ONSET DETECTION FOR PIANO MUSIC TRANSCRIPTION BASED ON NEURAL NETWORKS

Hanqing Wen
Dept. of Electrical and Computer Engineering
University of Rochester
hwen4@ur.rochester.edu

ABSTRACT

Onset detection refers to the task of determining the physical starting time of notes or other musical events as they occur in a music recording. Various kinds of onset detection methods have been proposed in recent years. The goal of this paper is to choose a relative appropriate method to do onset detection. The neural network is discussed, especially the advanced bidirectional long short-term memory. And an important method, spectral-based method with the concept of spectral flux is introduced.

Key Words: onset detection, neural network, spectral flux, music transcription

1. INTRODUCTION

Previously, it only could be done by musicians that reading a song on sheet music and then playing it on an instrument. With machines getting smarter, they have also been designed to complete the tasks just as professionals. This kind of work is called music transcription, and piano music transcription for which music is played by piano. Thus, onset detection is a significant procedure of segmenting and transcribing music.

The structure of this paper can be mainly illustrated as mentioned above all. In this paper, it gives a brief overview of the state of the art in onset detection in Section 2, and Section 3 provides an introduction to artificial neural networks, which is combined to the proposed onset detection method of this paper. Section 4 clarifies the whole process of onset-detected research in detail. Experimental results for data sets and conclusion are provided in Section 5 and Section 6 separately.

1.1 Background and Motivations

Generally, the start of an acoustic event is marked by an onset. It is useful to apply note onset detection and localization in a number of analysis and indexing techniques for musical signals [1]. Short areas of a note containing rapid changes of the signal spectral content. Detecting onsets is not trivial, especially when analyzing complex mixtures. Applications of note onset detection systems include time stretching, audio coding and synthesis.

Locating the position of onset is an important part of music segmentation and music transcription, and thus forms the basis for multiple high level automatic retrieval tasks [2]. An onset marks the initial time of an acoustic event. In contrast to music information retrieval studies which focus on beat and tempo detection via the analysis of periodicities, an onset detector faces the challenge of detecting single events, which need not follow a periodic pattern [3]. Recent onset detection methods have matured to a level where reasonable robustness is obtained for polyphonic music [4]. However, the methods are designed specially or tuned to some certain specific kinds of onsets, such as pitched or percussive onsets, and lack the ability to perform well for music with mixed onset types.

Therefore, multiple methods need to be combined or a method has to be selected depending on the type of onsets to be analyzed [5].

1.2 Spectral Flux

Spectral flux is a measure of how quickly the power spectrum of a signal is changing, calculated by comparing the power spectrum for one frame against the power spectrum from the previous frame. And if this is restricted to the positive changes and summed across all frequency bins, it gives the onset function SF [7]:

$$\mathrm{SF(n)} = \sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1} H(|X(n,k)| - |X(n-1,k)|) \qquad (1.1)$$

More precisely, it is usually calculated as the 2-norm (also known as the Euclidean distance) between the two normalised spectra.

Calculated this way, the spectral flux is not dependent upon overall power (since the spectra are normalized), nor on phase considerations (since only the magnitudes are compared). The spectral flux can be used to determine the timbre of an audio signal, or in onset detection, among other things.

1.3 Definition and Comparison

Generally speaking, onset detection is the task of determining the physical starting time of notes or other musical events as they occur in a music recording [6]. In practice, however, the notion of an onset can be rather vague and is related to other concepts such as attack or transient. There is often a sudden increase of energy at the beginning of a musical tone, as Figure 1.1 shows.

The attack of a note refers to the phase where the sound builds up which typically goes along with a sharply increasing amplitude envelope.

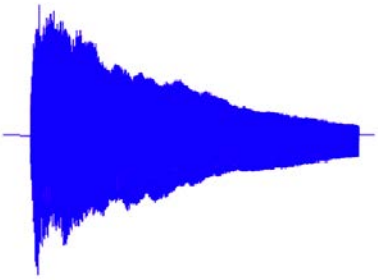

Figure 1.1. Waveform of a single note played by piano

The concept of a transient is more difficult to grasp. A transient may be described as a noise like sound component of short duration and high amplitude typically occurring at the beginning of a musical tone or a more general sound event.

