MUSICAL ONSET DETECTION BASED ON ADAPTIVE LINEAR PREDICTION

Wan-Chi Lee and C.-C. Jay Kuo
Integrated Media Systems Center and Department of Electrical Engineering
University of Southern California, Los Angeles, CA 90089-2564
E-mails: wanchile@usc.edu, cckuo@sipi.usc.edu

ABSTRACT

A new musical onset detection technique based on adaptive linear prediction theory is proposed in this work. We decompose a music signal into multiple sub-bands, and then apply a forward linear prediction error filter (LPEF) to model the narrow-band signal in each band, respectively. To enhance the modeling performance, the coefficients of the LPEF are updated with the least-mean-squares (LMS) algorithm. Under this framework, the onset detection problem can be formulated as the peak-error location problem. Peak selection algorithms are applied to prediction errors to locate the onset time. It is shown by experimental results that the proposed algorithm outperforms several well known existing methods for onset detection.

1. INTRODUCTION

Onset detection is an important problem in musical signal analysis. It is an essential step towards many advanced music analysis tasks, including tempo analysis, beat tracking, and automatic transcription. The onset information is also useful to temporal segmentation. Since the music signal is event-based, segmenting it into individual note events greatly facilitate editing and analysis of audio.

Due to its importance, several methods have been proposed to address the onset detection problem in recent years. Scheirer [1] used the amplitude envelops of signals in several frequency bands to determine onsets. However, since his goal was to retrieve the high-level information such as beat and tempo in music, onset detection only served as a preprocessing stage in his overall system where not every onset has to be located accurately. Klapuri [2] attempted to tackle the one-by-one onset detection problem based on the psychoacoustic model. More recently, since soft onsets cannot be easily identified by the amplitude change, new methods have been proposed using the phase information as well as the energy envelope. Duxbury *et al.* [3] developed a system that takes distributions of the phase deviation and the spectral magnitude difference into account. Another algorithm proposed by Bello *et al.* [4] also tried to combine the phase and energy information for onset detection by extracting features from the complex short-time Fourier transform (STFT) domain. There are techniques originating from a different viewpoint. For example, Abdallah and Plumbley [5] used the probability model and the independent component analysis (ICA) to analyze music signals for onset detection. A tutorial on onset detection techniques can be found in [6]. Although many algorithms have been developed so far, their performance is still not satisfactory in dealing with a complex mixture of many music sounds as shown in [7], where their accuracy was low for large corpus.

A new approach for onset detection based on adaptive linear prediction is proposed in this work. Linear prediction has been widely used for the modeling and analysis of time series. It is especially popular in speech signal processing such as speech synthesis and coding. Its application to music signal modeling was reported in [8]. However, to the best of our knowledge, there is no previous work that applies linear prediction to the onset detection problem. Here, we derive a detection mechanism by passing the audio signal through an adaptive linear prediction error filter (LPEF) with the following rationale. When a signal is modeled by linear prediction, it is assumed to be stationary or quasi-stationary. However, at the note boundary, the stationary assumption fails to hold and the prediction error increases significantly. Consequently, the onset can be located by analyzing the prediction error.

The rest of the paper is organized as follows. An overview of the onset detection system is presented in Sec. 2. The linear prediction error filter and onset selection are then discussed in Sec. 3. Experimental results are conducted in Sec. 4 to demonstrate the superior performance of the proposed algorithm as compared with other existing techniques. Concluding remarks are given in Sec. 5.

2. OVERVIEW OF PROPOSED SYSTEM

Generally speaking, an onset detection system consists of two main modules. The first module processes the input audio signal and converts it to a 1-D detection function (or a time series) that exhibits peaks where the properties of the signal changes, *i.e.* where the onset happens. Its sampling rate is usually much lower as compared to the original signal. The second module then finds the peaks of the detection function and the onset time can be determined accordingly.

