Improving Audio Onset Detection for String Instruments by Incorporating Visual Modality

Grigoris Bastas[1,3], Aggelos Gkiokas[2], Vassilis Katsouros[1], and Petros Maragos[3]

[1] Institute for Language and Speech Processing (ILSP), Athena R.C., Athens, Greece
{g.bastas, vsk}@athenarc.gr
[2] Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain
aggelos.gkiokas@upf.edu
[3] School of Electrical and Computer Engineering, National Technical University of Athens, Athens, Greece
maragos@cs.ntua.gr

Abstract. This paper presents a method for enhancing music audio onset detection in the context of live music performance recordings. As a structural element of our method we utilize a cascade of Temporal Convolutional Neural Networks (TCNs). Conventional frame based spectral representations are used as audio input features, whereas, post-processed body skeletons obtained with OpenPose constitute the visual input source. The network is trained and evaluated on monophonic string recordings from the University of Rochester Multi-Modal Music Performance (URMP) Dataset. Experimental results indicate that our model outperforms audio-based state-of-the-art methods and, additionally, that the visual component enhances detection performance.

Keywords: onset detection · audio-visual analysis · TCN.

1 Introduction

Onset detection is one of the most fundamental problems in the field of Music Information Retrieval (MIR). The state-of-the-art for audio onset detection [8] applies a Convolutional Neural Network (CNN) on spectral representations. However, music is not always experienced by humans solely through the aural modality. For instance, the produced sounds of many musical instruments correspond to certain visible movements and specific positioning of the instrument player's hands. Regarding the bowed string instruments, bowing motions are comparatively easily detectable and are strongly correlated with note onsets.

In the recent years, deep learning methods for modality fusion have gained increased interest [7]. Several innovative information extraction techniques that rely particularly on fusing audio and visual sources of music have been evolved [2], opening new areas for experimentation and further advancements. Audio-visual analysis focusing on onset detection for string ensembles has been conducted by Li et al. [3] to form a basis for score-informed audio-visual source association. Audio-visual source association has also been handled using vibrato

analysis [6]. In [4], the visual information was reduced to keypoints representing body and finger joints using OpenPose. The vibrato and bow stroke approaches have been combined permitting the generalization of the analysis on woodwind and brass instruments.

In this paper we deploy Temporal Convolutional Neural Networks (TCNs) and we demonstrate that the use of the visual modality can enhance the onset detection method. We focus on bowed string instruments, where the hand and body movement can provide cues on the beginning of the onsets.

2 Method Description

The main architecture employed in this work is a non-causal variant of the TCN model proposed in [1]. The main advantage of TCNs is that, by applying dilated 1D convolutions, they can handle temporal information by conditioning each prediction on an adequately long input, ensuring small added computational burden and large number of trainable parameters at the same time. In this configuration, the dilation factor increases from one layer l to the next by 2^l. At each layer, we apply 150 convolutional filters of size 5 and dropout with probability 0.25. One such network with 6 layers is applied on the visual source (TCN-Visual) and one with 4 layers on the audio (TCN-Audio), both followed by a linear layer with a softmax activation function predicting probabilities of occurring and non-occurring onsets. Our fusion architecture relies on concatenating the outputs of the two models and feeding them to an output network as presented in Fig. 1. Two different output networks were employed: a 4-layer TCN and a 1-layer fully connected network. The predicted onset locations were picked after computing local maxima of the activation function using centered moving maximum with a window size of 5 consecutive frames. Such values were taken under consideration provided that they exceeded a threshold of 0.5.

Our models are trained and tested on monophonic musical performance recordings drawn from the University of Rochester Multi-Modal Music Performance (URMP) Dataset [5] which also provides onset annotations. The raw audio input of 48kHz is further processed and represented in the form of mel spectrograms with 40 frequency bands, hop size of 512 samples and frame length of 2048. As for the visual modality, we chose to use OpenPose for 2D pose estimation and we kept body skeletons comprised of 11 keypoints. Lower body joints, from the knees and below, as well as keypoints corresponding to ears and eyes, were all discarded since they are often occluded and they don't add further musical information. In order to create continuous skeletons that match the audio frame rate, in certain frames, we eliminated specific keypoints that induced unnatural movements, by following the post-processing steps from [4], and we upsampled our data. In frames where certain joints were occluded or eliminated, the keypoints were recreated using linear interpolation between valid frame instances. Standard scaling per feature was applied for each separate performance. Finally, keypoint velocities and accelerations were appended to the feature vectors thus leveraging a 66-dimensional representation.

