TempoExpress: An Expressivity-Preserving Musical Tempo Transformation System

Maarten Grachten and **Josep-Lluís Arcos** and **Ramon López de Mántaras**
IIIA-CSIC - Artificial Intelligence Research Institute
CSIC - Spanish Council for Scientific Research
Campus UAB, 08193 Bellaterra, Catalonia, Spain.
Vox: +34-93-5809570, Fax: +34-93-5809661
Email: {mantaras,arcos,maarten}@iiia.csic.es

Abstract

The research described in this paper focuses on global tempo transformations of monophonic audio recordings of saxophone jazz performances. More concretely, we have investigated the problem of how a performance played at a particular tempo can be automatically rendered at another tempo while preserving its expressivity. To do so we have developed a case-based reasoning system called *TempoExpress*. The results we have obtained have been extensively compared against a standard technique called uniform time stretching (UTS), and show that our approach is superior to UTS.

Introduction

In this paper we summarize the results obtained with a case-based reasoning system called *TempoExpress*, described in (Grachten, Arcos, & López de Mántaras 2004; 2006). *TempoExpress* preserves the expressivity of recorded performances while changing their tempo. That is, ideally listeners should not be able to notice from the expressivity of a performance that has been tempo transformed by *TempoExpress* that its tempo has been scaled up or down from another tempo. The system deals with monophonic audio recordings of expressive saxophone performances of jazz standards. The paper is organized as follows: The first section puts into context the problem of generating expressive music and tempo transformation. Next we briefly summarize the *TempoExpress* system and we describe the core result based on an extensive comparison of *TempoExpress* against uniform time stretching – a standard technique for changing the tempo in which note durations and timings are scaled by a constant factor proportional to the tempo change. Finally we present some conclusions.

The problem of generating expressive music

It has been long established that when humans perform music from score, the result is never a literal, mechanical rendering of the score (the so-called nominal performance). As far as performance deviations are intentional (that is, they originate from cognitive and affective sources as opposed to e.g. motor sources), they are commonly thought of as conveying musical expressivity, which forms an important aspect of music. Two main functions of musical expressivity are generally recognized. Firstly, expressivity is used to clarify the musical structure (in the broad sense of the word: this includes for example metrical structure (Sloboda 1983), but also the phrasing of a musical piece (Gabrielsson 1987), and harmonic structure (Palmer 1996)). Secondly, expressivity is used as a way of communicating, or accentuating affective content (Juslin 2001; Gabrielsson 1995).

The field of expressive music research comprises a rich and heterogeneous number of studies. Some studies are aimed at verbalizing knowledge of musical experts on expressive music performance. For example, Friberg et al. have developed Director Musices (DM), a system that allows for automatic expressive rendering of MIDI scores (Friberg *et al.* 2000). DM uses a set of expressive performance rules that have been formulated with the help of a musical expert using an analysis-by-synthesis approach (Sundberg, Friberg, & Frydén 1991). Widmer (Widmer 2000) has used machine learning techniques like Bayesian classifiers, decision trees, and nearest neighbor methods, to induce expressive performance rules from a large set of classical piano recordings. In another study by Widmer (Widmer 2002), the focus was on discovery of simple and robust performance principles rather than obtaining a model for performance generation. Hazan et al. (Hazan *et al.* 2006) have proposed an evolutionary generative regression tree model for expressive rendering of melodies. The model is learned by an evolutionary process over a population of candidate models. In the work of Desain and Honing and co-workers, the focus is on the cognitive validation of computational models for music perception and musical expressivity. They have pointed out that expressivity has an intrinsically perceptual aspect, in the sense that one can only talk about expressivity when the performance itself defines the standard (e.g. a rhythm) from which the listener is able to perceive the expressive deviations (Honing 2002). In more recent work, Honing showed that listeners were able to identify the original version from a performance and a uniformly time stretched version of the performance, based on timing aspects of the music (Honing 2006). Timmers et al. have proposed a model for the timing of grace notes, that predicts how the duration of certain types of grace notes

behaves under tempo change, and how their durations relate to the duration of the surrounding notes (Timmers *et al.* 2002). A precedent of the use of a case-based reasoning system for generating expressive music performances is the SaxEx system (Arcos, López de Mántaras, & Serra 1998; López de Mántaras & Arcos 2002). The goal of SaxEx is to generate expressive melody performances from an inexpressive performance, allowing user control over the nature of the expressivity, in terms of expressive labels like 'tender', 'aggressive', 'sad', and 'joyful'. Another case-based reasoning system is Kagurame (Suzuki 2003). This system renders expressive performances of MIDI scores, given performance conditions that specify the desired characteristics of the performance. Although the task of Kagurame is performance generation, rather than performance transformation (as in the work presented here), it has some sub tasks in common with our approach, such as performance to score matching, segmentation of the score, and melody comparison for retrieval. Recently, Tobudic and Widmer (Tobudic & Widmer 2004) have proposed a case-based approach to expressive phrasing, that predicts local tempo and dynamics and showed it outperformed a straight-forward k-NN approach.