In deriving the detection function, it is useful to decompose a musical signal into several sub-bands and analyze the information in each sub-band separately [1,2,3,

6]. This can be explained by the fact that a broadband signal has a larger degree of freedom so that its modeling is more complicated than a narrow-band signal. Different modeling and decision techniques can be applied to signals in each sub-band. Afterwards, their outputs can be integrated to form a single decision [3].

The proposed detection system is shown in Fig. 1a. A filter bank is used at the first stage to decompose the target signal into 6 non-overlapping bands of varying bandwidths. The 1st filter covers a band with frequencies lower than note C5 (*i.e.* 523 Hz) and the bandwidth is around 500Hz. The 2nd and the 3rd bands are one-octave band-pass filters and their bandwidths are about 500Hz and 1000Hz, respectively. The next two filters, each of which covers 1000 Hz, form the 4th and the 5th bands. The 6th band contains the region with frequencies higher than note B7 (*i.e.* 3951 Hz).

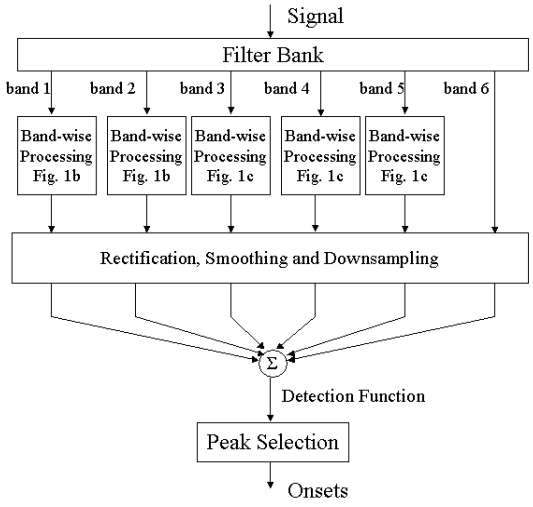


Figure 1a. System Architecture

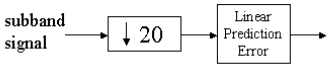


Figure 1b. Processing for bands 1 and 2

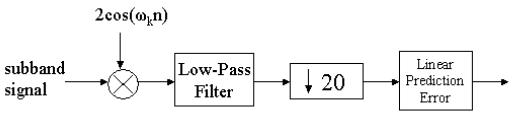


Figure 1c. Processing for bands 3, 4 and 5

Features extracted from sub-bands are shown in Figs. 1b and 1c. The first five bands are decimated (or down-sampled) by a factor of 20 to 1, which is then followed by the analysis of a linear prediction error filter (LPEF). This filter will be explained in detail in Sec. 3. The decimation before linear prediction is an important step in this system. First, it can reduce the amount of computation. Second, the decimation makes the poles of the underlying signal to spread more evenly over the entire frequency range. When poles of a signal concentrate at a small region, it is more difficult to find good linear prediction coefficients with adaptive coefficient update [9].

The input audio signal is sampled at a rate of 44KHz. The frequency ranges of the first two bands are lower than 1.1KHz so that they can be decimated directly without any aliasing effect as shown in Fig. 1b. In contrast, signals in the other 3 bands contain frequency components higher than 1.1KHz, and their cutoff frequencies are not an integer multiple of 1.1KHz. Thus, we have to pay special attention to their processing. As shown in Fig. 1c, these signals are multiplied by a sinusoidal wave so that their respective spectrum is shifted to the low-frequency range. The frequency of the sinusoidal wave depends on the lower cut-off frequency of the particular band. Furthermore, we apply a low-pass filter with a cut-off frequency of 1.1KHz before the decimation operation.

The only sub-band without linear prediction analysis is the last one of the highest frequency range. Generally, there is only little energy left in this high frequency band. The signal in this band does not have strong harmonic components (or no obvious peaks in the spectrum), which implies that the auto-regressive (AR) model, or linear prediction, is not suitable for this band. Thus, we simply adopt the amplitude envelope of the signal in this band. It was observed in [10] that percussion onsets can be well modeled as bursts of white noise, which will result in energy rise in all frequencies. This is especially obvious in the high frequency region. Then, the envelope of this band can serve as a good detection function for percussion onsets.