However, also the release or offset of a sustained note may contain a transient-like component.

In transient regions, the signal evolves quickly in some unpredictable and rather chaotic way. Taking the piano case as an example, the transient corresponds to the initial phase where a key is hit, the damper is raised, the hammer strikes the strings, the strings start to vibrate, and the vibrations are transmitted to the large soundboard that starts to resonate and finally yields a steady and sustained sound.

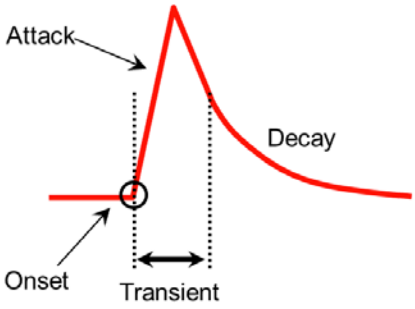

Figure 1.2. Note’s idealized amplitude envelope

Opposed to the attack and transient, the onset of a note refers to single instant (rather than a period) that marks the beginning of the transient, or the earliest time points at which the transient can be reliably detected.

Although offset estimation is essential in generating accurate transcriptions, it is likely of lesser perceptual importance than accurate onset detection. In addition, the problem of offset detection is obscured by relative energy decay and pedaling effects. In order to account for this and to reduce the influence of note duration on the performance results, it is widely used to report an evaluation of note onset detection.

2. EXISTING REDUCTION FUNCTIONS

To detect note onsets in the signal, the general idea is to capture sudden changes that often mark the beginning of transient regions. Note onset candidates could be depend on finding time positions where the signal’s amplitude envelope begins to increase when having a pronounced attack phase. It is much more challenging that the detection of onsets in the case of non-percussive music with soft onsets and blurred note transitions.

In this section, three different approaches for onset detection are reviewed: an energy-based function (Section 2.1), a spectral-based function (Section 2.2), and a phase-based function (Section 2.3).

2.1 Energy-based Function

A straightforward way to detect note onsets is to transform the signal into a local energy function that indicates for each time instance the local energy of the signal and then to look for sudden changes in this function.

The energy-based function $\Delta_{Energy} : \mathbb{Z} \to \mathbb{R}$ given by

$$\Delta_{Energy}(\mathrm{n}) := |\,\mathrm{E}_w^x(\mathrm{n}+1) - \mathrm{E}_w^x(\mathrm{n})\,|_{\geq 0} \qquad (2.1)$$

for $n \in \mathbb{Z}$, in which the local energy of x with regard to w is defined to be the function $E_w^x : \mathbb{Z} \to \mathbb{R}$ given by

$$E_w^x(\mathrm{n}) := \sum_{m=-M}^{M} |\mathrm{x}(\mathrm{n}+\mathrm{m})\,\mathrm{w}(\mathrm{m})|^2 = \sum_{m\in\mathbb{Z}} |\,\mathrm{x}(\mathrm{m})\,\mathrm{w}(\mathrm{m}-\mathrm{n})|^2 \quad (2.2)$$

2.2 Spectral-based Function

It is much harder to detect onsets when playing polyphonic music with simultaneously occurring sound events. A musical event of low intensity may be masked by an event of high intensity, and it is generally hard to detect all onsets when using purely energy-based methods. In particular, transients could be easy to be detectable in the higher frequency region since the energy of harmonic sources is concentrated more in the lower part of the spectrum.

Based on the consideration mentioned above, the idea of spectral-based novelty detection is to first convert the signal into a time-frequency representation and then to capture spectral changes in the frequency content.

As alternative of using a decibel scale, one often applies in audio processing a step also referred to as logarithmic compression, which works as follows. Let $\gamma \in \mathbb{R}_0$ be a positive constant and Γ_γ: $\mathbb{R}_0 \to \mathbb{R}_0$ a function defined by

$$\Gamma_\gamma(\nu) := \log(1+\gamma\cdot\nu) \qquad (2.3)$$

which yields a positive value $\Gamma_\gamma(\nu)$ for any positive value $\nu \in \mathbb{R}$.

