EXPRESSIVE MODIFICATIONS OF MUSICAL AUDIO RECORDINGS: PRELIMINARY RESULTS

Marco Fabiani & Anders Friberg
Royal Institute of Technology (KTH), Stockholm (Sweden)
Speech, Music and Hearing (TMH), Music Acoustics Group

ABSTRACT

A system is described that aims to modify the performance of a musical recording (classical music) by changing the basic performance parameters tempo, sound level and tone duration. The input audio file is aligned with the corresponding score, which also contains extra information defining rule-based modifications of these parameters. The signal is decomposed using analysis-synthesis techniques to separate and modify each tone independently. The user can control the performance by changing the quantity of performance rules or by directly modifying the parameters values. A prototype Matlab implementation of the system performs expressive tempo and articulation modifications of monophonic and simple polyphonic audio recordings.

1. INTRODUCTION

A music performance represents the interpretation that a musician (or a computer in our case) gives to a score. To obtain different performances, the musician often follows some principles related to structural features of the score (e.g. musical phrases). The KTH rules system for musical performance [5] models such principles in a quantitative way in order to play back a nominal score expressively using a sequencer and a synthesizer. This approach often produces rather unnatural results if compared to human performances, mostly because of the quality of the synthesizer. Therefore we propose here an alternative approach: directly modify a recorded human performance. This modification can be rule-based or controlled in other ways. Our aim is to obtain a system that can be used both for the analysis of music performance, as well as a tool to modify a performance in a controlled way, in order to be used for example in live interactive systems similar for example to [4].

We will concentrate on the musical parameters tempo, dynamics (sound level) and articulation (tone duration) which have been found to be crucial for performance expression [8]. To modify these parameters a few problems need to be solved. First of all, the rules system requires the division of the musical piece into phrases. This process is usually done by hand since automatic methods are not yet reliable enough. A second problem, since we aim at modifying each note independently, is the separation of each tone, or at least chord, from the mixture (especially in a polyphonic case). Finally, the modifications should be accurate and possibly avoid artifacts. To solve the first two problems, we propose to combine the use of score files aligned with the audio file. The third problem is approached using analysis-synthesis techniques.

2. SYSTEM OVERVIEW

The system can be divided into three main parts, as shown in Figure 1. In the first part (a), the audio file is analyzed and aligned with the score. This is done prior to the performance and the information stored for later use. The audio analysis separates the harmonic part of the signal from the noise and transients. In the second part (b, modification and synthesis), tempo, tone duration and sound level are modified according to the output values of the control block (c). These values are defined for example by the rules system or in other ways (movements detected by cameras, sensors or other controllers).

3. SCORE ALIGNMENT

In order to use the information provided in it, the score needs to be aligned with the audio file. This is done by defining a number of corresponding fixed points in the two files, which in our system are tone onsets. Tone onsets are automatically extracted, with the possibility to correct errors by hand, since an accurate alignment is crucial for analysis. Accurate automatic onset detection is still an open problem. Many algorithms have been proposed...
5. MODIFICATIONS AND SYNTHESIS

Previous attempts of expressive modifications of tempo can be found, for example, [6, 7]. Our aim is to go beyond tempo and modify each tone independently in order to change also the articulation (note length relative to the Inter Onset Interval) and the sound level. Sound level and tone duration modifications are performed on the spectrogram by modifying, adding and removing peaks according to the mapping obtained during the analysis.

To obtain a realistic sound level modification both the amplitude of the tone and its timbre need to be changed (usually louder tones have a brighter sound, [11]). This can be simulated by adding or removing entire partials tracks (see [3]) as well as changing the spectral shape of the tone in the time-frequency domain. Another solution is the use of an appropriate filter (e.g. shelf filter with variable slope) in the time domain. In case we want to make the timbre brighter, higher partials might not be present in the original recording. The filter solution will then only raise the noise level, while the spectrogram approach, on the other hand, allows synthesizing higher harmonics which are not present. It is also important to notice that in order to change the sound level according to the rule values, we need to know the original sound level of each note in the recording.

Modifications of tone duration (articulation) are performed by shortening or extending the partials tracks associated with that tone in the same way that the timbre is modified (e.g. by removing or adding peaks in the spectrogram).

For time scale modifications we implemented a simple approach in our prototype system: to make tempo faster frames are dropped from the spectrogram, while to make tempo slower some frames are used twice. This is done only in the sustain part of the tone in order to preserve tone attacks and avoid attack duplication.

Once the modification of the spectrogram is complete, the new performance is synthesized by inverting the analysis operation. Particular attention has to be paid to the phase response when modifying the signal in the time-frequency domain. Most of the artifacts (as for example “phasiness”, also known as “loss of presence”) are caused by a non-consistent phase response, meaning that no real signal exactly corresponds to that particular combination of magnitude and phase frequency response. To solve the problem we try to compute the inverse transform using only the spectrogram by iteratively minimizing an error function (see [13]). An alternative method would be to correct the original phase response.