After getting linear prediction errors in bands 1-5 and the amplitude envelope of band 6, we determine their low-frequency envelopes through rectification, smoothing and decimation by a factor of 20 to 1 as shown in Fig. 1a. For each band, an envelope of a sampling rate of 110 Hz is obtained. We will simply add these six envelopes together to form the final detection function.

At the last stage, a peak selection algorithm is used to locate the onset time. There are several candidate algorithms available for this purpose. Here, we use the median filter dynamic threshold method, which is similar to the one used in [3] and [4]. Let D[n] denote the detection function. A data location *n* is called a peak if

$$D[n] > \delta + median(D[n-i] \text{ to } D[n+j]), \qquad (1)$$

where δ is a preset threshold and parameters *i* and *j* control the length of the sliding window. To detect a rising edge effectively, we demand $i > j$. In our experiments, *i* and *j* are chosen to be 5 and 2, respectively.

It is worthwhile to point out that two onsets in a music signal can hardly be distinguished by human being if the distance between them is less than 60ms. Thus, if two peaks are found to be within a window of 60ms, the latter one will be discarded. This can be viewed as a post-processing step applied to detected onsets.

3. LINEAR PREDICTION ERROR FILTER (LPEF)

3.1 Basic LPEF Algorithm

To model a signal using the linear prediction technique, we assume that the signal is generated by an AR process. This

actually provides a good way to synthesize musical sounds consisting of several harmonic components. Mathematically, the AR process can be written as:

$$x[n] = \sum_{k=1}^{p} a_k x[n-k] + v[n], \quad (2)$$

where p is the model order, a_k are the forward prediction coefficients and *v[n]* is a noise-like signal which is independent of *x[n]*. If coefficients a_k are known, we can calculate *v[n]* by passing *x[n]* through an FIR filter with coefficients a_k. However, these coefficients are usually unknown, and they can be estimated by minimizing the energy of *v[n]*. There are several ways to compute the prediction coefficients. Since the music signal is not stationary in the long run, the prediction coefficients should be updated with time. Here, we use an adaptive algorithm to track these coefficients whenever a new sample of *x[n]* arrives.

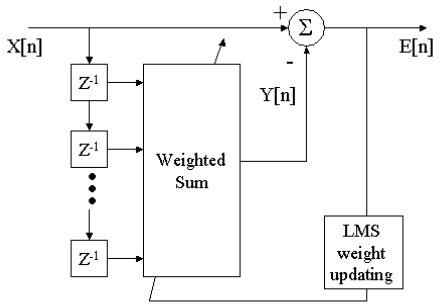


Figure 2. Linear Prediction Error Filter

Fig. 2 shows the structure of a linear prediction error filter (LPEF), where *x[n]* is the input signal and *e[n]* is the output prediction error. The LMS (Least-Mean-Squares) method is used as the adaptive algorithm for weight coefficient update. The LMS-based weight update iteration can be written as

$$\begin{aligned} y[n] &= \mathbf{w}^T[n]\mathbf{u}[n] \\ e[n] &= x[n] - y[n] \\ \mathbf{w}[n+1] &= \mathbf{w}[n] + \mu e[n]\mathbf{u}[n] \end{aligned}, \quad (3)$$

where μ is the step size, $\mathbf{u}[n] = (x[n-1], x[n-2], \ldots, x[n-p])^T$ and $\mathbf{w}[n] = (a_1, a_2, \ldots, a_p)^T$.

The prediction error *e[n]* derived by this filter provides an approximation to *v[n]* in Eq. (2). For a stationary input signal, the weight vector $\mathbf{w}[n]$ converges, and *e[n]* is close to *v[n]*. If the AR model is accurate, the energy of *e[n]* will be small. However, at the onset point where the statistical property of the input signal changes abruptly, the weight (or the linear prediction coefficients a_k) cannot be changed immediately. Then, the prediction error increases due to the poor modeling effect of the existing AR process. This leads to a peak in the energy of the filter output.

