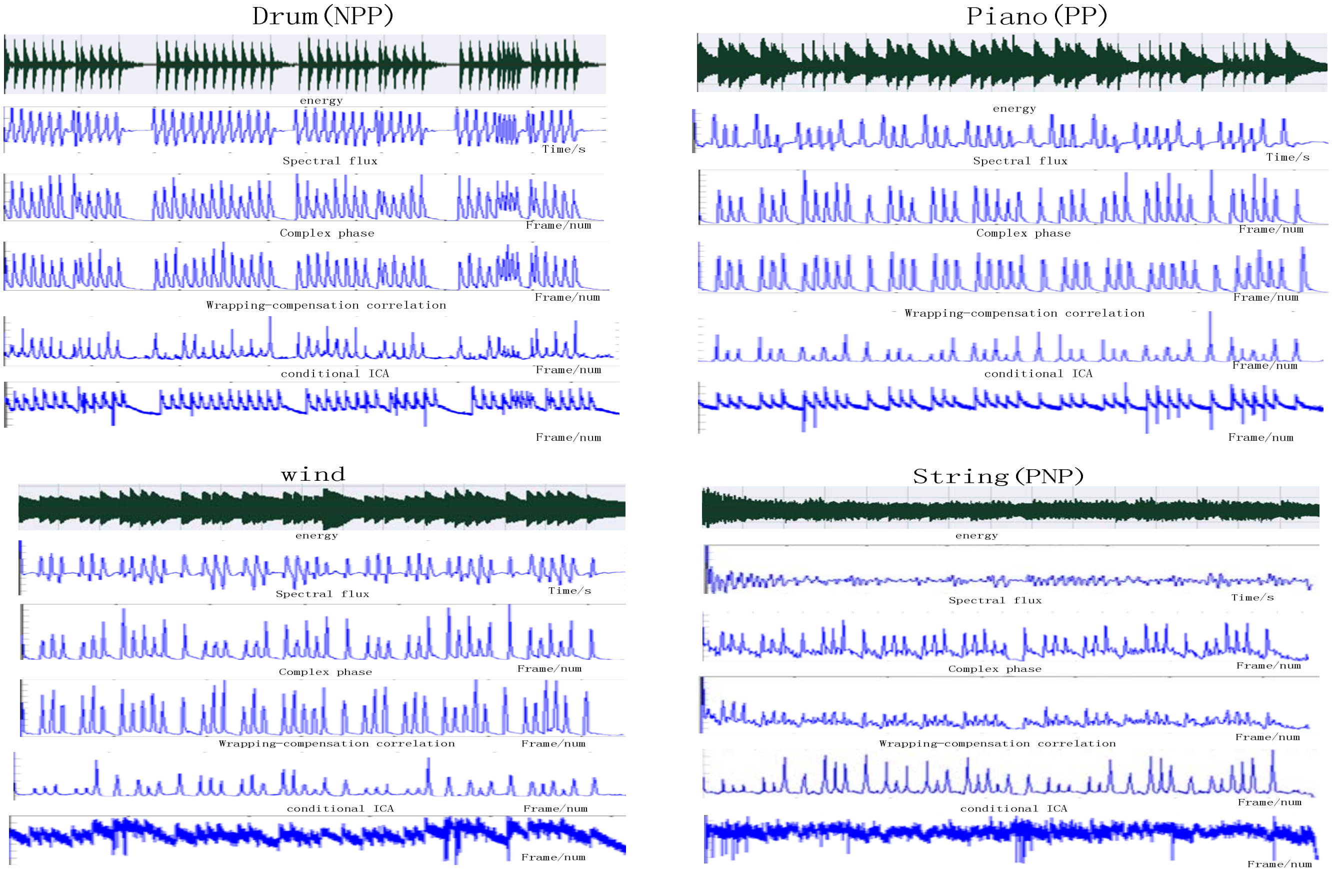

Fig.6. onset detection function

4) Wrapping-compensation correlation

Literature [4] proposed wrapping-compensation correlation detection algorithm. This algorithm analyzes only the following significant changes in the band under 8KHz and the band are divided into three sub-bands, which is used to detect the correlation value of adjacent frames. The formula below is used to generate detection function (10):

