Automatic Annotation of Timbre Variation for Musical Instruments

Goffredo Haus, Luca A. Ludovico, and Giorgio Presti

Laboratorio di Informatica Musicale
Dipartimento di Informatica "Giovanni Degli Antoni"
Università degli Studi di Milano
Via Comelico, 39 - 20135 Milano (Italy)
{goffredo.haus,luca.ludovico,giorgio.presti}@unimi.it

Abstract. This paper proposes a preprocessing technique for the automatic transcription of performances produced by a musical instrument (or other sound source) capable of timbre variations. Voice recognition techniques will be exploited to gather information about timbre, then a clustering approach will be used to reduce data cardinality, and, finally, data dimensionality will be further reduced using multi-dimensional scaling to create labels as points in a data-driven timbre-space. A graphical visualization of the achieved results will be implemented in order to verify the achievement of the initial requirements. A MATLAB toolkit performing the operations described in this paper is publicly available to test the effectiveness of the proposed approach.

Keywords: automatic transcription; performance; timbre; music score

1 Introduction

The automatic transcription of a musical performance from its audio recording is a relevant goal in sound and music computing. Already addressed in a number of scientific works, from many points of view it is still considered an open research problem.

Concerning the state of the art, different approaches have been experimented. As a criterion to group reference literature, we considered the kind of information they start from, thus identifying 2 categories: *pitch-based approaches* and *timbre-based techniques*. In this context, the term *timbre* refers to the nuances and colorations of the overall spectral profile, rather than the actual harmonic structure only distinguishing different instruments. In this way, we can also distinguish among timbral modulations produced by the same sound source.

The former category mainly addresses polyphonic music and includes proposals based on pitch tracking (e.g., [8] and [9]), pitch salience and MIDI representations (e.g., [15] and [16]), and the detection of musical structures (e.g., [10]). For our purposes, a particularly relevant research is the one described in [14], aiming at the automatic transcription of piano music by first assigning a pitch-based label to each audio frame, and then applying hidden Markov models (HMMs).

The latter category focuses on timbre-based approaches. Given the aforementioned definition of timbre, we are not considering those research works based only on the distinction among stochastic and deterministic spectrum components, such as [17]. Conversely, our survey embraces works addressing timbre as coarse spectral profile. Articles compliant with this definition include the automatic transcription of drum loops based on onset detection, feature extraction and classification through HMMs and support vector machines (SVMs) [4], possibly analyzing also audiovisual features [5], and the transcription of expressive oral percussive performances with a similar approach [6].

In addition, it is worth citing review papers which analyze the limitations of current methods and identify directions for future research. Promising approaches include the combination of several processing principles and the extraction of various types of musical information (such as the key, metrical structure, and ensemble) to feed that into a model that provides context for the note detection process. Examples falling into this category are [12], [2], and [3].

These lists do not claim to be complete. Rather, the mentioned papers should be considered as the reference literature we have analyzed to formulate a novel approach, originally conceived only for monophonic and single-pitch musical instruments and afterwards extended to cover other categories of sound sources.

Our proposal is similar to timbre-based techniques, with some noticeable differences:

- Known techniques mainly use mel-frequency cepstral coefficients (MFCCs), timbral features and spectral bands. Conversely, our proposal explores the adoption of linear predictive coding (LPC) and real cepstral coefficients, since these DSP techniques do not involve perceptual aspects. In any case, thanks to the modularity of our approach, this analysis procedure can be substituted by MFCCs with a very limited impact on code;
- Common approaches in literature adopt *classifiers* such as HMMs, SVMs, and other machine learning techniques, while we will propose *hierarchical clustering* and multidimensional scaling in order to reduce data dimensionality with no prior information about classes;
- In the process of extraction of musical information we explicitly focus on timbre, where other proposals privilege key, ensemble, structure, and so on;
- The idea of using labeling in the process of automatic transcription is not completely new; for instance, it was experimented in [14]. Nevertheless, the adoption of labels based on timbre rather than pitch is quite original. Please note that we are going to talk about *labels* instead of *features* as they do not present any musical meaning.

In this context, we started developing an algorithm to simplify the task of automatic transcription. Initially, the audio track to be analyzed had to contain the sound of a single, monophonic instrument, capable of timbre variations.

