INTERACTIVE MULTIMODAL MUSIC TRANSCRIPTION

José M. Iñesta and Carlos Pérez-Sancho

University of Alicante, Spain

ABSTRACT

Automatic music transcription has usually been performed as an autonomous task and its evaluation has been made in terms of precision, recall, accuracy, etc. Nevertheless, in this work, assuming that the state of the art is far from being perfect, it is considered as an interactive one, where an expert user is assisted in its work by a transcription tool. In this context, the performance evaluation of the system turns into an assessment of how many user interactions are needed to complete the work. The strategy is that the user interactions can be used by the system to improve its performance in an adaptive way, thus minimizing the workload. Also, a multimodal approach has been implemented, in such a way that different sources of information, like onsets, beats, and meter, are used to detect notes in a musical audio excerpt. The system is focused on monotimbral polyphonic transcription.

Index Terms— Music transcription, human-computer interaction, multi-modal transcription.

1. INTRODUCTION AND PREVIOUS WORK

The goal of automatic music transcription is to extract a human readable representation, like a musical score or a piano-roll, from an audio signal containing a music performance. This process requires to perform an audio processing stage to get a piano-roll representation coding the pitches, onsets and durations of the notes, and a piano-roll to score stage if the music representation using a given notation is needed. This latter stage requires tempo, meter, and tonality estimation.

The state-of-the-art technology is far from being perfect, so, although there are good approaches (see [1]), it is unreal to think that we can have a system able to make a fully automatic music transcription, even in a restricted domain. Former works [2, 3] show that the success rate expected for monotimbral polyphonic music transcription is not particularly satisfactory due to the complexity of this task. Therefore, human post-processing is still necessary to correct the results obtained from unattended transcription systems. This is why some authors claim to focus on collaborative human-computer approaches to solve multimedia processing and recognition tasks [4, 5].

In most of the former studies, music transcription system performances have usually been evaluated in terms of precision, recall, accuracy, F-measure, and other pattern recognition and information retrieval measurements [3] (the Music Information Retrieval Evaluation eXchange, MIREX[1], is a good example of this). Nevertheless, in an interactive system like that presented in this paper, the performance evaluation turns into an assessment of how many user interactions are needed to complete the work.

The proposed strategy makes use of the time-dependency of the output, in such a way that a user interaction at a given point (time) is used by the system in two ways: validates the left hand side of the transcription, relative to that point, and improves the output at the right hand side by recomputing it in an adaptive way, making use of the user's feedback. The aim is to minimize this way the number of user interactions to complete the work.

Another issue in the proposed approach is how to use different sources of information that can be derived from the audio signal to improve the output in a multimodal scheme. We will consider informations like onsets and beat estimations to complement the multiple pitch estimation engine.

Due to the vast of possible transcription scenarios it is very important to impose a number of constraints to the problem to be solved. Therefore, we are going to focus in the monotimbral polyphonic problem, so one or more notes can sound at a given time, but emitted by a single instrument. Should more than one instrument sound at the same time the system would perform as well, but without timbre separation.

The next section brings a system overview, including descriptions of its core transcription engine and the auxiliary modules under a multimodal approach. Next, its current interaction capabilities are discussed, and finally, some results, conclusions and further development lines are presented.

2. SYSTEM ENGINES AND MULTIMODALITY

The proposed system can work on a multimodal basis, combining the informations provided by a number of engines, including multiple f_0 and tracking, onset detector, tempo es-

This work was supported by the project DRIMS (TIN2009-14247-C02), the Consolider Ingenio 2010 research programme (MIPRCV, CSD2007-00018), and the PASCAL2 Network of Excellence, IST-2007-216886.

[1] See http://www.music-ir.org/mirex

timator and beat tracker. Now we will briefly introduce the tools utilized.

2.1. Multiple fundamental frequency estimation

A multiple fundamental frequency (f_0) estimation method is the core component of a music transcription system. It infers the pitches of the harmonic sounds from the input signal and how they evolve along time. This is a challenging task that has been studied in the literature using different strategies both in the time [6] and the frequency [7] domains.

