November-December 2019
ISSN: 0193-4120 Page No. 6256 - 6259

Music Onset Detection Using Convolutional Neural Network

R B Kulkarni
R B Kulkarni, Associate Professor, IT Department, Government College of Engineering, Karad, Maharashtra, India.
(Email: raj.kulkarni@gcekarad.ac.in)

Article Info
Volume 81
Page Number: 6256 - 6259
Publication Issue:
November-December 2019

Article History
Article Received: *5 March 2019*
Revised: *18 May 2019*
Accepted: *24 September 2019*
Publication: *28 December 2019*

Abstract
Onset detection is a primary task in audio processing for any higher-level audio processing such as music information retrieval (MIR) or automatic speech recognition (ASR). Onset detection using data driven approach is hard due to labeled data scarcity. In this work we use some in build dataset for training our convolutional neural network (CNN) work and make some test data for Nepalese traditional music. The CNN with raw waveform of input audio signal performs well in this study for onset detection. This network performs well in diversified audio type where 50 millisecond windows are set in each audio file to identify the presence or absence of onset.

Keywords: *Onset detection, MIR, ASR, raw waveform*

I. INTRODUCTION

Onset is the time where the slope is the highest, during the attack time. *Onset detection* is the process of finding the starting points of all musically relevant events in an audio signal [1]. In visual image, edge detection highlights the sharp oriented edges in an image. Similarly, onset detection in audio are characterized by a swift change of spectral content over time. The onset detection process not directly useful but it is the first step for higher-level music analysis algorithm.

The onset detection is an old unsolved problem because of some difficulties in music. The real-world music data are complex polyphonic, and the onset can be appearing in very short time, mostly seen on violin or flute music. Other difficulties can be possible due to masking, simultaneous or quasi-simultaneous notes, unknown number of onset in a chord, and effect of noise due to low quality recording. The onset become more hard problem if some ambiguous events such as vibrato, glissando, varied dynamics, embouchure and articulation are appeared in music that continuously lower or raise the pitch of a note or chord. The onset detection is a subjective problem. According to the change in energy level in spectrogram, the onset can be characterized as hard onset and soft onset, as described in [2] and [3]. The sudden change in energy of hard onset can easily detected by energy-based algorithms with time-frequency representation. In soft onset detection, the energy changes gradually and that makes it more difficult to detect because music signals often contain noises. The onset detection algorithms are generally divided into two categories: energy-based onset detection and pitch-based onset detection. The energy-based onset detection is preferable for hard onset and pitch-based onset detection algorithms are preferred for soft onset detection.

Onset detection is old problem and many algorithms are proposed for its solution. MIREX 2012 [9] was an open challenge for onset detection. The superflux method [4] and its extensions [5] and [6] are robust algorithm in onset detection. The spectral flux is defined as the sum (or mean) of the

November-December 2019
ISSN: 0193-4120 Page No. 6256 - 6259

first order differences of the magnitude spectrogram of an audio file. Superflux algorithm with vibrato suppression in [5] is a widely used method for onset detection. The vibrato is rapid or slight variation in pitch in singing or playing some musical instruments. Onset detection with linear prediction and sinusoidal modeling in [7] and S-Transform in [8] describes another way of onset detection. In data driven approaches, recurrent neural network (RNN) based onset detection in [10] and convolutional neural network (CNN) based onset detection. in [11] and [12] shows relatively better performance compared to the other conventional methods.

In this paper our study focuses on musical onset detection using convolutional neural network with raw waveform input. The old train of data driven approach use spectrogram as input but it loss the phase information of audio. Recent trend [13] and [14] in audio and speech processing, the raw waveform is fed directly into the deep neural network (DNN) which provides a more thorough end-to-end process by completely abandoning the feature extraction step and preserve the phase information. Our experiments also show better results for onset detection using the raw waveform in very shallow CNN.

II. DATASET

In this research, we used freely available data set for training which is available at [15] and more details in [16]. The Sound Onset Labellizer software also available at [15] whose purpose is to manually put notes onsets on .wav musical files (mono). The software comes along with the small annotated database for onset detection that was part of the MIREX 2005 [17] Onset detection task. The database is proposed with two labels sets: "goodlabels.zip" contains labels that have been validated by the three annotators but may miss some that are difficult to annotate precisely, and "labelsPL.zip" contains the onsets annotated by only one annotator. This dataset includes variety of music such as instruments sounds (guitar, piano, violin, cello, saxophone, trumpet and clarinet) and music sound (classic, pop, jazz, rock and techno).