The original approach could be applied both to pitched and to non-pitched musical instruments. For a non-pitched instrument (also known as unpitched, indefinite-pitch or untuned instrument), no discernible pitch can be heard. Many non-pitched instruments are part of the percussion family, such as snare drums, crash cymbals, whistles, maracas, cowbells, and triangles. Many non-pitched instruments do or can produce a sound with a recognizable fundamental frequency, thus becoming pitched. Similarly, pitched instruments can be modified or deliberately played in order to produce unpitched sounds, as for prepared piano performances. Even if the technique proposed here is timbre- rather than pitch-based, it could to be able to enhance pitch detection, in accordance with the results described in [10].

In particular, we selected the jew's harp as a monophonic instrument capable of multiple and heterogeneous timbral effects. The jew's harp is a lamellophone instrument in the category of plucked idiophones, which may consist of a flexible metal or bamboo tongue or reed attached to a frame. The tongue/reed is placed in the performer's mouth, between the lips of the player, and plucked with the finger to produce a note (see Figure 1). The sound is then amplified and modulated by the player's skull and mouth cavities.

The jew's harp – also known in English as the *jaw harp*, *juice harp*, *mouth harp*, *Ozark harp* or *trump* – is a traditional musical instrument that belongs to many different cultures and traditions, as demonstrated by the high number of names it can assume: *guimbarde* in France, *Maultrommel* in Germany, *scacciapensieri* in Italy (with many regional variants, such as *marranzanu* in Sicily, *malarruni* in Calabria, *trunfa* in Sardinia), *koukin* in Japan, *munnharpa* or *munnharpe* in Norway, *morsing* in South India, *changu* in the Sindh province of Pakistan, *xomus* or *khomus* in the Tyva Republic and the Sakha Republic in Russian Federation, and so on.

We are particularly interested in this instrument due to the wide range of timbre effects it can produce through ad-hoc playing techniques, including tremolo, vibrato, articulation by breathing and with the tongue. An early experimentation of our algorithm has been conducted on jew's harp performances, in order to benchmark the results of the algorithmic computations. The resulting achievements will be discussed in the following.

2 The Proposed Approach

From a broad perspective, the complete process bringing from sound analysis to score transcription should consist in 5 steps:

1. Pre-processing of the input to optimize the dataset;
2. Labelling of small frames according to timbral properties;
3. Classification of segments based on labels sequences and onset timings;
4. Binding of segments to gesture symbols;
5. Reconstruction of rhythmic notation.

In this work – focusing on annotation rather than transcription – we will discuss only the first two steps, which bring from the analysis of an audio signal

Fig. 1. *Khomus* players strucking the instrument between their lips.

to the recognition of a sequence of well-defined sound entities. The aggregation of sound entities into meaningful music events and their translation into instrument-specific notation would require the availability of ground-truth data and will be matter of future works, as explained in Section 3.

2.1 Input Preprocessing

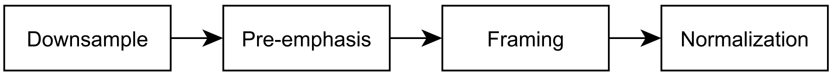

Fig. 2. Preprocessing operations.

When analyzing signal, it is a common practice to perform some preprocessing operation (see Figure 2), especially when big datasets need to be inspected and a reasonable processing time is required.

First of all, the loaded performance is converted into a monophonic signal, since stereo information brings an increase of complexity which is not necessarily bound to an increase of performance.

In [11] the authors suggested that sample rate, in a complex polyphonic context, can be drastically reduced from 44.1 kHz down to 11.025 kHz affecting music-similarity computation results by a limited amount, but allowing to significantly save computing resources. Similarly, the relevant part of the spectrum for preserving voice intelligibility is the band up to 4-5 kHz, as evidenced by the fact that in voice-recognition applications higher frequencies are usually discarded. This approach can be applied to our scenario as well, by downsampling the signal to 11.025 kHz.

At this point, the signal has been subdivided into overlapping frames that are analyzed independently as small segments of a sound event. Concerning speech recognition, the typical window size for this segmentation task is about 20 ms long, with an overlap of less than half frame. For our purposes empirical tests showed that a window size of about 25 ms with an overlap of 1/2 of the window produces good results. Nevertheless, a fine tuning of these parameters will be performed once the whole synchronization algorithm is complete.