The f_0 estimator used here is based on a former work [8, 9]. It analyzes the spectrum adopting hypotheses that restrict the problem domain, addressing a feasible solution to it. These hypotheses are based on the expectation that most musical sounds (apart from percussion instruments) have a harmonic spectrum and their spectral envelopes tend to vary smoothly as a function of frequency [10]. We can use this property to separate notes contributing to the analyzed spectrum, maximizing the probabilities of smooth spectral envelopes with high harmonic amplitudes.

The output at each audio frame t consists of a set of pitch combinations $\mathcal{C}_k(t)$ with associated scores, $\mathcal{S}(\mathcal{C}_k(t))$, derived from their harmonic amplitude and spectral smoothness. The set of notes selected at t is

$$\arg\max_k \{\ \mathcal{S}(\mathcal{C}_k(t))\ \} \stackrel{\text{def}}{=} \mathcal{C}_1(t)\ .$$

In order to get a more stable detection, a short context is considered, K, where the scores of combinations with the same pitches for the $2K+1$ neighboring frames are added, obtaining a new score $\tilde{\mathcal{S}}$ for each combination:

$$\tilde{\mathcal{S}}(\mathcal{C}_k(t)) = \sum_{j=t-K}^{t+K} \mathcal{S}(\mathcal{C}_k(j))$$

and eventually selecting at t the combination with highest $\tilde{\mathcal{S}}$. These scores are used by the system to suggest other possibilities when an error is detected by the user. In this situation, when the system is asked for a new combination, all the combinations for frame t are sorted in terms of their scores, limited by a maximum amount, M: $\tilde{\mathcal{S}}(\mathcal{C}_1(t)) > \tilde{\mathcal{S}}(\mathcal{C}_2(t)) > ... > \tilde{\mathcal{S}}(\mathcal{C}_M(t))$, providing a catalogue of likely sets of notes where the user can pick the right one from.

2.2. Note onset detection

Onset detectors can infer note onsets from the harmonic content of the audio signal, but they can also be directly estimated from the input signal using strategies independent from the former [11], which typically tries to find strong energy variations. Onsets provide a temporal segmentation of the audio that can help the transcription in such a way that new pitches can only appear at onset times.

In this particular case, the method described in [12] is used to compute the onset times o_i in the audio signal. Energy variations between frames are detected using a one semitone band-pass filterbank in the frequency domain. The onsets are detected (see [12]) through local maxima of the onset energy function $O(t)$ when it is over a given threshold θ. This parameter will be used to adapt the algorithm to the characteristics of each particular signal using the user's feedback.

2.3. Beat detection

Beats can be defined as a sense of equally spaced temporal units related to perceived rhythm [13]. Like onsets, they are another source of information for transcription, and their positions can be detected from the audio, providing also the tempo of the piece and musical meaning to the note durations.

For beat tracking and tempo estimation the BeatRoot [14] algorithm was used. This tool tries to detect the implicit (or audible) tempo of the piece by inferring a list of more or less steady pulse times, that in the ideal case will correspond to beats of the music piece time signature.

2.4. Multimodal transcription

This type of transcription combines the frame-based multiple f_0 estimation with other information sources, like onsets and beats, to improve the results. The main idea is to segment the audio signal in terms of pitched energy variation local maxima (provided as onset times) or music pulses (related to the estimated beats). After sound segmentation, note sets are merged in wider, more meaningful, segments.

2.4.1. Transcription based on onsets

The onset times partition the signal in a number of sound segments between consecutive onsets. New notes can not appear within these time segments, so the combination $\mathcal{C}(t_i)$ is the same for all the frames t in the segment $o_i o_{i+1}$. The combinations selected by the multiple f_0 estimation method for all the frames in each segment are merged, combining their scores, yielding a stable set of pitches along the segment.

2.4.2. Transcription based on pulses

Similarly to onsets, the transcription module can use this pulse information and combine it with the raw f_0 estimation data to compute note candidates within each region. The beats are also used to predict and track the tempo.

The time gap between two pulses can be considered as a segment for merging pitch sets detected within that region. In this case, a divisor number q (for *quantization*) can also be used to consider pulse fractions or multiples (a quarter note, a half note, etc.). Therefore, although the beats are detected, the actual segments where the note combination must be kept are the $1/q$ division of the beat.