$$d(t_0)=1/\sum_{f=t_0-[w/2]}^{t_0+[w/2]}\prod_{b=1}^{b_0}\left(\left(\frac{c_b(t_0,f)}{\sqrt{c_b(t_0,t_0)c_b(f,f)}}\right)^{Wb}\right) \quad (10)$$

In above formulation, w represents the relevant frame number, which is set to 13. b0 representing the number of sub-bands, that is set to 3, t0 means the current frame, Cb is the covariance of the two frames of the same sub-band signal, Wb is the compensation coefficient value, which is set to 0.2 times of semitone, which can effectively eliminate the vibrato interference of bowed string instrument.

During the stable period, the pitch is stable, and it has a strong correlation when instrument is played. By analyzing the correlation of adjacent frame sand use compensation coefficients, this method is suitable for the detection of non pitched percussion instrument category (bowed string).

5) Conditional ICA

The probability model considered the uncertainty of onset point is highest. By computing the entropy of the signal x, the onset of the note can be detected [17] [18]. The music signal can be divided into multidimensional data, and the signal can be seen as linear synthesis of independent components by the following formula (11):

$$\mathrm{x} = As \quad (11)$$

Whereby the entropy of the signal can be represented by the following formula (12):

$$S(x)=\log(\det A)-\sum_{i=1}^{n}\log f_i(s_i) \quad (12)$$

Where n represents the number of independent components,$f_i$means the edge probability density function of the ith independent component. The experiments show that the density function can be represented approximately by the following equation (13)[17]:

$$f(s)\propto \exp-|s|^{\alpha},\alpha\approx 0.3 \quad (13)$$

Unconditional ICA system does not consider the correlation of signal, thus multidimensional signal X is then decomposed into two parts X1 and X2 in each frame, constituting a conditional ICA detection, as shown inFig.4.

Table 1 onset detection result

	Piano			Drum			wind			string		
	f-m	f-recall	f-pre	f-m	f-recall	f-pre	f-m	f-recall	f-pre	f-m	f-recall	f-pre
Eg	0.882	0.860	0.905	0.984	0.976	0.993	0.687	0.745	0.637	0.648	0.244	0.503
Sf	0.917	0.891	0.943	0.972	0.967	0.976	0.850	0.901	0.804	0.890	0.862	0.920
Cp	0.912	0.899	0.925	0.973	0.970	0.976	0.772	0.844	0.711	0.846	0.787	0.913
Wc	0.764	0.730	0.801	0.533	0.732	0.421	0.577	0.657	0.514	0.903	0.876	0.933
ICA	0.904	0.931	0.879	0.993	0.995	0.991	0.571	0.789	0.448	0.368	0.713	0.248

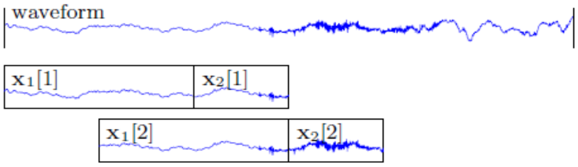

Fig.4. conditional ICA

Signals X and X1 is decomposed to obtain the independent component, and then calculate its entropy value. The entropy of the total signal is formulated as below (14):

$$S(x) = \log P(x1) - \log P(x) \qquad (14)$$

Estimating the maximum non-Gaussian value based on negative entropy can get the independent component of signal. For the signal that can't be decomposed, the detection method performs poorly.