An important issue when performing music is the effect of tempo on expressivity. It has been argued that temporal aspects of performance scale uniformly when tempo changes (Repp 1994). That is, the durations of all performed notes maintain their relative proportions. This hypothesis is called relational invariance (of timing under tempo changes). However, counter-evidence for this hypothesis has been provided (Desain & Honing 1994; Timmers *et al.* 2002), and a recent study shows that listeners are able to determine above chance-level whether audio-recordings of jazz and classical performances are uniformly time stretched or original recordings, based solely on expressive aspects of the performances (Honing 2006). Our approach also experimentally refutes the relational invariance hypothesis by comparing the automatic transformations generated by *TempoExpress* against uniform time stretching.

TempoExpress

Given a MIDI score of a phrase from a jazz standard, and given a monophonic audio recording of a saxophone performance of that phrase at a particular tempo (the source tempo), and given a number specifying the target tempo, the task of the system is to render the audio recording at the target tempo, adjusting the expressive parameters of the performance to be in accordance with that tempo.

TempoExpress solves tempo transformation problems by case-based reasoning. Problem solving in case-based reasoning is achieved by identifying and retrieving a problem (or a set of problems) most similar to the problem that is to be solved from a case base of previously solved problems (also called cases), and adapting the corresponding solution to construct the solution for the current problem.

To realize a tempo transformation of an audio recording of an input performance, *TempoExpress* needs an XML file containing the melodic description of the recorded audio performance, a MIDI file specifying the score, and the target tempo to which the performance should be transformed (the tempo is specified in the number of beats per minute, or BPM). The result of the tempo transformation is an an XML file containing the modified melodic description, that is used as the basis for synthesis of the transformed performance. For the audio analysis (that generates the XML file containing the melodic description of the input audio performance) and for the audio synthesis, *TempoExpress* relies on an external system for melodic content extraction from audio, developed by Gómez et al. (Gómez *et al.* 2003b). This system performs pitch and onset detection to generate a melodic description of the recorded audio performance, the format of which complies with an extension of the MPEG7 standard for multimedia content description (Gómez *et al.* 2003a).

We apply the edit-distance (Levenshtein 1966) in the retrieval step in order to assess the similarity between the cases in the case base (human performed jazz phrases at different tempos) and the input performance whose tempo has to be transformed. To do so, firstly the cases whose performances are all at tempos very different from the source tempo are filtered out. Secondly, the cases with phrases that are melodically similar to the input performance (according to the edit-distance) are retrieved from the case base. The melodic similarity measure we have developed for this is based on abstract representations of the melody (Grachten, Arcos, & López de Mántaras 2005) and has recently won a contest for symbolic melodic similarity computation (MIREX 2005).

In the reuse step, a solution is generated based on the retrieved cases. In order to increase the utility of the retrieved material, the retrieved phrases are split into smaller segments using a melodic segmentation algorithm (Temperley 2001). As a result, it is not necessary for the input phrase and the retrieved phrase to match as a whole. Instead, matching segments can be reused from various retrieved phrases. This leads to the generation of *partial* solutions for the input problem. To obtain the complete solution, we apply *constructive adaptation* (Plaza & Arcos 2002), a reuse technique that constructs complete solutions by searching the space of partial solutions.

The solution of a tempo-transformation consists in a performance annotation. This performance annotation is a sequence of changes that must be applied to the score in order to render the score expressively. The result of applying these transformations is a sequence of performed notes, the output performance, which can be directly translated to a melodic description at the target tempo, suitable to be used as a directive to synthesize audio for the transformed performance.

To our knowledge, all of the performance rendering systems mentioned in the previous section deal with predicting expressive values like timing and dynamics for the notes in the score. Contrastingly, *TempoExpress* not only predicts values for timing and dynamics, but also deals with more extensive forms of musical expressivity, like note insertions, deletions, consolidations, fragmentations, and ornamentations.

