# MUSICAL ONSET DETECTION WITH LINEAR PREDICTION AND JOINT FEATURES

**Wan-Chi Lee, Yu Shiu and C.-C. Jay Kuo**
Ming Hsieh Department of Electrical Engineering
University of Southern California, Los Angeles, CA 90089-2564
E-mails: wanchile@usc.edu, yshiu@usc.edu and cckuo@sipi.usc.edu

## ABSTRACT

Two approaches to musical onset detection by detecting the significant change of joint energy and phase features are proposed in this work. In the first algorithm, to capture the energy deviation of an audio waveform along time, we compute the error of a forward linear prediction filter as the energy feature. By further adding phase information in the second method, onsets in several music type can be detected better.

## 1 INTRODUCTION

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. We have applied adaptive linear prediction with subband analysis on music signal and showed that the forward linear prediction error can serve as a good detection function to localize the onset position in [1]. We derived 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.

In addition to the linear prediction method, we also investigated other approaches to the onset detection. To deal with onsets without obvious energy rise, we utilize the phase based based method. We have proposed a modified phase detection function that is more robust [2]. We show that the phase deviation weighted by its associated magnitude provides a better feature. In our research, we will also described the limitation of the phase feature. Due to the limitation, the phase feature cannot provide satisfying detection results while used alone. However, it can serve as an auxiliary feature and help in those cases where energy features is not sufficient.

we describe the two proposed features for onset detection, which are based on linear prediction and phase change, in depth. A complete detection system that integrates both in a sophisticated way is also described. In our system, the energy based detection is used primarily for onset detection. A evaluation procedure will be used to predict the accuracy of the feature. In many cases such as percussion sounds, the energy based feature alone will be enough for accurate onset detection and the phase detection function is not needed. The system will screen out the situations in which the phase feature are necessary. If two features are both needed, the system will perform the detection based on joint feature.

## 2 LINEAR PREDICTION ERROR FILTER(LPEF)

The most common and simplest way of prediction is the linear prediction. The value of current sample is estimated by a linear combination of several previous samples. The difference between the estimation and the true value becomes error. 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:

$$x[n] = \sum_{k=1}^{p} a_k x[n-k] + v[n], \tag{1}$$

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 ak, 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.

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

$$\begin{aligned} y[n] &= \mathbf{W}^T[n]\mathbf{u}[n] \\ e[n] &= x[n] - y[n] \\ \mathbf{w}[n+1] &= w[n] + \mu e[n]\mathbf{u}[n] \end{aligned} \tag{2}$$

where $\mu$ 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. 1. For the LEPF on a stationary input signal, the weight vector $w[n]$ 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.

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. 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. In our system, the audio signal is divided into 6 bands and linear prediction filters are applied to each.

## 3 DERIVATION OF ROBUST PHASE FEATURE

For an audio signal, $s(m)$, its STFT can be computed as

$$S(n,k) = \sum_{t=-\infty}^{\infty} s(t) w(t - nR) e^{-j2\pi tk/N}, \tag{3}$$

where $n$ is the frame index, $k = 0, 1, \ldots, N-1$ is the frequency bin index, $N$ is the DFT size, $w(t)$ is the analysis window and $R$ is the hop size (or the shift step-size of $w(t)$). We can rewrite $S(n,k)$ into the polar form as

$$S(n,k) = |S(n,k)| e^{j\tilde{\phi}(n,k)},$$

where $|S(n,k)|$ and $\tilde{\phi}(n,k)$ are the magnitude and the unwrapped phase of bin $k$.

$$\Delta\tilde{\phi}(n,k) = \tilde{\phi}(n,k) - \tilde{\phi}(n-1,k). \tag{4}$$

In practical implementation, only the wrapped phase difference

$$\Delta\phi(n,k) = princarg(\Delta\tilde{\phi}(n,k)) \tag{5}$$

can be obtained.

$$\Delta^2\phi(n,k) = princarg(\Delta\phi(n,k) - \Delta\phi(n-1,k)). \tag{6}$$

In this work, we define a new phase detection function as

$$P(n) = \frac{\sum_{k=1}^{N/2-1} |S(n,k)||\Delta^2\phi(n,k)|}{\sum_{k=1}^{N/2-1} |S(n,k)|}, \tag{7}$$

## 4 ONSET DETECTION

### 4.1 Post Processing of Feature

We observe from experiments that onsets generally occur at the rising edges, rather than the peaks, of the magnitude of the linear prediction error. Hence, we define a function $D_e(n)$ by

$$D_e(n) = \max_{m=[n,n+M1]} E(n) - \min_{m=[n-M2,n]} E(n), \tag{8}$$

where $E(n)$ is the linear prediction error. A rising edge in $E(n)$ will result in a peak in $D_e(n)$. Thus, thresholding can be applied to $D_e(n)$ for onset detection. Similarly, the same procedure is applied to phase detection function $P(n)$ and a function $D_p(n)$ indicating the rising edges in $P(n)$ is derived. Afterward onset decision are based on the two functions $D_e(n)$ and $D_p(n)$.

### 4.2 Onset Detection

So far, ways to extract phase and energy features are described. The next step is to find onsets from joint features. To decide whether the linear prediction feature is good enough or not, we develop a evaluation procedure. $D_e(n)$ will be normalized first. In the normalization, the cross-correlation of $D_e(n)$ and a pattern vector, $Ptrn$, is calculated. $Ptrn$ contains the shape of the typical peaks in the $D_e(n)$ corresponding an onset, which can be trained from onsets of isolated notes. The normalized curve $Corr(n)$ is defined as the correlation coefficients of two vector, $[D_e(n), \ldots, [D_e(n+L-1)]$ and $Ptrn$, where $L$ is the length of $Ptrn$. Since the correlation coefficients is normalized, the value of this curve is between $-1$ and 1. Generally, a peak in $D_e(n)$ result in a peak in $Corr(n)$. If the shapes of all peaks in the $D_e(n)$ are very similar to the typical pattern, the value of peaks in $Corr(n)$ will be very close to 1. This indicates peaks in $D_e(n)$ are all in very clear shape and $D_e(n)$ alone is enough for onset detection. Therefore, by the average peak value of $Corr(n)$, we can estimate the detection ability of the curve $D_e(n)$. In the practical implementation, we use a peak pattern in the length of 15 points. All local maximum larger than 0.3 in $Corr(n)$ are averaged to derived the average peak value. If average peak value is larger than 0.8, only linear prediction feature will be used. Otherwise, joint decision based on two features will be adopted.

If the result in the previous stage shows that only one feature is need in the decision, the detection by single feature is done through a threholding on $D_e(n)$. Since the normalization has been done, a fixed threshold can be used for whole detection function. A post processing will remove the detected onsets that are too close to each other. When two feature are both needed, different criteria can be used to separate onset points from others on the 2D feature space composed of $D_e(n)$ and $D_p(n)$. Supervised training can be used but we simply thresholding on the multiplication of $D_e(n)$ and $D_p(n)$ here.

## 5 REFERENCES

[1] W.-C. Lee and C.-C. Jay Kuo ”Musical Onset Detection based on Adaptive Linear Prediction”, *Proc. Internaional Conference on Multimedia and Expo (ICME06)*, Toronto, Canada, 2006.

[2] W.-C. Lee and C.-C. Jay Kuo ”Musical Onset Detection based on Joint Feature”, *Proc. Internaional Conference on Multimedia and Expo (ICME07)*, Beijing, China, 2007.