3.2 Discussion on the Design of LPEF

A. Step Size in LMS

The choice of step size μ is critical to the performance of the LPEF. Generally speaking, we want the weight to converge as fast as possible for a stationary input so that the occurrence of a non-stationary signal can be detected quickly. Then, the step size can be selected based on the normalized LMS. That is, we set

$$\mu = 1 / \mathbf{u}^T[n]\mathbf{u}[n], \quad (4)$$

where $\mathbf{u}[n]$ is given in Eq. (3). The choice in (4) makes the step size as large as possible while the convergence behavior is guaranteed. However, the large step size given in (4) will cause some problems in our application. First, if the weight vector converges too fast, the error increases due to the model mismatch at the onset point may become less obvious. Second, a large step size may make *e[n]* deviate from *v[n]*, to result in a larger value in the steady region [9]. In a nearly silent region, this effect is even worse since the magnitude of $\mathbf{u}[n]$ is extremely small. All these properties are undesirable for peak selection applied to the detection function. Thus, we want to choose a smaller step size in our current context.

We obtain a low-pass filtered version of $\mathbf{u}^T[n]\mathbf{u}[n]$, which is denoted by *r[n]*, and choose step size μ via

$$\mu = \min\left(\frac{0.1}{r[n]}, \frac{1}{\mathbf{u}^T[n]\mathbf{u}[n]}\right), \quad (5)$$

The low-pass filtering is used to prevent $\mathbf{u}^T[n]\mathbf{u}[n]$ from approaching zero, which leads to a very large step size. The factor 0.1 is determined empirically. The second term in (5) is added to guarantee the convergence of the LMS algorithm.

B. Model Order Selection

The order of the AR model, denoted by p in (2), is another important parameter. In principle, any signal can be model by an AR-model if an infinite order is adopted. However, a larger order will demand higher computational complexity. It is observed that an order number p between 25~50 works well for each band. In our experiments, we use *p=40* for the first sub-band, and *p=30* for the remaining five sub-bands.

4. EXPERIMENTAL RESULTS

To evaluate the proposed algorithm, we test it in a database containing multiple music types, including 7 types of solo musical instrument performance (piano, guitar, violin, cello, trumpet, woodwind, etc.) and 7 files of complex music mixtures. Part of our audio signals and labels are from a public database [11]. We also label other music signals using the tool provided in [11]. All audio files in our database are mono PCM sampled at 44.1KHz with 16-bit resolution per sample. All audio samples are normalized to the range [-1,1] before onset detection. The lengths of audio files are 7~30 seconds. There are 24 files in total, containing 1143 onsets. In our experiments, an automatically detected onset is viewed as "correct" if its distance to a labeled one is less than 50ms. This margin allows the inaccuracy of hand labeling. One ground-truth onset can only be matched by one detected onset, and vice versa. Thus, doubled onsets

(two detection for one true onset) and merged onsets (one detected onset for two true onsets) will be considered as errors. The numbers of files and onsets in each category are listed Table 1.

Table 1. Performance comparison of onset detection algorithms

Signal	Files	No. of Onsets	Preci-Sion(%)	Recall (%)	F-measure	F Complex
Solo Piano	4	205	98	99	98	91
Solo Guitar	2	99	84	93	88	84
Solo Violin	3	208	84	90	87	74
Solo Cello	2	120	82	78	80	67
Solo Trumpet	1	60	84	95	89	80
Solo woodwind	2	102	67	93	78	72
Solo others	3	55	58	96	72	68
Complex Mixture	7	328	72	72	72	63
Total	24	1177	79	86	83	73

Well known performance metrics for onset detection include: Precision, Recall and the F-measure. Let CD, FP and FN denote the numbers of correct detection, false positive and false negative, respectively. These three metrics can be expressed mathematically as

- Precision = CD / (CD+FP),
- Recall= CD / (CD+FN)
- F-measure =2 * Precision * Recall / (Precision + Recall)

The results are shown in Table 1. The threshold δ in Eq. (1) is set to 0.01 for all audio files. It is worthwhile to mention that the high false positive rate in solo woodwind is due to many detected offsets, which is actually good for some applications such as segmentation and transcription.