III. EXPERIMENT AND ANALYSIS

A. *Instrument recognition test*

Training and testing dataset is produced by software Cubase, involving Piano, bowed string, wind, drum, all together 80 audios. MFCC feature is extracted for each type, and the data for each instrument category is divided into 20 parts. The first 7thparts, which is 35% of the data, is used as the training set, and the last 13th parts, 65% of the data, is used as the test set. The test result is shown as Fig.5 (abscissa means the index of test copies, ordinate means the music category number):

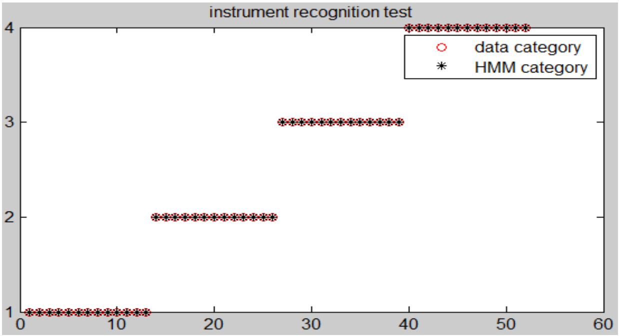

Fig.5. instrument recognition

The HMM category represents the category that data should be classified. The data category represents test results. If the two results are consistent, it means that the correct recognition. Fig.5 shows that it can be correctly classified.

B. *Onset test*

Referring to Bello [1] , we build the dataset by ourselves. Test dataset is produced by loading 4 kinds of timbre on 10 MIDI files using Cubase. It contains a total of 40 wave files (2036 onsets). Mix set consists of 15 wave files, which randomly selected from 40 waves, contain variety of music types. The correct onset times are located manually(in seconds).The error toleration window is set to 50ms, and the sampling frequency is 44100 Hz, quantization is 16 bit.

Onset researchers use the F-measure, Precision (P), Recall (R) values to evaluate the performance of onset detection [1]. We also use them as criteria. They are calculated as the following equation (15, 16, 17):

$$F\text{-measure} = 2PR/(P+R) \qquad (15)$$

$$P = TP/(TP+FP) \qquad (16)$$

$$R = TP/(TP+FN) \qquad (17)$$

In the above formulation, TP represents undetected present onsets, FP stands falsely detected un-present onsets, FN denotes undetected present onset.

Fig.6 shows the detection function of five onset algorithms on four types of musical instruments, with time-domain waveform of the instrument. Methods represent energy, spectral flux, the complex phase, wrapping-compensation correlation, conditional ICA detection from top to bottom. Abscissa means the frames or seconds, Ordinate means the detection magnitude. Musical instruments from top to bottom successively are drum(NPP), piano (PP), wind, string (PNP).We can see clearly from the figure, the detection function of energy and conditions ICA algorithms becomes flat on wind and bowed string instrument. Wrapping-compensation correlation method can generate clear waveform on string instrument, but it performs poorly on other instruments.

Table 1 shows the results of five algorithms on four types of musical instruments. Eg, Sf, Cp, Wc are short for energy, spectral flux, the complex phase, wrapping-compensation correlation detection respectively. Seen from the above test results, we know the most suitable method for drum is ICA detection. For wind and piano instrument, spectral flux is a fit approach. For bowed string instrument, wrapping-compensation correlation is appropriate. Thus, we can determine the detection method used after the instrument is recognized.

Table 2 shows the results on the mix set of single onset detection algorithm and proposed detection. In the following table, we can clearly see that the accuracy is improved after adding instrument classifier into onset detection.

Table 2 result on mix set

	f-m	f-recall	f-pre
Eg	0.754	0.696	0.821
Sf	0.909	0.896	0.922
Cp	0.874	0.858	0.890
Wc	0.694	0.748	0.648
ICA	0.680	0.842	0.571
proposed	0.929	0.910	0.948

IV. CONCLUSION

Adaptively using onset detection algorithms for music which is played by different types of instrument, the performance is superior to only a certain kind of onset detection. Due to the complexity of wind instruments, we can't produce good detected waveform. How to effectively detect such instrument is still a tough task in the future research work.