To enhance weak spectral components, we apply a logarithmic compression to the spectral coefficients. To obtained the compressed spectrogram, the function Γ_γ of (2.3) to the magnitude spectrogram $|\chi|$. This yields

$$y := \Gamma_\gamma(|\chi|) = \log(1+\gamma\cdot|\chi|) \tag{2.4}$$

for a suitable constant $\gamma > 1$.

Next, it is necessary to compute the discrete temporal derivative of the compressed spectrum y. Similar to the energy-based novelty function, we only consider the positive differences (increase in intensity) and discard negative ones. This yields the spectral-based function $\Delta_{Spectral} : \mathbb{Z} \to \mathbb{R}$ defined by

$$\Delta_{Spectral}(\mathrm{n}) := \sum_{k=0}^{K} |\,\mathrm{y(n+1,k)} - \mathrm{y(n,k)}\,|_{\geq 0} \tag{2.5}$$

for $n \in \mathbb{Z}$.

The introduced spectral-based approach has more advantages than purely energy-based one. That’s because the resulting novelty function measures spectral changes, which yields much more refined information than purely energy-based approaches. And it is the most popular novelty function, so as applied in the proposed method.

2.3 Phase-based Function

Besides magnitude of the spectral coefficients, the phases of the complex coefficients are also an important source of information for various audio analysis and synthesis tasks. In particular, the phase information can be exploited for estimating an instantaneous frequency for each analysis window and frequency bin, which leads to an improved frequency resolution. In the following, it shows how the phase information can be used for onset detection.

As before, let $\chi(\mathrm{n,k}) \in \mathbb{C}$ be the complex-valued Fourier coefficient of frequency index $k \in [0:\mathrm{K}]$ and time frame $n \in \mathbb{Z}$. Using the polar coordinate representation, this complex coefficient can be written as

$$\chi(\mathrm{n,k}) = |\,\chi(\mathrm{n,k})\,|\exp(2\pi i\varphi(\mathrm{n,k})) \tag{2.6}$$

with the phase $\varphi(\mathrm{n,k}) \in [0,1)$. As the second order difference, the simultaneous disturbance of the values $\varphi''(\mathrm{n,k})$ for $k \in [0:\mathrm{K}]$ is a good indicator for note onsets. Motivated by this observation, the phase-based novelty function Δ_{Phase} defined by

$$\Delta_{Phase}(\mathrm{n}) = \sum_{k=0}^{K} |\,\varphi''(\mathrm{n,k})\,| \tag{2.7}$$

for $n \in \mathbb{Z}$.

3. NEURAL NETWORKS

In machine learning and related fields, artificial neural networks (ANNs) are computational models inspired by biological neural networks (the central nervous systems of animals, in particular the brain) and are used to estimate or approximate functions that can depend on a large number of inputs and are generally unknown. Artificial neural networks are generally presented as systems of interconnected "neurons" which can compute values from inputs, and are capable of machine learning as well as pattern recognition thanks to their adaptive nature.

For example, a neural network for handwriting recognition is defined by a set of input neurons which may be activated by the pixels of an input image. After being weighted and transformed by a function (determined by the network's designer), the activations of these neurons are then passed on to other neurons. This process is repeated until finally, an output neuron is activated. This determines which character was read.

Like other machine learning methods-systems that learn from data-neural networks have been used to solve a wide variety of tasks that are hard to solve using ordinary rule-based programming, including computer vision and speech recognition.

The neural time series is applied in MATLAB’s toolbox, which can solve a nonlinear time series problem with a dynamic neural network. Prediction is a kind of dynamic filtering, in which past values of one or more time series are used to predict future values. Dynamic neural networks, which include tapped delay lines are used for nonlinear filtering and prediction. This tool makes it possible to solve three categories of nonlinear time series problems shown below.

The first type is nonlinear autoregressive with external (exogenous) input (NARX), which simple structure figured by

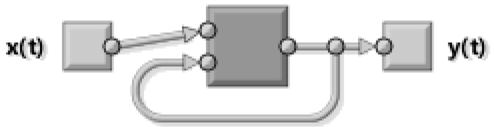

Figure 3.1. NARX structure

Predict series y (t) given d past values of y (t) and another series x (t), which defined by

$$\mathrm{y(t)} = f(\mathrm{x(t-1)},\ldots,\mathrm{x(t-d)},\mathrm{y(t-1)},\ldots,\mathrm{y(t-d)}) \tag{3.1}$$

and used for the proposed onset detector.

