

# A prototype system for rule-based expressive modifications of audio recordings

**Marco Fabiani and Anders Friberg**

Department of Speech, Music, and Hearing, Royal Institute of Technology, Sweden

A prototype system is described that aims to modify a musical recording in an expressive way using a set of performance rules controlling tempo, sound level, and articulation. The audio signal is aligned with an enhanced score file containing performance rules information. A time-frequency transformation is applied, and the peaks in the spectrogram, representing the harmonics of each tone, are tracked and associated with the corresponding note in the score. New values for tempo, note lengths, and sound levels are computed based on rules and user decisions. The spectrogram is modified by adding, subtracting, and scaling spectral peaks to change the original tone's length and sound level. For tempo variations, a time scale modification algorithm is integrated in the time domain re-synthesis process. The prototype is developed in Matlab. An intuitive graphical user interface (GUI) is provided that allows the user to choose parameters, listen, and visualize the audio signals involved, as well as perform the modifications. Experiments have been performed on monophonic and simple polyphonic recordings of classical music for piano and guitar.

*Keywords:* automatic music performance; performance rules; musical expression; emotions; audio signal processing

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, harmony, melody). The KTH rules system for musical performance (Friberg *et al.* 2006) models such principles in a quantitative way in order to control three main musical parameters: tempo, articulation, and sound level. The rules are used to play back MIDI files

expressively (Friberg 2006). The result sounds often unnatural, mostly because of the quality of the synthesizer.

We propose a different approach to automatic music performance in order to obtain a more realistic result: directly modify a recorded human performance. Previous attempts to make automatic expressive modifications of tempo have been suggested by, for example, Gouyon *et al.* (2003) and Janer *et al.* (2006). Interactive virtual conducting systems are other examples of expressive tempo and sound level modifications (Borchers *et al.* 2004). In this case the modifications are not automatic but controlled by the user. In our system, the modifications of the audio signal are done on a note basis, allowing also changes of the length of single tones (articulation).

We also take into account timbre variations of acoustic instruments when changing the sound level (Luce 1975). The whole process should avoid noticeable artifacts and work on monophonic and polyphonic recordings.

## METHOD

The system can be divided into three main sections as shown in Figure 1. In section (a), the audio signal is aligned with the score file, transformed into the time-frequency domain, and analyzed. In section (b), the modifications on the spectrogram, as well as the synthesis of the modified time domain signal, are performed. Note lengths, sound level, and tempo are computed in section (c) using rules values and inputs from the user.

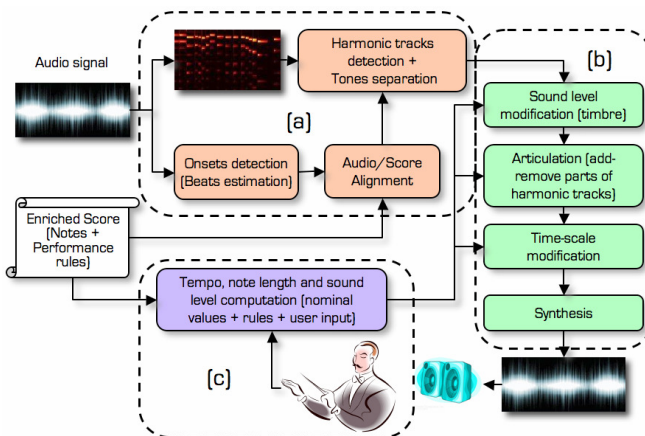


Figure 1. Schematic representation of the system. (See full color version at [www.performance-science.org](http://www.performance-science.org).)

## Level score alignment and signal analysis

In order to modify the performance on a note basis, each tone needs to be separated from the rest of the signal. In polyphonic recordings, tones can also overlap. A tone produced by an acoustic instrument is usually harmonic, with a large number of partials. To modify the single tone in the time-frequency domain, the partials need to be associated with their corresponding fundamental and note in the score.