6. PERFORMANCE CONTROL

The control part is where the user interacts with the system in order to obtain the desired performance. The control over the parameters can be done directly by the user, or can be partly automated by using the KTH rule system for music performance [5].
pDM [4] is a system for both rule based and directly controlled performance modifications of symbolic music files based on the KTH rule system, which consists of 19 rules: 14 rules influence tempo, 11 influence sound level and 5 influence articulation. Each rule produces a default parameter change for each note based on context-dependent information such as rhythm, phrasing, pitch, and expressive signs in the score. pDM uses a score file in which these default values are stored together with the notes. The final values for tempo, sound level and articulation are obtained by computing a weighted sum of the default values given by the rules. The weights are controlled by the user using for example sliders or previously defined in default sets for standard performances (e.g. happy, angry, sad, . . . ). Our system works in a similar fashion, but is using audio recordings instead of symbolic data and a synthesizer. The audio file is aligned with the pDM score in order to use the rule values. The same functionalities present in pDM are implemented so that a pDM score file can be used for applying each rule individually to the each tone in the audio recording.

7. PRELIMINARY RESULTS

In order to test the timing accuracy of the system, an expressive performance was transformed into a dead-pan performance and the resulting onset accuracy was measured. The audio input was a piano rendering of the first 11 bars of Chopin’s Etude op.10 no.3. It was a recording of a grand piano (Bösendorfer CEUS), controlled via MIDI. The performance contained large tempo variations and originated from an average of 22 real performances used in the RENCON contest 2006. A recorded deadpan version was also available (denoted Deadpan), which was used in an informal listening test in a comparison with the generated dead-pan (denoted Flattened).

In Figure 2, the resulting Inter Beat Interval for the original performance version, the recorded dead-pan version and the generated dead-pan version are shown. Note that the IBI (Inter Beat Interval) values are approximate because of the relative uncertainty of onset detection: this may explain the small mean error ($\text{std} = 9.0 \text{ms}$) for the Deadpan performance. It is indeed evident that the Flattened version is very close to the Deadpan version, regarding both overall tempo (mean IBI $m = 399.9 \text{ms}$) and error ($\text{std} = 9.7 \text{ms}$), as opposed to the original performance ($m = 472.9 \text{ms}$ and $\text{std} = 92.6 \text{ms}$). The informal listening test corroborated the measured results.

Some preliminary tests were done for the articulation modification. The length of the tones were shortened (e.g. to obtain a more staccato performance) by canceling the magnitude peaks of the tone’s partials in the time-frequency representation. The length of the single note was computed using rule values from the pDM score. The final result of this modification strongly depends on the quality of the analysis/separation operation. Artifacts were audible, caused by the fact that tones overlap both in frequency and in time: when canceling a part of a tone we might cancel parts of other tones, as well as modify the residual part (e.g. noise and transients). Furthermore, as mentioned in Section 5, artifacts were also caused by the non-consistent phase response.

The last parameter to modify is the sound level. The measurement of the sound level of each note is not trivial for a polyphonic recording. Many factors have to be taken into consideration, like for example masking effect and note overlapping. We tried to compute the sound level of each single tone by summing the energy of the partials we extracted during the analysis, but the results were not very satisfactory, especially for piano recordings.

Another problem that we tried to solve is the phase incoherence. Since modifications of the spectrogram are done at different levels and different times, we decided to discard the original phase response and use only the magnitude response. We thus implemented the algorithm proposed in [13]. We are currently also trying to use the original phase response for the unmodified parts and a generated one for the modified parts.

Finally, all parts were integrated into a prototype system controlled through a simple GUI (Figure 3). The user can: load audio and score files; automatically extract tone onsets and correct errors by hand; transform and analyze the signal using different window sizes and overlapping ratios; define rule values using sliders or choose performance presets; change the overall tempo; playback the original recording and the generated performance.

8. DISCUSSION AND FUTURE WORK

A system prototype was presented that can be used to modify the expression of a recorded musical performance. Possible applications of this system could be to gener-
Figure 3. GUI of the prototype system developed in Matlab.

imate realistic and precisely controllable performances of the same piece to be used for listening experiments (e.g. psychology, neuroscience). We can also think about an orchestra conducting system or other interactive applications. Given that some of the current problems can be solved, the system should sound more natural when compared to other performance systems based on MIDI sequencers and synthesizers [4]. It should also allow the user to use any kind of recording, which is not possible using other conducting systems [10] based on audio recordings, as well as allowing modifications of articulation and sound level on a note basis.

Many open problems need to be solved to obtain audio performances which are free from artifacts. The analysis-synthesis process needs to be improved in order to obtain better tone separation. An open issue is the choice of the most suitable time-stretching approach. We also need to find a solution to remove the phase artifacts introduced in the synthesis stage. Finally, after testing the various parts of the system using Matlab, we would like to implement a real-time version. This is not infeasible given that the analysis part of the computations can be done in advance and only the modification/synthesis part would run online.

9. REFERENCES