We make our own test data set for Nepalese traditional music (called Lok-Dohori song). We included one song for test by per-estimating the onset point by free available software tools and manual detection.

We compare the manual and software prediction of onset points and correct the wrong predictions. The labels for test data set used in out CNN network evaluation for unknown data.

III. METHODOLOGY

This section covers data Pre-processing and convolutional neural network architecture for onset detection and network performance.

3.1 Preparation and Audio Augmentation

The given audio file (.wav) is represented in waveform format using python programming. The large waveform representation of each audio file is segmented in 50 milliseconds (that is 2205 window size with sampling rate 44100). The data sample in dataset are also augmented by noise addition as described in [18]. To reduce the data dimension of each input data files, down sampling is done by half of total available datapoints during data reading.

3.2 Network Architecture

We propose a shallow convolutional neural network for onset detection. The network only has three convolutional layer and two fully connected layer. The pooling layer is in successive convolutional layer. The detailed overview of the proposed network is illustrated in Fig. 1.

November-December 2019
ISSN: 0193-4120 Page No. 6256 - 6259

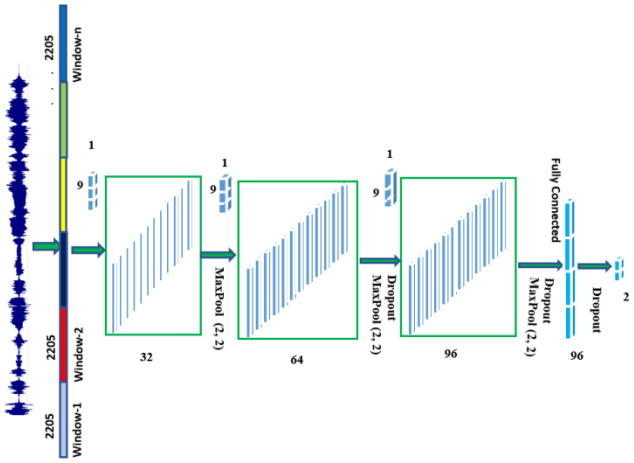

Fig. 1: Convolutional Neural Network with 1D raw audio waveform input in sliced windowed format

The dataset [15] includes 17 audio files and corresponding onset label files as a sample. We perform windowing operation in each raw waveform representation of each audio file to get the number of data samples. The size of window is very important here because if window size is too small then the probability onset in this chunk of audio is very low or no probability and if it is very large then probability of presence of onset is always one. According to the proper length of onset described in [19], we choose 50 millisecond of audio length. This setting leads the total data samples of 19164. This dataset has 57019 true onset points, that is 33.60 percentage of true data in whole dataset. The dropout also applied all the layers except the first convolutional layer. The testing dataset has 7648 true onset points, that is 7.112 percentage of true data in whole dataset. For testing case we use 50 millisecond of audio length for onset point detection..

IV. RESULTS AND DISCUSSIONS

The whole experiment was done using NVIDIA5 GeForce GTX 1080 Ti GPU. The proposed CNN with raw waveform input got the maximum training accuracy 0.9488 percentage, evaluation accuracy 0.9318 percentage and testing accuracy 0.928 percentage. The training and testing loss and accuracy curve of our network is illustrated in Fig. 2 and Fig. 3.