The system uses an enhanced score file containing performance rules values. The score notes are also used in combination with the spectrogram to analyze the audio signal in order to separate the harmonic components of each tone. The score is aligned with the audio signal using tone onset positions, which can be extracted automatically (using a simple algorithm based on an edge detection filter), defined by hand or a combination of the two.

The signal is divided into overlapping time windows and transformed into a time-frequency representation using the method proposed by Ferreira and Sinha (2005). This method allows for accurate estimation of the frequency of spectral peaks. For each time window, the expected tone fundamental frequency and its partials are computed according to the notes in the score. The peaks in the spectrogram are detected and associated with the corresponding note in the score.

## Modifications and synthesis

The KTH rule system concentrates on the modification of tempo, sound level, and articulation; three acoustic parameters that have been found to be crucial for performance expression (Juslin 2000). In our prototype, the modifications are performed in the frequency domain using an analysis-synthesis approach in this order: articulation, sound level, and tempo.

Articulation is changed by lengthening (*staccato* to *legato*) or shortening (*legato* to *staccato*) the harmonic tracks corresponding to the tone. Using Ferreira's (2001) method, we interpolate the magnitude of the frequency peak and the two adjacent frequency bins and subtract them. In the same way, we can interpolate the magnitude of new peaks and add them to lengthen a tone.

Acoustic instruments usually sound brighter (e.g. higher partials are present) when played loud compared to when played soft. Therefore, to obtain a realistic sound level modification, the timbre also needs to be changed. Addition and subtraction of partials can be done using the same method applied for articulation. In addition, knowledge about the original sound level of each tone is needed in order to apply the correct amplitude

scaling. Measurement of the single tone level in a polyphonic recording is a complex problem that we have not yet solved. For this reason, in the prototype system, sound level modifications are currently not performed.

The modification of tempo is integrated in the synthesis algorithm. As mentioned earlier, the transformation to time-frequency domain is performed by first dividing the audio signal into overlapping time windows, separated by hop-size  $R_a$ . A common way to do time scale modifications (Laroche and Dolson 1999) is to modify the synthesis hop-size,  $R_s$ , so that the reconstructed time windows are more or less largely spaced (time scale expansion or compression). When  $R_s$  becomes too small or too large, audible artifacts are introduced. To avoid this problem, we either discard some frames or use the same frame twice. This approach has the side effect that it may also smear sharp tone attacks. By using  $R_s=R_a$  within tone attacks, we avoid this effect.

A major drawback of direct modifications of the spectrogram is phase incoherence, which introduces artifacts known as “phasiness” or “loss of presence.” The inverse transformation to a time domain signal requires both magnitude (spectrogram) and phase responses. Since only the magnitude is modified, the combination with the original phase response usually does not produce a real signal. Solutions have been proposed that try to correct the phase response to maintain coherence (Laroche and Dolson 1999) for time scale modifications. In our case, the problem is more complicated as we need to keep track of additions and subtractions of frequency peaks. For this reason, we decided to discard the original phase information and reconstruct the time domain signal from the magnitude only using the Real Time Iterative Spectrogram Inversion (RTISI) method (Beauregard *et al.* 2005). This algorithm also smears sharp tone onsets. Since we do not modify the magnitude response of the time frames containing onset data, for these frames we use the original phase response to prevent smearing, while for modified frames we use RTISI.

### **Performance values computation**

The modifications of the performance are based on a new value of sound level and length for each note, as well as a series of tempo values (usually one for each Inter Onset Interval). These values are the sum of the nominal value from the score and a delta value obtained from a weighted sum of the values of the rules. The weights are individually defined by the user or saved in default sets (e.g. happy, sad, angry, tender performance). There are 19 rules

in the system and each rule influences one or more of the acoustic parameters. For a more detailed explanation, refer to Friberg (2006).

## RESULTS

The system described above, with some limitations, has been implemented using Matlab. The user is provided with a simple graphical user interface (GUI) to load audio files and score files. The waveform is visualized together with tone onset points. These points can be detected automatically and manually corrected in case of errors in the detection. The user can choose some analysis parameters such as window and hop size. The user controls the overall tempo and performance parameters from the default sets or by using sliders for each rule. Before performing the synthesis, it is possible to choose whether or not to modify articulation. The sound level modification has not yet been implemented.