One additional advantage of performing pulse-based transcription is that the note durations acquire musical meaning. Without beats, note positions and durations are conditioned by the frame duration, which depend on mathematical aspects of the short-time Fourier transform. Thus, pulse detection is required if the user aims to get a music score as the final output, otherwise only a piano roll can be eventually obtained.

3. USER INTERACTION

During the music transcription task, interaction with a human expert is needed to validate or correct the initial solution given by the system. This interaction should be done from left to right in the timeline and it should be possible to automatically propagate certain proposed changes to the rest of the transcription. Different types of interactions have been considered with onsets, pulses, and transcribed notes. This framework has been tested under a graphical user interface (GUI) developed in this project [15].

3.1. Interaction with onsets

At the first stage, onsets are detected according to a given initial threshold θ_0. When the user works with the GUI, he can hear/see the detected onsets and the sound source in order to decide whether the onsets were detected correctly. Onsets can be added (false negatives, FN), deleted (false positives, FP) or moved by the user. When an interaction is made with an onset o_i, the onset energy at that onset interaction point $O(t_i)$ is utilized to set the threshold to a new value θ_i with which, considering the left-to-right timeline validation approach, the onset detection will be re-computed for $t > t_i$. This way,

$$\theta_i = \begin{cases} O(t_i) - \varepsilon & \text{if } o_i \text{ was a FN} \\ O(t_i) + \varepsilon & \text{if } o_i \text{ was a FP} \end{cases}$$

where ε is a very small number with respect to the $O(t)$ variation range.

If the onset at t_i is moved to a new time t_i', limited by $o_{i-1} \leq t_i' \leq o_{i+1}$, it will be considered as a FN at t_i' if it is moved to the right ($t_i' > t_i$), and a FP at t_i if it is moved to the left ($t_i' < t_i$)).

If the onset computation and corrections are made before computing the transcription they are only used to adapt the onset detection dynamically to the particular nature of the source sound, until it is correct. But if they are performed once the transcription is made, an onset edition at o_i implies a recalculation of the transcription between o_{i-1} and o_{i+1}. For that, the M sets of notes $\{\mathcal{C}_k(t)\}_{k=1}^{M}$, with the highest scores $\tilde{\mathcal{S}}(\mathcal{C}_k(t))$ in $o_{i-1} \leq t < o_{i+1}$ are presented to the user in order to decide which is the best transcription for that new inter-onset interval $o_{i-1}o_{i+1}$.

This interaction at t_i is propagated to the rest of the transcription, $t \geq o_{i+1}$. The onset edition will induce a onset recalculation that may modify the segmentation, in which case the transcription will be recomputed with the new inter-onset intervals. Here, the edition outcome is taken into account in such a way that if the same set of pitches is found at $t_j \geq t_{i+1}$ before recomputation, $\mathcal{C}(t_j) = \mathcal{C}(t_i)$, it is changed by that selected by the user at the edition point $\mathcal{C}'(t_j) = \mathcal{C}_k(t_i)$, under a 'the user is always right' policy. If the user selected the most probable combination at t_i, $\mathcal{C}_1(t_i)$, the decision will be taken by the system as $\mathcal{C}'(t_j) = \mathcal{C}_1(t_j)$.

3.2. Interaction with pulses

The interaction with the pulses is much more limited, since beats can not be inserted or deleted, due to their rhythmic nature. Anyway, a different relationship between the beats computed by the engine and the user diagnosis (for example, 1 detected beat can be actually 2 pulses) can be set. Also, the time signature, initially set to the most frequent one (4/4), can be established at any bar. Pulse editions at a pulse time t_p have impact in the transcription in terms of a re-computation for $t \geq t_p$ in the new context, given the quantization constraint.

4. RESULTS

Significant improvements have been found in the transcription accuracy when the multi-modal approach was utilized. Next, a representative example of a monotimbral polyphonic sound is shown. This particular case corresponds to the initial seconds of the Van Halen's song *Jump*, taken from the original recording.