The second one is nonlinear autoregressive (NAR), which simple structure figured by

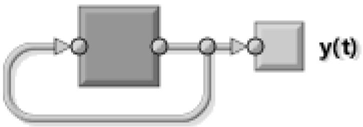

Figure 3.2. NAR structure

Its predict series y (t) given d past values of y (t), which defined by

$$y(t) = f(y(t-1),\ldots, y(t-d)) \tag{3.2}$$

The third type is nonlinear input-output, which simple structure figured by

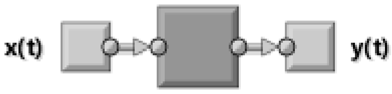

Figure 3.3. Input-output structure

The predict series y (t) given d past values of series x (t), which defined by

$$y(t) = f(x(t-1),\ldots, x(x-d)) \tag{3.3}$$

NARX solutions are more accurate than this solution. Only use this solution if past values of y (t) will not be available when deployed.

4. PROPOSED METHOD

This section describes the proposed approach for onset detection in music signals, which is based on recurrent neural networks. In contrast to previous approaches it is able to model the context an onset occurs in. The properties of an onset and the amount of relevant context are thereby learned from the data set used for training. The music data is transformed to the frequency domain. The obtained spectral flux and their corresponding ground truth values are used as inputs and targets to the recurrent neural network, which produces an onset activation function at its output. Figure 4.1 shows this basic signal flow. The individual blocks are described in more detail in the following sections.

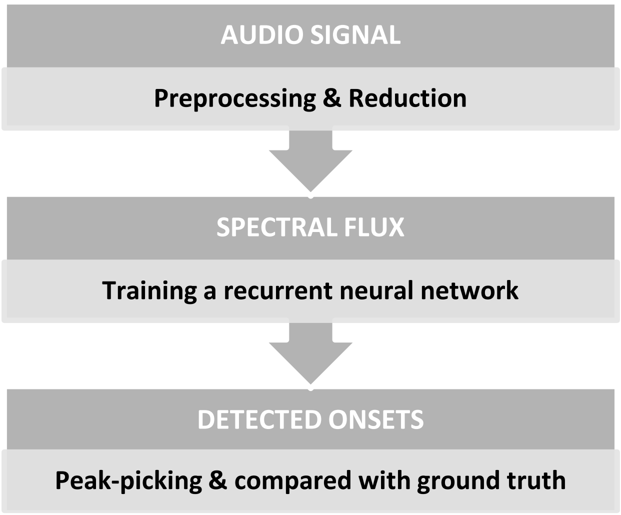
AUDIO SIGNAL
Preprocessing & Reduction
SPECTRAL FLUX
Training a recurrent neural network
DETECTED ONSETS
Peak-picking & compared with ground truth

Figure 4.1 Signal flow of onset detector

Therefore, the proposed method could be roughly illustrated as three main steps: preprocessing and reduction, NARX neural network and peak-picking.

4.1 Preprocessing and Reduction

An onset detection function is a function whose peaks are intended to coincide with the times of note onsets. Onset detection functions usually have a low sampling rate (e.g. 100Hz) compared to audio signals; thus they achieve a high level of data reduction whilst preserving the necessary information about onsets. Most onset detection functions are based on the idea of detecting changes in one or more properties of the audio signal. But audio signals, whether composed of natural or synthetic sounds, are in a continual state of change, so the task of onset detection also involves distinguishing between the various types of change, such as onsets, offsets, vibrato, amplitude modulation and noise.

If an audio signal is observed in the time-frequency plane, the onset of a new sound has noticeable energy in the frequency bands in which the sound is not masked by other simultaneous components.

Thus an increase in energy (or amplitude) within some frequency bands is a simple indicator of an onset. Alternatively, when considering that the phase of the signal in various frequency bands, it is unlikely that the frequency components of the new sound are in phase with previous sounds, so irregularities in the phase of various frequency components can also indicate the presence of an onset. Further, the phase and energy (or magnitude) can be combined in various ways to produce more complex onset detection functions.