Finally, a normalization of each frame is called for to bring all signals to the same variance scale.

2.2 Clustering Small Frames by Timbral Properties

Clustering is performed upon timbral properties, and the steps of the process are shown in Figure 3.

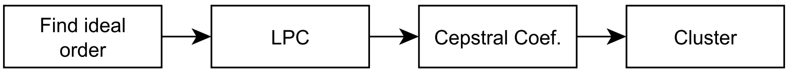

Fig. 3. Clustering operations.

Timbral properties are extracted as cepstral coefficients of the autoregressive (AR) model of each frame. Both Yule-Walker and modified covariance methods were adopted to compute AR coefficients, with no significant difference, so Yule-Walker was chosen [7]. As an optimal number of coefficients, literature suggests an order $N = 1 + 0.001 \cdot f_s$, where f_s is the sample rate. In this case, $N = 12$. To test if this rule of thumb applies to our scenario, a Minimum Description Length (MDL) test is run over each frame of a reference audio recording for the musical instrument to track. Results are illustrated in Figure 4, where gray lines show the scores of different frames, while the minimum score of each frame is highlighted by a pale blue dot. The closer to the median behavior a line is, the darker it gets.

According to MDL, most frames are best modeled with an order $N \in [6 \ldots 17]$. Figure 5 highlights the frames correctly explained by the rule of thumb of $N = 12$ through areas of different color, and those left unexplained by choosing $N = 17$. Thus, an order $N = 17$ is sufficient to model 95% of the frames.

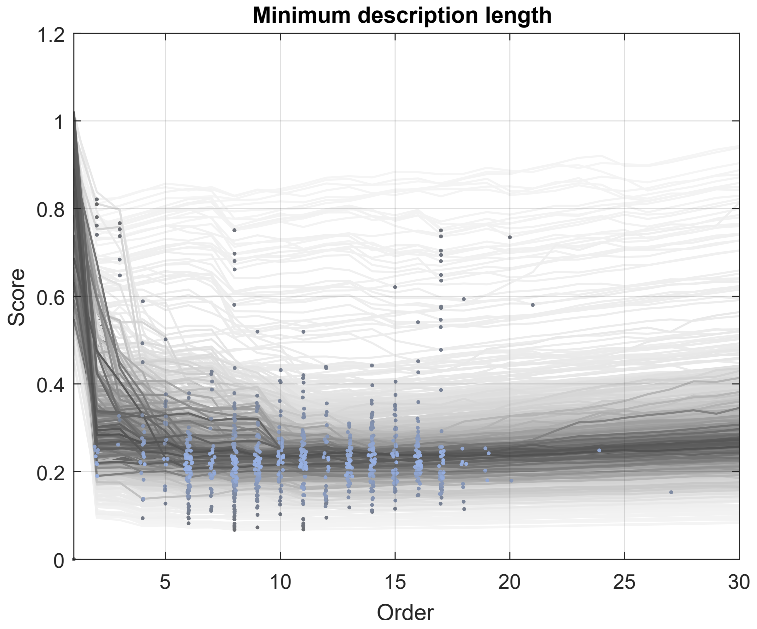

Fig. 4. Results of a Minimum Description Length (MDL) test over each frame of a reference recording.