In Figure 1 a portion of a frame-based transcription analysis using $K = 2$ neighbor frames is presented. Note the presence of very short notes. Most of them are actually fragments of a note with the same pitch preceding or following them. Some of these situations could be filtered out by merging or deleting too short notes through parameters that can be controlled by the user with the GUI. There are also false positive notes that have been fired due to local energy fluctuations in the spectrogram.

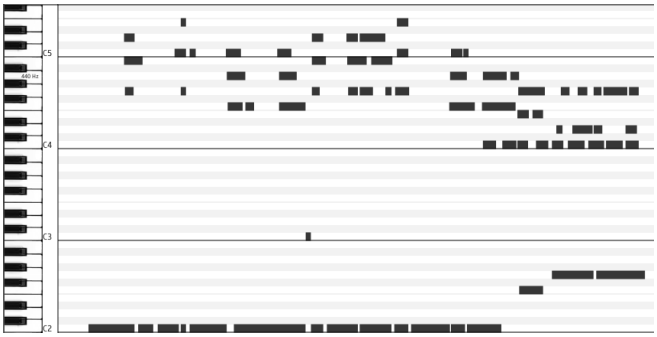

Fig. 1. Example of pitch evolution along time using transcription with frame by frame analysis.

In Figure 2 the same sound transcription example is pre-

sented, but here it is based on the detected onsets. Each vertical line marks where an onset was detected. This sequence of onsets has been validated by the user.

The number of user interactions needed to reach to a correct sequence of onsets is much lower in general compared to what the user should edit without propagation (that would be equal to the number of FP+FN). A comparison to other approaches to onset detection, like a median adaptive thresholding [16], showed no significant differences in terms of the number of interactions needed.

Note that the transcribed notes are stable and the very short notes that appeared in the frame-based transcription are removed. Also note that in this scheme silences are not allowed, since the sets of notes are maintained for all the frames in an inter-onset interval.

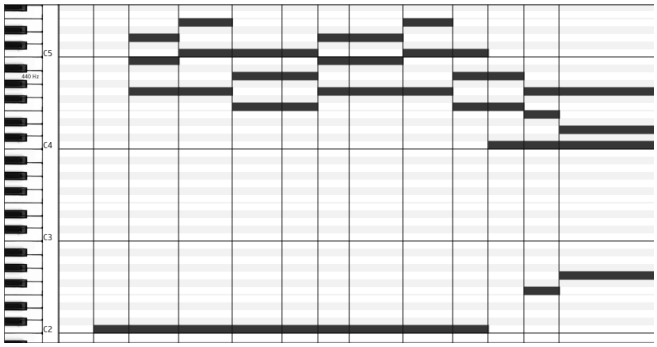

Fig. 2. Transcription of the same sound sample in Fig. 1, but assisted by onsets (vertical lines).

In Figure 3 the transcription of the same sound, now based on the detected pulses, using a quantization grid of $q = 2$, is shown. Every vertical line marks the instant where a pulse has been detected. The darker lines mark the beginning of bars whereas lighter ones mark the beats en each bar. The outcome is closer than that with onsets to the original score, but presents a higher rate of false positives, due to the fact that tempo-based segmentation does not fit perfectly with the player's performance, something that onset times do.

5. CONCLUSIONS AND FURTHER WORK

A new approach to sound to score music transcription has been presented in this paper. The inherent multimodality of the task is driven through the collaborative work of the f_0 detection and tracking engine with onset or beat computing modules. Both aspects have shown their ability to produce a more accurate and stable transcription compared to the raw frame-based output.

A more extensive evaluation is now intended, but it needs of accurately timed databases and the definition of a metric that can assess the goodness of the performance. This is not a trivial task. For example, from a qualitative point of view, the

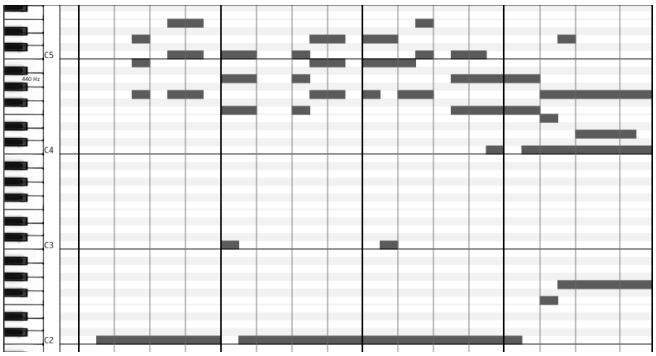

Fig. 3. Example of transcription based on pulses using $q = 2$ (quantization to a eighth note) in a 4/4 meter.

onset-based transcription is perceived as clearly better than the frame-based one when the result as a piano-roll is the objective pursued, because it is based in the physical features an timing of the signal. On the other hand, the beat-based transcription seems much better when one aims the transcribed scores, so the evaluation seems to be objective-dependent and, what is worse, subjective.