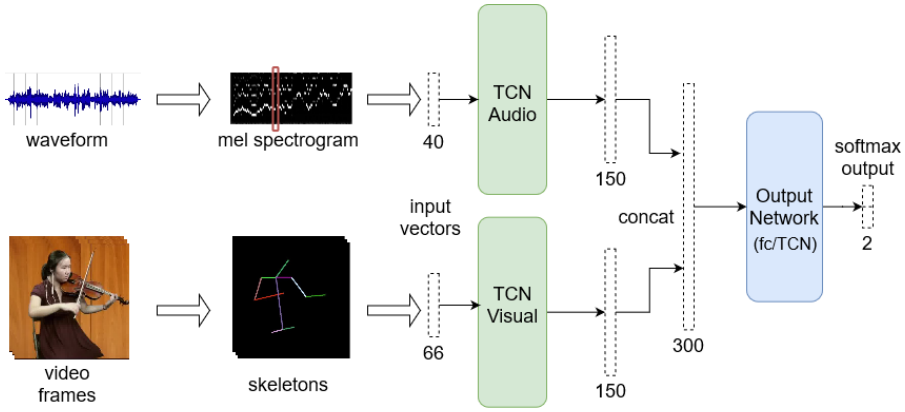

Fig. 1. Fusion model that combines the outputs of the pre-trained visual and audio sub-models by concatenating and feeding them to an output network (fc or TCN).

3 Experimental Setup

We evaluate the dataset using 8-fold cross-validation by computing the F measure for the predictions in each performance, with a tolerance window ±50 ms around the ground truth values. In the first phase of our experiments, two separate models were tested, one trained on the visual and one on the audio input, using cross-entropy loss. As presented in Table 1, the audio sub-model outperforms the state-of-the-art (CNN-Audio) on URMP dataset. Its visual counterpart naturally yields lower, yet notable results, and exhibits lower stability, as reflected by the relatively high standard deviation among different folds.

In the second phase, the pre-trained sub-models are reloaded and an additional network is fed with the their concatenated output vectors. Separate experiments were conducted in order to test four distinct fusion strategies and their potential to improve the performance in onset detection. The first strategy was based on a cascade of TCN models (TCN-Fusion), where the loaded pre-trained sub-models were rendered free to update their weights while training the whole cascade model. The same arrangement was deployed in a second experiment, this time keeping the sub-model parameters freezed (TCN-Fusion-Frzd). In the alternative layout, with the linear network used in the output instead of the TCN, as previously the sub-model parameters were at first left unfreezed (TCN-LinO-Fusion). However, freezing the pre-trained sub-models (TCN-LinO-Fusion-Frzd) proved slightly more beneficial in this arrangement.

As displayed in Table 2, TCN-Fusion achieves the most notable enhancement (+0.5%) of the onset detector among the tested fusion strategies, in terms of average scores. TCN-LinO-Fusion and TCN-LinO-Fusion-Frzd also outperform the audio sub-model, with little difference from TCN-Fusion which suffered, early on, from over-training. TCN-Fusion-Frzd was the only among the four models to exhibit no enhancement of the detection performance.

Table 1. Performance of models trained on distinct modalities with 8-fold cross-validation.

Models	F measure	
	Mean	**Std.**
TCN-Visual	0.64	5.82%
TCN-Audio	**0.921**	1.80%
CNN-Audio[8]	0.886	1.19%

Table 2. Performance of fusion models for 8-fold cross-validation.

Fusion Models	F measure	
	Mean	**Std.**
TCN-Fusion	**0.926**	1.99%
TCN-Fusion-Frzd	0.898	3.54%
TCN-LinO-Fusion	0.923	2.22%
TCN-LinO-Fusion-Frzd	0.925	1.66%

4 Conclusions

The audio-visual onset detection exhibited a non negligible improvement over the models which were trained solely on one source. This fact entails that the visual model captured information that the audio model alone couldn't.

As future work, the need to experiment with new fusion strategies is one of our priorities. The same is true about improving the performance of the visual model alone. This can have a positive impact on the fusion model. Experimenting with polyphonic performances is another possible path which could give us the opportunity to push further the limits of audio-visual onset detection analysis.

References

1. Bai, S., Kolter, J.Z., Koltun, V.: An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271 (2018)
2. Duan, Z., Essid, S., Liem, C.C., Richard, G., Sharma, G.: Audiovisual analysis of music performances: Overview of an emerging field. IEEE Signal Processing Magazine **36**(1), 63–73 (2018)
3. Li, B., Dinesh, K., Duan, Z., Sharma, G.: See and listen: Score-informed association of sound tracks to players in chamber music performance videos. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2906–2910. IEEE (2017)
4. Li, B., Dinesh, K., Xu, C., Sharma, G., Duan, Z.: Online audio-visual source association for chamber music performances. Transactions of the International Society for Music Information Retrieval **2**(1) (2019)
5. Li, B., Liu, X., Dinesh, K., Duan, Z., Sharma, G.: Creating a multitrack classical music performance dataset for multimodal music analysis: Challenges, insights, and applications. IEEE Transactions on Multimedia **21**(2), 522–535 (2018)
6. Li, B., Xu, C., Duan, Z.: Audiovisual source association for string ensembles through multi-modal vibrato analysis. Proc. Sound and Music Computing (SMC) (2017)
7. Ramachandram, D., Taylor, G.W.: Deep multimodal learning: A survey on recent advances and trends. IEEE Signal Processing Magazine **34**(6), 96–108 (2017)
8. Schlüter, J., Böck, S.: Musical onset detection with convolutional neural networks. In: 6th international workshop on machine learning and music (MML), Prague, Czech Republic (2013)