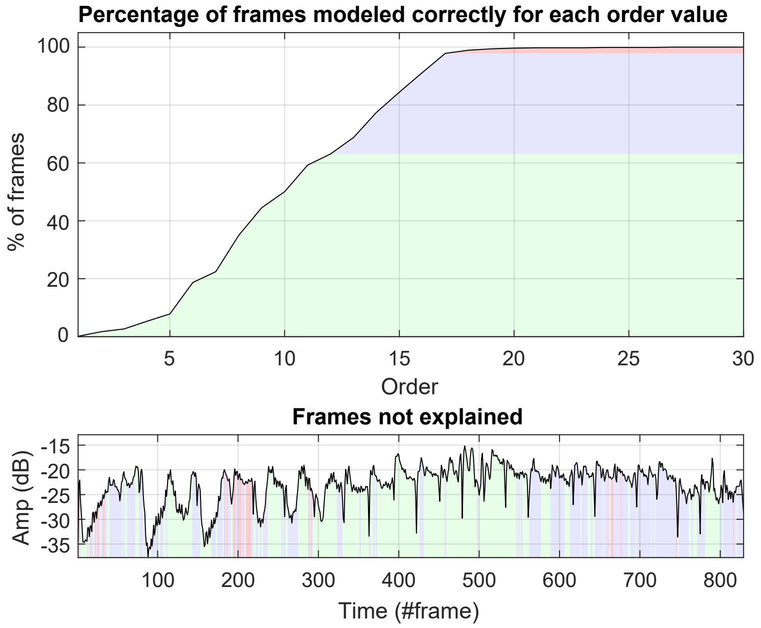

Fig. 5. Percentage of reference frames correctly modeled (above) and the corresponding signal envelope over time (below).

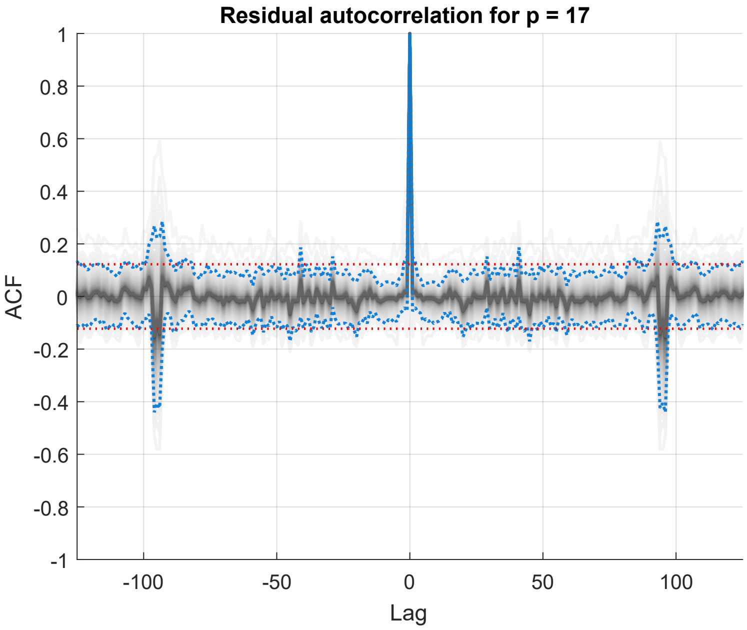

Fig. 6. Autocorrelation function of the residuals of all frames using 17^{th} order models.

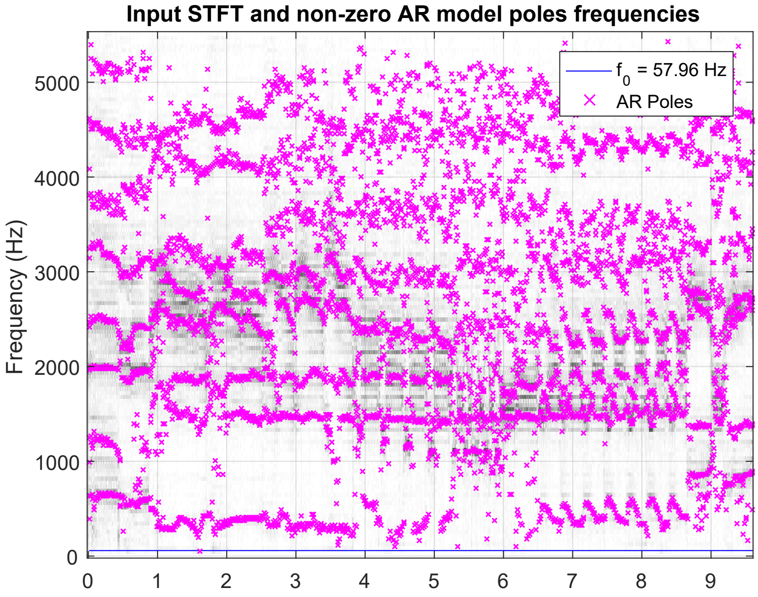

Fig. 7. AR poles against input sonogram. Frequency f_0 is highlighted as a continuous line on the bottom of the diagram: no poles are used to match f_0.