As inputs, the raw piano pieces with a sampling rate of fs = 44.1 kHz is used. To reduce the computational complexity, stereo signals are converted to a monaural signal by averaging both channels. The discrete input audio signal is segmented into overlapping frames by a hamming window in which the window size N = 2048 (46 milliseconds at a sampling rate of 44100 hertz) and hop size h = 128 (2.9 milliseconds, or 93.75% overlap). As targets, the ground truth onsets included in every corresponding midi music files are read and then smoothed by Gaussian window.

In this paper, spectral flux is used for reduction which belongs to the spectral-based function.

4.2 NARX Neural Network

Motivated by the high performance of the onset detection method of Lacoste and Eck [8], this paper has investigated by applying an artificial neural network (ANN) based approach. Such networks were proven to work well on other audio detection tasks, such as speech recognition and enhancement.

The recurrent neural network has three hidden layers. And the first one contains eighteen neurons and the second one has fifteen. Its complete structure could be figured as below.

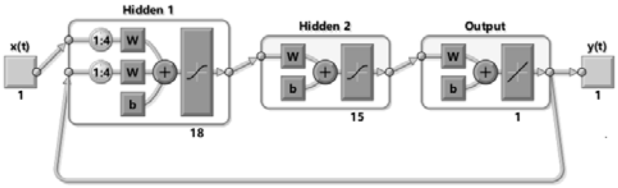

Figure 4.2 proposed neural network's structure

For network training, supervised learning with early stopping is used. Each audio sequence is presented frame by frame (in correct temporal order) to the network. Standard gradient descent with back propagation of the output errors is used to iteratively update the network weights. And the network loop type is "closed".

To prevent over-fitting, the performance (cross entropy error, cf. [9]) on a separate validation set is evaluated after each training iteration (epoch). If no improvement of this performance over 20 epochs is observed, the training is stopped and the network with the best performance on the validation set is used as the final network. The gradient descent algorithm requires the network weights to be initialized with nonzero values. Its initialization value is the weights with a random Gaussian distribution with mean 0 and standard deviation 0.1. The training data, from five piano pieces written by different composers randomly, and the test piece is selected randomly as well.

4.3 Peak-picking

The onsets are selected from the detection function by a peak-picking algorithm which finds local maxima in the detection function, subject to various constraints. The thresholds and constraints used in peak-picking have a large impact on the results, specifically on the ratio of false positives to false negatives. For example, a higher threshold generally reduces the number of false positives and increases the number of false negatives. The best values for thresholds are dependent on the application and the relative undesirability of false positives and false negatives. It is difficult to generate threshold values automatically, so the threshold is picked very carefully.

A network obtained after training as described in the previous section is able to classify each frame into two classes: "onset" and "non-onset". The standard method of choosing the output node with the highest activation to determine the frame class has not proven effective. Hence, only the output activation of the "onset" class is used.

In this paper, the threshold for "onset" probability is 0.5 and for difference between the found onsets and ground truth points is 50 milliseconds

5. RESULTS

Weights and bias are got after training the neural network, then the result of a testing piece can tell the performance of the trained network.

Figure 5.1 is a part of processed testing data compared with the ground truth. Their horizontal axis is presented by frame number. Although as a segment, it demonstrates that most onsets are found whereas some are missed which there is no corresponding onset related to the found points in upper plot.

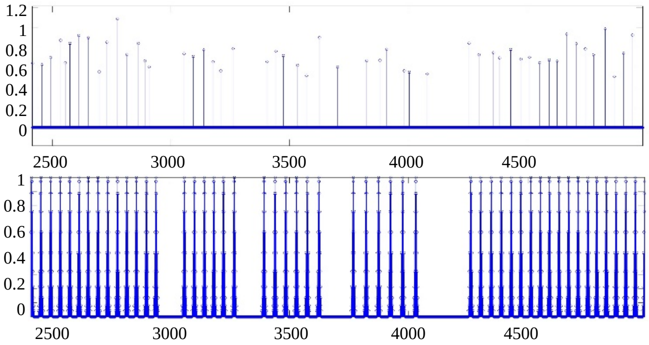

Figure 5.1 A segment of comparison in frame

It shows the found onsets at the first panel in frame, while the ground truth at the second panel.