A few experiments using monophonic recordings of a theme from Haydn's *F Major Quartet* (Op. 74, No. 2), played with piano and guitar, showed good results for tempo modifications (sharp attacks are preserved). For articulation, a sort of reverberation effect is introduced in the silenced parts when the analysis is not able to extract all the frequency peaks. In the case of polyphonic recordings, the tempo modification does not introduce extreme artifacts, but the articulation is rather noisy, as the separation of partials becomes more complex with overlapping tones.

## DISCUSSION

In this paper, we presented a system that aims to modify a musical recording (tempo, sound level, and articulation of each single tone) in order to obtain an automatic performance comparable to a human performance in terms of expressivity and sound quality. The main problem is the separation of each single tone from the rest of the recording. We use a time-frequency representation and extract harmonic tracks corresponding to each tone. This is not yet reliable enough, and we are investigating how to improve the tone separation. The articulation of single notes strongly depends on the quality of the separation. Another open problem is that of measuring the sound level of the single tone in order to modify it consistently. A more reliable onset detection algorithm is also needed.

Possible applications for this system are in music cognition studies, where stimuli are usually artificial sounding MIDI files. Another example is the implementation of an advanced home conducting system that can work with any available recording.

### Address for correspondence

Marco Fabiani, Department of Speech, Music, and Hearing (TMH), Royal Institute of Technology (KTH), Lindstedtsv. 24, Stockholm, SE-10044, Sweden; *Email*: himork@kth.se.

### References

- Borchers J., Lee E., and Samminger W. (2004). Personal orchestra: A real-time audio/video system for interactive conducting. *Multimedia Systems*, 9, pp. 458-465.
- Beauregard G. T., Zhu X., and Wyse L. (2005). An efficient algorithm for real-time spectrogram inversion. *Proceedings of the 8<sup>th</sup> International Conference on Digital Audio Effects* (pp. 116-118). Madrid, Spain: Universidad Politécnica de Madrid
- Ferreira A. J. S. (2001). Combined spectral envelope normalization and subtraction of sinusoidal components in the ODFT and MDCT frequency domains. *Proceedings of the 2001 IEEE Workshop in Applications of Signal Processing in Audio and Acoustics* (pp. 51-54). New Paltz, New York, USA: IEEE.
- Ferreira A. J. S. and Sinha E. (2005). Accurate and robust frequency estimation in the ODFT domain. *Proceedings of the 2005 IEEE Workshop in Applications of Signal Processing in Audio and Acoustics* (203-206). New Paltz, New York, USA: IEEE.
- Friberg A. (2006). pDM: An expressive sequencer with real-time control of the KTH music performance rules. *Computer Music Journal*, 30, pp. 37-48.
- Friberg A., Bresin R., and Sundberg J. (2006). Overview of the KTH rule system for music performance. *Advances in Cognitive Psychology*, 2, pp. 145-161.
- Gouyon F., Fabig L., and Bonada J. (2003). Rhythmic expressiveness transformations of audio recordings: Swing modifications. *Proceedings of the International Conference on Digital Audio Effects (DAFX03)*. London: Queen Mary, University of London.
- Janer J., Bonada J., and Jorda S. (2006). Groovator: An implementation of real-time rhythm transformations. *Proceedings of the 121<sup>st</sup> Convention of the Audio Engineering Society*. San Francisco, California, USA: AES.
- Juslin P. N. (2000). Cue utilization in communication of emotion in music performance: Relating performance to perception. *Journal of Experimental Psychology: Human Perception and Performance*, 26, pp. 1797-1813.
- Laroche J. and Dolson M. (1999). Improved Phase Vocoder. Time-scale modification of audio. *IEEE Transactions on Speech and Audio signal processing*, 7, pp. 323-332.
- Luce D. A. (1975). Dynamic spectrum changes of orchestral instruments. *Journal of the Audio Engineering Society*, 23, pp. 565-568.