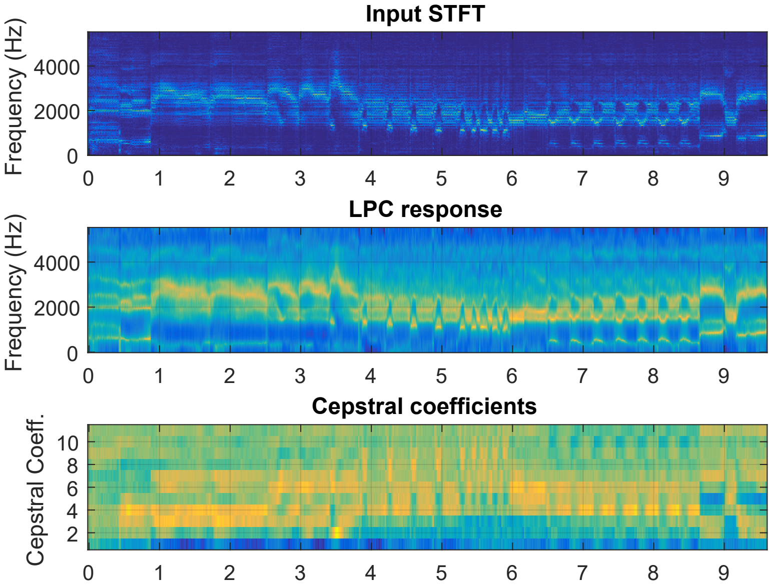

Fig. 8. The input spectrum and the extracted features: the spectral envelope based on autoregressive modeling, and cepstral coefficients respectively.

The autocorrelation function of the residuals is inspected to be sure that the selected order whitens the signal properly. Results are shown in Figure 6, where those residuals close to the median behavior are darker. Dashed lines contains 95% of the frames, whereas white Gaussian noise is contained almost entirely between the horizontal dotted red lines. Peaks near ± 100 samples lag are relative to the recording pitch.

Since Figure 6 highlights peaks in correspondence with the fundamental frequency of the khomus reed, a visual inspection is performed to be sure that no poles are wasted by tracking it. In the example adopted here and illustrated in Figure 7, the fundamental frequency is $f_0 = 57.96$ Hz, corresponding to a slightly detuned A♯.

The features extracted from LPC coefficients are: the cepstral coefficients (CCs), and the magnitude of autoregressive coefficients frequency response (ARMs). The former feature is used for the clustering task, whereas the latter for labeling purposes. In particular, CCs are limited to the first 12 elements; element 1 is discarded, too, being related to the overall energy of the frame. This step is shown in Figure 8.

Frames are finally clustered together based on CCs, using an agglomerative hierarchical clustering strategy run over each performance. The main idea is to discretize the stream of frames, or – in other words – to reduce the dimension-

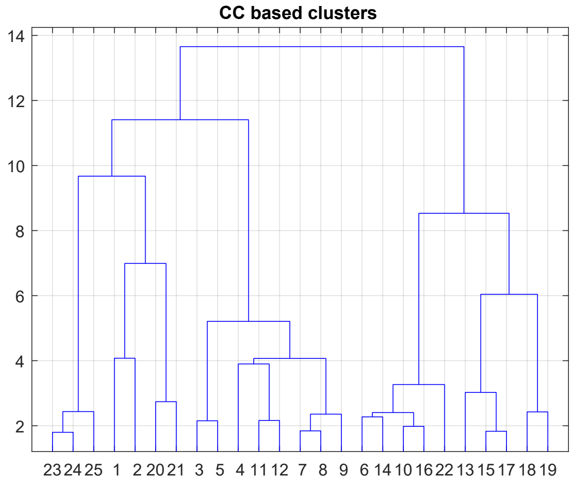

Fig. 9. Dendrogram of the clusters based on cepstral coefficients.

ality of CCs from 11 real values to 1 class, based on n-dim Euclidean distance. The number of symbols can be chosen so as to match the number of expected timbral classes. Results can be represented through a dendrogram of the cepstral coefficients computed on the reference recording. In the case of Figure 9, the tree has been cut to present 25 clusters, corresponding to the number of timbral classes expected for this example. These clusters are the input for the next steps of the transcription process.

The pairwise log-spectral distance of cluster centroids is used to run multidimensional scaling and gather 3 scores for each of them (represented in Figure 10).

