Onset detection with the user in the learning loop

Jose J. Valero-Mas, José M. Iñesta, and Carlos Pérez-Sancho

Department of Software and Computing Systems
University of Alicante
{jjvalero,inesta,cperez}@dlsi.ua.es

Abstract. An initial study in the context of interactive onset detection is presented. In cases in which high accuracy is a must, users are required to correct the results of a given initial onset detection. Considering this fact, we propose to include the user as an active part of the process so as to perform a supervised detection. In this paradigm, accuracy now relies on the user's expertise and the systems must adequate its performance to lower the human effort. Preliminar experimentation without considering user interaction shows an acceptable performance when using instance-based pattern recognition models trained as onset detectors. This fact opens a path for interactive approaches in which the user progressively adequates the model as this type of classifiers can be easily updated without the need for a retraining process.

1 Introduction

Onset detection, the process of automatically retrieving the starting points of musical events in audio streams, has been largely addressed in the Music Information Retrieval (MIR) field due to its usefulness in diverse applications as real-time accompaniment or audio compression, among others.

Although much research effort has been devoted to this task, state-of-the-art onset detectors are still far from being capable of retrieving perfect results [1]. While this fact does not generally limit their applicability, in cases in which precision is a must, human intervention is required for correcting the results. Given this fact, the user could be considered as an active part of the detection rather than a verification agent. As a consequence, the accuracy of the systems now relies on the user's expertise, being now the goal to optimize the interaction scheme and reduce the amount of human workload as much as possible [2].

In our case, we address the onset detection issue from a Pattern Recognition (PR) perspective. On that premise, the task can be seen as a supervised classification problem in which each time frame of a given analysis window must be labelled as containing or not an onset (two-class problem). Each instance is given by the time frame analysis of a music piece from which a set of audio descriptors, which shall act as the features for the PR system, are extracted.

Initial experimentation has been performed with static models not considering user interaction so as to find a proper combination of audio features and classifier for the task. Future experimentation considers the use of Incremental Machine Learning techniques so as to build classification models capable of adapting their performance each time the user performs a correction.

2 Initial experimentation

Experimentation so far has mainly focused on assessing audio descriptors with different classification schemes so as to find the most promising combination.

Audio description was obtained using the MIR.EDU [4] Vamp plug-in with the Sonic Annotator[1] tool. Out of the 14 descriptors provided, we used a subset of 10 comprising the frame-based low-level ones: Zero Crossing Rate and RMS Energy were obtained for the temporal domain description while spectral description was given by Centroid, Spread, Skewness, Kurtosis, Rolloff, Flatness, Crest and Flux. These descriptors were normalized to a maximum value of 1 and first order derivatives were computed out of them. Analysis parameters of blocksize and stepsize were fixed to 92.9 ms and 46.6 ms respectively.

Trees (J48), Neural Networks (MLP) and Nearest Neighbour (NN) techniques were considered for the classification. All of them were tested with the default Weka implementation except for NN in which a k-d Tree implementation was used for speeding-up the process and k neighbor values of 1, 3 and 5 were considered. Additionally, Approximate Nearest Neighbor search was also tested using the FLANN library [3] with the aforementioned k parametrization.

Evaluation was performed using 10-Fold cross validation scheme on a subset of the RWC dataset comprising 15 monotimbral audio pieces (5 from real piano recordings and 10 from synthesized MIDI) with a total of 11,553 onsets. F-measure, with a 50 ms tolerance time-window as in the Music Information Retriveal eXchange (MIREX)[2], is used as the evaluation measure.

Results obtained, which can be checked in Figure 1, prove PR as a way of addressing onset detection. Descriptors used are capable of describing onset events as, on average, F-measure is over a 0.5 value. In general, normalization does not carry a performance improvement, but the inclusion of the derivatives does remarkably enhance the detection process. Furthermore, instance-based classification, represented by the NN schemes, shows a competitive performance over the more complex models MLP and J48 in spite of their conceptual simplicity.

3 Conclusions

State-of-the-art onset detection algorithms are still far from retrieving perfect results, thus requiring human corrections in situations where accuracy is a must. In sight of this, users could be regarded as active elements in the detection, thus implying that results would now rely on their expertise. Incremental Machine

[1] http://www.vamp-plugins.org/sonic-annotator/

[2] http://www.music-ir.org/mirex/wiki/MIREX_HOME

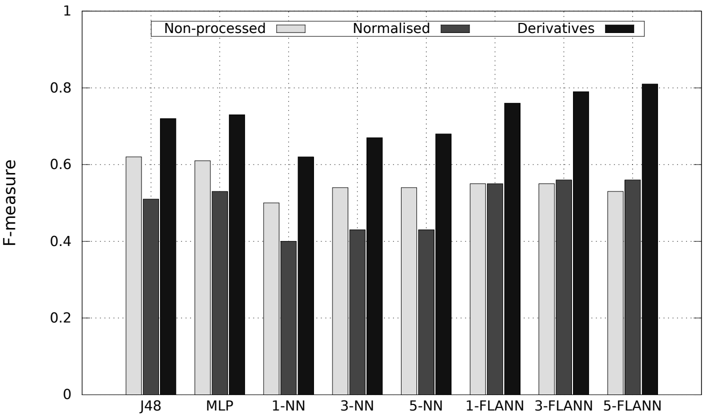

Fig. 1. F-measure results obtained for the different classification schemes proposed.

Learning techniques could be considered for, given a onset detection model, include this expert knowledge in the detection loop and reduce the user workload.

Initial results addressed in this paper prove that onset detection can be addressed as a classfication task. Low-level audio description shows an acceptable performance in the detection, especially when combined with temporal derivatives. Moreover, instance-based classifiers depict competitive accuracy results when compared to more complex models as Neural Networks, opening the path for the use of low-complexity incremental learning techniques.

Acknowledgements

This work has been funded by the Spanish Ministry project TIN2013-48152-C2-1-R and Universidad de Alicante through FPU program (UAFPU2014–5883).

References

1. Bello, J., Daudet, L., Abdallah, S., Duxbury, C., Davies, M., Sandler, M.B.: A tutorial on onset detection in music signals. Speech and Audio Processing, IEEE Transactions on (2005)
2. Iñesta, J.M., Pérez-Sancho, C.: Interactive multimodal music transcription. In: Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, BC, Canada, May 26-31 (2013)
3. Muja, M., Lowe, D.G.: Scalable nearest neighbor algorithms for high dimensional data. Pattern Analysis and Machine Intelligence, IEEE Transactions on (2014)
4. Salamon, J., Gómez, E.: Mir.edu: An open-source library for teaching sound and music description. In: Proceedings of the 15th International Society for Music Information Retrieval (ISMIR), Tapei, Taiwan (2014)