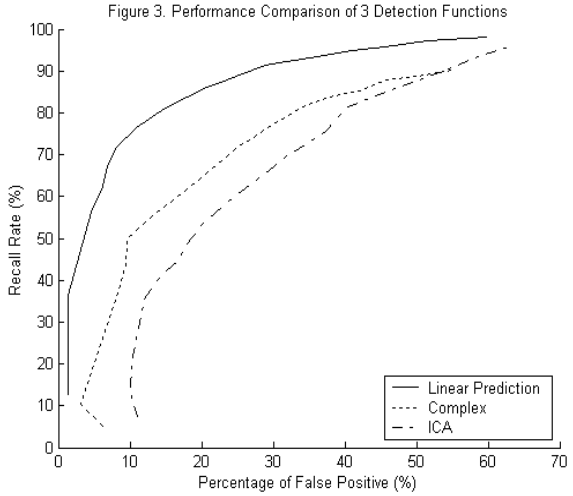
Since the test audio files used in previous work are different from ours, the performance cannot be compared directly. For the comparison purpose, we implemented two methods; namely, the complex STFT domain method [4] and the ICA method [5]. The last column in Table 1 is the F-measure value of the complex STFT domain method [4]. Since a larger F-measure implies better detection performance, we see that the proposed method outperforms the complex STFT domain method for all test audio files.

Although the three methods have different detection functions, all of them demand a peak selection algorithm in the final stage by choosing threshold δ. To compare the effectiveness of the three detection functions, we plot performance curves by varying threshold values. The recall rate versus the false positive rate for each of the three methods is plotted in Fig. 3. It is clear the proposed method gives the best results while ICA the worst.

5. CONCLUSION AND FUTURE WORK

An adaptive linear prediction error filtering method was proposed for onset detection in musical signals. It has superior performance as compared with previous methods. We will continue to improve the proposed scheme and make more tests to show the advantages of the new method.

Figure 3. Performance Comparison of 3 Detection Functions



6. REFERENCES

[1] E. D. Scheirer, "Tempo and beat analysis of acoustic musical signals," J. Acoust. Soc. Amer., vol. 103, no. 1, pp. 588–601, Jan. 1998.

[2] A. Klapuri, "Sound onset detection by applying psychoacoustic knowledge," in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP-99), Phoenix, AZ, 1999, pp. 115–118.

[3] C. Duxbury, J. P. Bello, M. Davies, and M. Sandler, "A combined phase and amplitude based approach to onset detection for audio segmentation," in Proc. 4th European Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS-03), London, U.K., Apr. 2003, pp. 275–280.

[4] J. P. Bello, C. Duxbury, M. Davies, and M. Sandler, "On the use of phase and energy for musical onset detection in the complex domain," IEEE Signal Proces. Lett., vol. 11, no. 6, pp. 553–556, Jun. 2004.

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

[6] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. Sandler, "A tutorial on onset detection in music signals," in IEEE Trans. on Speech and Audio Processing, Vol. 13, No. 5, Part 2, pp. 1035-1047, Sep. 2005.

[7] Music Information Retrieval Evaluation eXchange (MIREX) 2005 Contest for Audio Onset Detection, http://www.music-ir.org/evaluation/mirex-results/

[8] I. Kauppinen and K. Roth, "An adaptive technique for modeling audio. signals," in Proc. Conf. on Digital Audio Effects (DAFX-01), Limerick, Ireland, Dec. 2001

[9] S. Haykin, Adaptive Filter Theory, 2nd ed. Prentice-Hall 1991

[10] P. Masri, "Computer Modeling of Sound for Transformation and Synthesis of Musical Signal," Ph.D. dissertation, Univ. of Bristol, Bristol, U.K., 1996.

[11] P. Leveau, L. Daudet et G. Richard, "Methodology and Tools for the evaluation of automatic onset detection algorithms in music", in Proc. International Symposium on Music Information Retrieval (ISMIR), Barcelone, Spain, Oct. 2004.