A comparison between pre peak-picking and post peak-picking of found onsets derived from the proposed method is shown in Figure 5.2. The blue waveform are found onsets before peak-picking in time domain, while the red ones are final detected onsets after peak-picking.

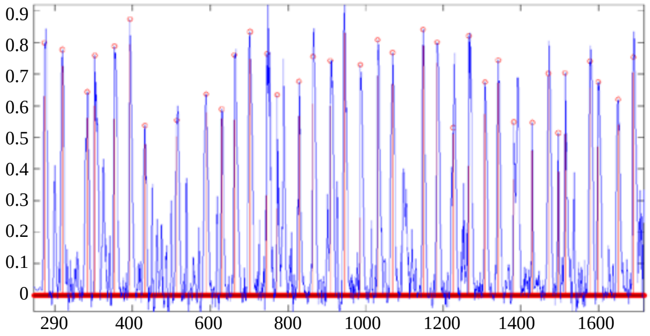

Figure 5.2 A segment of found onsets in time

In this figure, it is unavoidable that some oscillation in the ground truth midi files as well.

Five piano music pieces are chosen randomly from different composers with the corresponding midi files for ground truth in dataset . That is a relative proper number for training the network, neither more nor less. The number of ground truth onsets is 739, which removes those onsets closer than 50 milliseconds due to oscillation. After peak-picking, there is 528 onsets are true positive (TP), 163 false positive points (FP) and 211 false negative (FN). Final results are shown in Table 5.1 below.

Measure name	**Result**
Precision	**71.45%**
Recall	**76.41%**
F-measure	**73.85%**

Table 5.1 Final results of onset detection

6. CONCLUSION

The paper gives an overview of the state-of-the-art onset detection approaches for piano music transcription. And researched music onset detection using spectral-based approach combined with a recurrent neural network.

For future work, applying the bidirectional long short-term memory is a vital key to improve the neural network so that for better accuracy to detect onsets.

7. ACKNOWLEDGMENT

The author want to give her deepest gratitude first goes to Dr. Andrea Cogliati and Prof. Zhiyao Duan for fruitful discussion, help and sharing data sets, and she also wants to thank three reviewers for their precious suggestions in improving this article.

8. REFERENCES

[1] F. Eyben, B. Schuller, and G. Rigoll: "Wearable Assistance for the Ballroom-dance Hobbyist–holistic Rhythm Analysis and Dance-style Classification," *IEEE Proc. of ICME 2007*, pp.92–95, 2007.

[2] F. Gouyon, A. Klapuri, S. Dixon, M. Alonso, G. Tzanetakis, C. Uhle, and P. Cano: " An Experimental Comparison of Audio Tempo Induction Algorithms," *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 14, No. 5, pp.1832–1844, 2006.

[3] S. Dixon: "Onset Detection Revisited," *In Proc. of DAFx-06*, pp.133–137, 2006.

[4] A. Röbel: "Onset Detection By Means Of Transient Peak Classification in Harmonic Bands," *MIREX*, 2009

[5] R. Zhou and J. Reiss: "Music Onset Detection Combining Energy-Based and Pitch-Based Approaches," *MIREX*, 2007.

[6] J. Bello, L. Daudet, S. Abdallah , C. Duxbury , M. Davies and M. Sandler: "A Tutorial on Onset Detection in Music Signals," *IEEE Transactions on Speech Audio Processing*, Vol. 13, No. 5, pp.1035-1047, 2005.

[7] P. Masri: "Computer modelling of sound for transformation and synthesis of musical signal," PhD thesis, University of Bristol, 1996.

[8] A. Lacoste and D. Eck: "A Supervised Classification Algorithm for Note Onset Detection," *EURASIP Journal on Advances in Signal Processing*, 2007.

[9] A. Graves: "Supervised Sequence Labelling with Recurrent Neural Networks," PhD thesis, Technische Universität München, 2008.

[10] F. Eyben, S. Böck, B. Schuller and A. Graves: "Universal Onset Detection with Bidirectional Long Short-Term Memory Neural Networks," *ISMIR*, 2010.