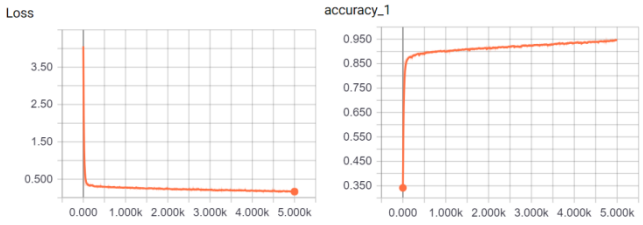

Fig. 2: Training Loss and accuracy for training dataset

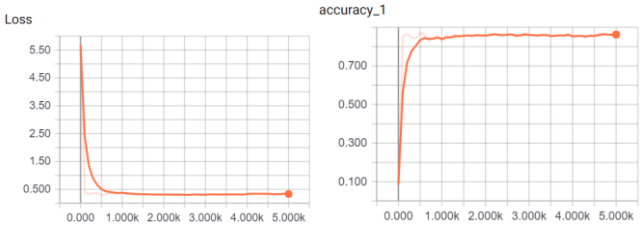

Fig. 3: Loss and accuracy for testing dataset

V. CONCLUSION AND FUTURE WORK

The onset detection using raw waveform in convolutional neural can perform well even the network is not much deep. The data can be increased by using augmentation to improve the network performance. In future it can be possible to detect the audio onset point more accurately by increasing the number of properly labeled dataset. It is possible to make hand crafted ground truth onset either observing the audio waveform and spectrogram or by using freely available onset software tools.

ACKNOWLEDGMENT

The research leading to these result, authors would like to thank Guru Technology Pvt Ltd www.nepguru.com for funding.

REFERENCES

[1] J. P. Bello, L. Daudet, and S. Abdallah,"A Tutorial on Onset Detection in Music Signals", IEEE Transactions on Speech and Audio Processing 2005, pp. 1035 - 1047.

[2] R. Zhou and J. D Reiss 'Music Onset Detection Combining Energy based and Pitch-Based Approaches", 2007, Mirex.

November-December 2019
ISSN: 0193-4120 Page No. 6256 - 6259

[3] H. L. Tan, Y. Zhu, L. Chaisorn, and S. Rahardja, "Audio Onset Detection using Energy-based and Pitch-based Processing" IEEE International Symposium on Circuits and Systems, 2010, pp. 3689 – 3692.

[4] P. Masri, Computer Modeling of Sound for Transformation and Synthesis of Musical Signals, Ph.D. thesis, University of Bristol, UK, 1996.

[5] S. Böck, and G. Widmer. "Maximum filter vibrato suppression for onset detection." 16th International Conference on Digital Audio Effects, 2013, Maynooth, Ireland.

[6] B. Stasiak and J. Mońko, "Analysis of time-frequency representations for musical onset detection with convolutional neural network", IEEE Federated Conference on Computer Science and Information Systems (FedCSIS), 2016, pp. 147 – 152

[7] J. Glover, V. Lazzarini, and J. Timoney, "Real-time detection of musical onsets with linear prediction and sinusoidal modeling," EURASIP Journal on Advances in Signal Processing, vol. 68, 2011.

[8] *N. Silva*, *C. Weeraddana*, "On Musical Onset Detection via the S-Transform" *arXiv:1712.02567*, 2017.

[9] "MIREX 2012 onset detection results", http: //nema.lis.illinois.edu/nema_out/ mirex2012/results/aod/, 2012, accessed 2013-03-27.

[10] F. Eyben, S. Böck, B. Schuller, and A. Graves, "Universal onset detection with bidirectional long short-term memory neural networks," in Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR 2010), August 2010, pp. 589–594.

[11] J. Schlüter, and S. Böck, "Improved Musical Onset Detection with Convolutional Neural Networks", 2014.

[12] B. Stasiak and J. Monko "Analysis of time-frequency representations for musical onset detection with convolutional neural network", IEEE Federated Conference on Computer Science and Information Systems (FedCSIS), 2016.

[13] W, Dai, C. Dai, S. Qu, J. Li, S. Das "Very deep convolutional neural networks for raw waveforms", Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on, pp. 421–425. IEEE, 2017.

[14] T. N Sainath, R. J Weiss, A. Senior, K. W. Wilson, and O. Vinyals, "Learning the speech front-end with raw waveform CLDNNs", Sixteenth Annual Conference of the International Speech Communication Association, 2015.

[15] "Sound Onset labellizer", http://www.tsi.telecom-paristech.fr/aao/en/2011/07/13/sound-onset-labellizer/, MIREX 2005.

[16] P. Leveau, L. Daudetm, and G. Richard, "Methodology and Tools for the Evaluation of Automatic Onset Detection Algorithms in Music", 2004.

[17] "Mirex 2005 Audio Artist", http://www.music-ir.org/mirex/wiki/2005:Audio_Artist_Identification .

[18] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," The IEEE Signal Processing Letters, 2016.