The interactive transcription GUI developed is conceived as a platform for interactive multimodal research in the context of sound transcription, but it also aims at providing a tool to help musicians, music educators, and students to transcribe a music piece with a minimum number of interactions.

6. REFERENCES

[1] Anssi Klapuri and Tuomas Virtanen, "Automatic music transcription," in *Handbook of Signal Processing in Acoustics*, David Havelock, Sonoko Kuwano, and Michael Vorlnder, Eds., pp. 277–303. Springer New York, 2009.

[2] A. de Cheveigné, "Multiple F0 estimation," in *Computationaly Auditory Scene Analysis: Principles, Algorithms and Applications*, D. Wand and G. J. Brown, Eds. Wiley-IEEE Press, 2006.

[3] M. Bay, A. F. Ehmann, and J. S. Downie, "Evaluation of multiple-f0 estimation and tracking systems," in *Proc. of the 10th International Conference on Music Information Retrieval (ISMIR)*, 2009, pp. 315–320.

[4] P. Maragos, A. Potamianos, and P. Gros, *Multimodal Processing and Interaction: Audio, Video, Text*, Springer, 2008.

[5] Alejandro H. Toselli, Enrique Vidal, and Francisco Casacuberta, *Multimodal Interactive Pattern Recognition and Applications*, Springer, 2011.

[6] J. Brown and B. Zhang, "Musical frequency tracking using the methods of conventional and 'narrowed' autocorrelation," *Journal of the Acoustical Society of America*, vol. 89, pp. 2346–2354, May 1991.

[7] Ryota Ikeuchi and Kazushi Ikeda, "An automatic music transcription based on translation of spectrum and sound path estimation," in *Neural Information Processing*, Bao-Liang Lu, Liqing Zhang, and James Kwok, Eds., vol. 7062 of *Lecture Notes in Computer Science*, pp. 532–540. Springer Berlin Heidelberg, 2011.

[8] A. Pertusa and J. M. Iñesta, "Efficient methods for joint estimation of multiple fundamental frequencies in music signals," *EURASIP Journal on Advances in Signal Processing*, vol. 2012, no. 2, pp. 27, February 2012.

[9] A. Pertusa and J. M. Iñesta, "Multiple fundamental frequency estimation using Gaussian smoothness," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Las Vegas, NV, 2008, pp. 105–108.

[10] A. Klapuri, "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness," *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 6, pp. 804–816, 2003.

[11] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler, "A tutorial on onset detection in music signals," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 5, pp. 1035–1047, September 2005.

[12] A. Pertusa, A. Klapuri, and J. M. Iñesta, "Recognition of note onsets in digital music using semitone bands," in *Proc. of the 10th Iberoamerican Congress on Pattern Recognition (CIARP 2005)*, A. Sanfeliu and M. Lazo, Eds. 2005, pp. 869–879, Springer-Verlag.

[13] S. Handel, *Listening: An Introduction to the Perception of Auditory Events*, Bradford Books / MIT Press, 1989.

[14] S. Dixon, "Evaluation of the Audio Beat Tracking System BeatRoot," *Journal of New Music Research*, vol. 36, pp. 39–50, 2007.

[15] T. Pérez-García, J.M. Iñesta, P.J. Ponce de León, and A. Pertusa, "A multimodal music transcription prototype," in *Proc. of International Conference on Multimodal Interaction, ICMI 2011*, Alicante, Spain, november 2011, pp. 315–318, ACM.

[16] Kris West, "Finding an optimal segmentation for audio genre classification," in *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR 2005)*, 2005, pp. 680–685.