ACKNOWLEDGMENT

This research is supported by the National Natural Science Foundation of China (Grant No. 61231015), National Key Technology Research and Development Program of the Ministry of Science and Technology of China (2013BAH66F01-01), the State Administration of Radio Film and Television of China (Grant No. 2011-11) and the Engineering Planning Project of Communication University of China (Grant No. 3132014XNG 1425).

REFERENCE

[1] Juan Pablo Bello, Laurent Daudet, SamerAbdallah, Chris Duxbury, Mike Davies, and Mark B. Sandler, Senior Member. A Tutorial on Onset Detection in Music Signals. IEEE Transaction on speech and audio processing, Vol.13,No.5, 2005, pp:1035-1047.

[2] Simon Dixon. Onset detection revisited. Proc of the 9th Int. Conference on Digital Audio Effects (DAFx'06), Montreal, Canada, September 18-20, 2006

[3] Zhou R, Reiss J D. Music onset detection combining energy-based and pitch-based approaches [J]. Proc. MIREX Audio Onset Detection Contest, 2007.

[4] Schleusing O, Zhang B, Wang Y. Onset detection in pitched non-percussive music using warping-compensated correlation.Acoustics, Speech and Signal Processing,ICASSP,2008: 117-120.

[5] Xiang-Bin S, Ming L, Jie-Hong W, et al. An Improved Onset Detection Algorithm. Internet Technology and Applications, International Conference on. IEEE, 2010: 1-4.

[6] Holzapfel A, Stylianou Y, Gedik A C, et al. Three dimensions of pitched instrument onset detection [J]. Audio, Speech, and Language Processing, IEEE Transactions on, 2010, 18(6): 1517-1527.

[7] Rosão C, Ribeiro R, de Matos D M. Influence of Peak Selection Methods on Onset Detection, ISMIR. 2012: 517-522.

[8] A Eronen. Comparison of features for musical instrument recognition. Proceedings ofWorkshop on Applications of Signal Processing to Audio and Acoustics. New York : 2001.19—22.

[9] Kostek B. Musical instrument classification and duet analysis employing music information retrieval techniques. Proceedings of the IEEE, 2004, 92(4): 712-729.

[10] Deng J D, Simmermacher C, Cranefield S. A study on feature analysis for musical instrument classification. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, 2008, 38(2): 429-438

[11] Amid E, Aghdam S R. Musical Instrument Classification Using Embedded Hidden Markov Models[J].2012

[12] Vergin R, O'shaughnessy D, Farhat A. Generalized mel frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition [J]. Speech and Audio Processing, IEEE Transactions on, 1999, 7(5): 525-532.

[13] Lee J, Chun J. Musical instruments recognition using hidden markov model[C]. Signals, Systems and Computers, 2002. Conference Record of the Thirty-Sixth Asilomar Conference on. IEEE, 2002, 1: 196-199.

[14] Wang H, Yan X, Min S, et al. Musical Information Retrieval Based on Bayesian Decision.Proceedings of the 2012 Second International Conference on Electric Information and Control Engineering-Volume 02. IEEE Computer Society, 2012: 997-1001.

[15] Team Projects created for ELEC 301. ELEC 301 Projects Fall 2006. Rice University, 2006.

[16] Duxbury C, Bello J P, Davies M, et al. Complex domain onset detection for musical signals. Digital Audio Effects Workshop. 2003 : 6-9.

[17] Abdallah S A, Plumbley M D. Probability as metadata: event detection in music using ICA as a conditional density model[C]. Proc. 4th Int. Symp. Independent Component Analysis and Signal Separation (ICA2003). 2003: 233-238.

[18] Abdallah S, Plumbley M D. Unsupervised onset detection: a probabilistic approach using ICA and a hidden Markov classifier[C]. Cambridge Music Processing Colloquium. 2003.