The 25 resulting clusters are described by triplets of numbers that can be used as a single label. The calculation of MDS over centroids makes the 3 dimensions preserve cluster similarity in the form of Euclidean distance. Nevertheless, equal labels coming from different music pieces could imply very different timbres. In order to solve this problem, it is possible to calculate MDS over the centroids of the whole dataset instead of focusing on single pieces.

For the sake of clarity, it is worth stressing that *labels* are not *features*, i.e., they do not carry any musical information, rather they reflect timbre similarities inside the dataset on which MDS is run.

Finally, the music work described by this sequence of labels can feed a pattern-recognition system, which represents the third step of the 5-stage transcription process described above.

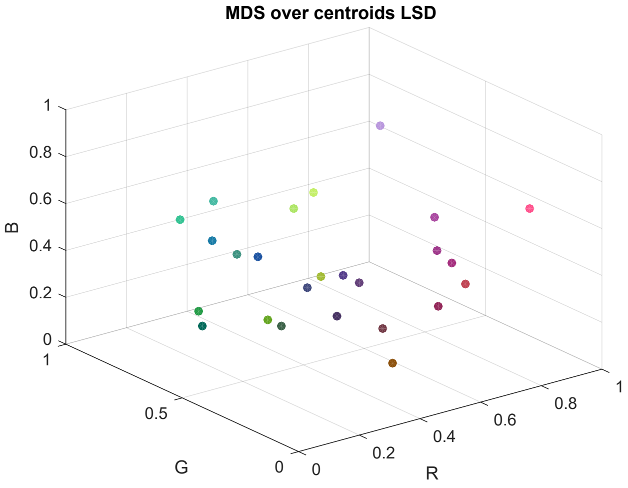

Fig. 10. Cluster centroids in the Multi-Dimensional Scaling subspace.

2.3 Visualization process

In order to test the validity of the aforesaid steps, a visualization strategy – which sub-steps are shown in Figure 11 – was implemented to evaluate performances in absence of ground-truth data.

Labels, i.e. MDS scores, were first normalized and then used as RGB values to color each frame on the base of the cluster it belongs to. Needless to say, labels could be represented through characters or any other symbolic representation instead of colors, but decoding symbols during real-time listening would be less human readable.

In this way, similar frames should have similar colors, and unique sounds should appear as similar for each instance in the performance, as intuitively shown in Figure 12.

Some videos presenting the results of the algorithm when applied to heterogeneous sound sources – including the mentioned jew's harp performance – can be found at the following URL: `http://www.lim.di.unimi.it/demo/labelSignal.php`

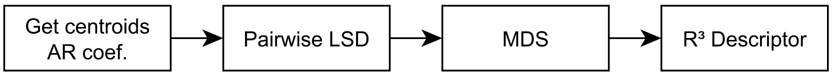

Fig. 11. Visualizing operations.

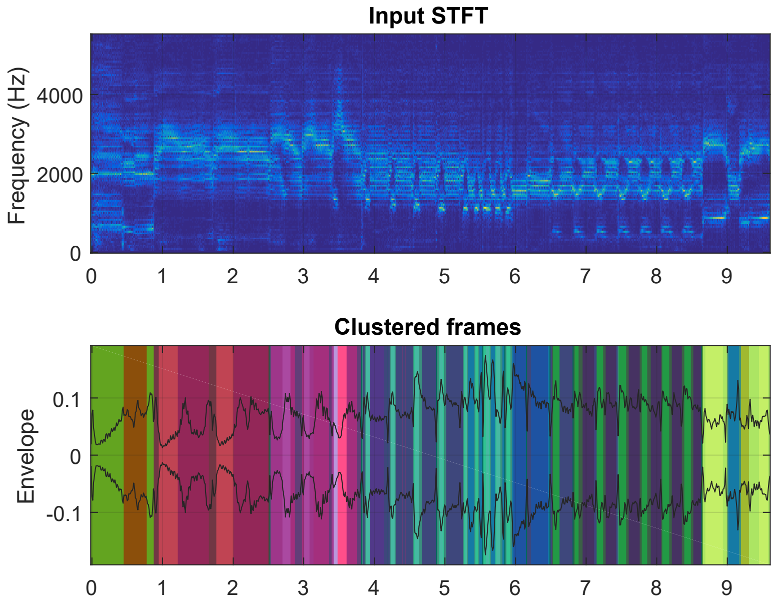

Fig. 12. Input spectrum (above) and color coded frames (below): similar frames present similar colors.

3 Future Work

Recalling the 5-step process described in Section 2, stage 3 – namely the clustering of the segments obtained so far in order to create meaningful sequences of labels and to produce onset timings – could be realized exploiting onset-detection techniques such as the ones described in [1] and [13]. A promising way to realize stages 4 and 5 is the one based on hidden Markov model (HMM) networks. This approach has been already adopted in some of the works mentioned in Section 1.

4 Conclusion

An approach to the automatic annotation of khomus performance has been presented, adopting preprocessing techniques and a qualitative evaluation strategy based on data visualization. Results seem promising so far, but a validation of the complete 5-step process and a fine tuning of the algorithm parameters will be performed when datasets accompanied by ground-truth score transcriptions will be available.

The MATLAB code implementing the algorithms described in this paper is available on GitHub at the following URL: `https://github.com/LIMUNIMI/labelSignal`. On the one side, this allows a user to replicate the tests performed by the authors, and on the other side to experiment with other sound files and parameter settings.

References

1. Barry, D., Fitzgerald, D., Coyle, E., Lawlor, B.: Drum source separation using percussive feature detection and spectral modulation. In: Proceedings of the IEE Irish Signals and Systems Conference 2005. pp. 13–17. IET (2005)
2. Benetos, E., Dixon, S., Giannoulis, D., Kirchhoff, H., Klapuri, A.: Automatic music transcription: Breaking the glass ceiling. In: ISMIR 2012. pp. 379–384. FEUP Edições (2012)
3. Benetos, E., Dixon, S., Giannoulis, D., Kirchhoff, H., Klapuri, A.: Automatic music transcription: challenges and future directions. Journal of Intelligent Information Systems 41(3), 407–434 (2013)
4. Gillet, O., Richard, G.: Automatic transcription of drum loops. In: Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on. vol. 4, pp. iv–iv. IEEE (2004)
5. Gillet, O., Richard, G.: Automatic transcription of drum sequences using audiovisual features. In: Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on. vol. 3, pp. iii–205. IEEE (2005)
6. Hazan, A.: Towards automatic transcription of expressive oral percussive performances. In: Proceedings of the 10th international conference on Intelligent user interfaces. pp. 296–298. ACM (2005)
7. Kay, S.M.: Modern spectral estimation. Pearson Education India (1988)
8. Marolt, M.: A connectionist approach to automatic transcription of polyphonic piano music. IEEE Transactions on Multimedia 6(3), 439–449 (2004)
9. Martin, K.D.: Automatic transcription of simple polyphonic music: Robust front end processing. In: Third Joint Meeting of the Acoustical Societies of America and Japan. Citeseer (1996)
10. Mauch, M., Noland, K., Dixon, S.: Using musical structure to enhance automatic chord transcription. In: ISMIR 2009. pp. 231–236 (2009)
11. Pachet, F., Aucouturier, J.J.: Improving timbre similarity: How high is the sky. Journal of negative results in speech and audio sciences 1(1), 1–13 (2004)
12. Plumbley, M.D., Abdallah, S.A., Bello, J.P., Davies, M.E., Monti, G., Sandler, M.B.: Automatic music transcription and audio source separation. Cybernetics &Systems 33(6), 603–627 (2002)
13. Presti, G., Mauro, D.A.: Continuous brightness estimation (CoBE) : implementation and its possible applications. In: Proceedings of the 10th International Symposium on Computer Music Multidisciplinary Research (CMMR), Marseille, France, October 15-18, 2013. pp. 967–974 (2013)
14. Raphael, C.: Automatic transcription of piano music. In: ISMIR (2002)
15. Ryynanen, M., Virtanen, T., Paulus, J., Klapuri, A.: Accompaniment separation and karaoke application based on automatic melody transcription. In: Multimedia and Expo, 2008 IEEE International Conference on. pp. 1417–1420. IEEE (2008)
16. Ryynänen, M.P., Klapuri, A.P.: Automatic transcription of melody, bass line, and chords in polyphonic music. Computer Music Journal 32(3), 72–86 (2008)
17. Serra, X., Smith, J.: Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition. Computer Music Journal 14(4), 12–24 (1990)