Results

In this section we describe results of an extensive comparison of *TempoExpress* against *uniform time stretching* (UTS), the standard technique for changing the tempo of audio recordings, in which the temporal aspects (such as note durations and timings) of the recording are scaled by a constant factor proportional to the tempo change.

For a given tempo transformation task, the correct solution is available as a target performance: a performance at the target tempo by a profesional musician, that is known to have appropriate expressive values for that tempo. The results of both tempo transformation approaches are evaluated by comparing them to the target performance. More specifically, let M_H^s be a melodic description of a performance of phrase p by a musician H at the source tempo s, and let M_H^t be a melodic description of a performance of p at the target tempo t by H. Using *TempoExpress* (TE), and UTS, we derive two melodic descriptions for the target tempo from M_H^s, respectively M_{TE}^t, and M_{UTS}^t.

We evaluate both derived descriptions by their similarity to the target description M_H^t. To compute the similarity we use a distance measure that has been modeled after human perceived similarity between musical performances. Ground truth for this was gathered through a web-survey in which human subjects rated the perceived dissimilarity between different performances of the same melodic fragment. The results of the survey were used to optimize the parameters of an edit-distance function for comparing melodic descriptions. The optimized distance function correctly predicts 85% of the survey responses.

In this way, the results of *TempoExpress* and UTS were compared for 6364 tempo-transformation problems, using 64 different melodic segments from 14 different phrases. The results are shown in figure 1. The figure shows the distance of both *TempoExpress* and UTS results to the target performances, as a function of tempo change (measured as the ratio of the target tempo to the source tempo). The lower plots show the significance value for the null hypothesis that the melodic descriptions generated by *TempoExpress* are not more similar or less similar to the target description than the melodic description generated using UTS (in other words, the hypothesis that *TempoExpress* does not give an improvement over UTS).

Firstly, observe that the plot in Figure 1 shows an increasing distance to the target performance with increasing tempo change (both for slowing down and for speeding up), for both tempo transformation techniques. This is evidence against the hypothesis of relational invariance discussed earlier in this paper. This hypothesis implies that the UTS curve should be horizontal, since under relational variance, tempo transformations are supposed to be achieved through mere uniform time stretching.

Secondly, a remarkable effect can be observed in the behavior of *TempoExpress* with respect to UTS, which is that *TempoExpress* improves the result of tempo transformation specially when slowing performances down. When speeding up, the distance to the target performance stays around the same level as with UTS. In the case of slowing down, the improvement with respect to UTS is mostly significant,

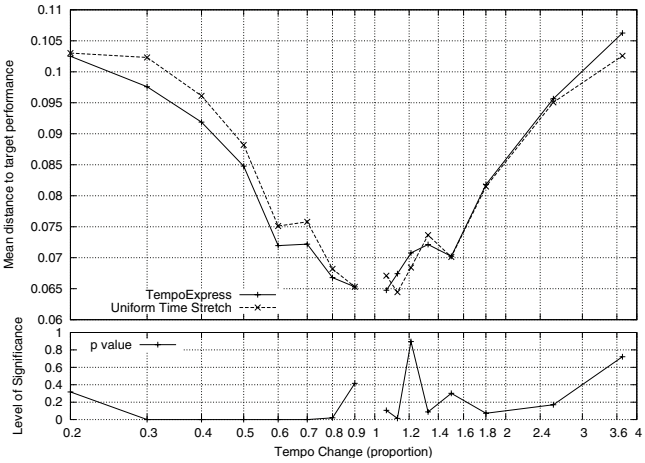

Figure 1: Performance of *TempoExpress* vs. UTS as a function of the ratio of target tempo to source tempo. The lower plot shows the probability of incorrectly rejecting H_0 for the Wilcoxon signed-rank tests

	mean distance to target		Wilcoxon signed-rank test		
	TempoExpress	UTS	p <>	z	*df*
tempo ↑	0.0791	0.0785	0.046	1.992	3181
tempo ↓	0.0760	0.0786	0.000	9.628	3181

Table 1: Overall comparison between *TempoExpress* and uniform time stretching, for upwards and downwards tempo transformations, respectively

as can be observed from the lower part of the plot. Note that the p-values are rather high for tempo change ratios close to 1, meaning that for those tempo changes, the difference between *TempoExpress* and UTS is not statistically significant. This is in accordance with the common sense that slight tempo changes do not require many changes, in other words, relational invariance approximately holds when the amount of tempo change is very small.

Table 1 summarizes the results for both tempo increase and decrease. Columns 2 and 3 show the average distance to the target performance for *TempoExpress* and UTS, averaged over all tempo increase problems, and tempo decrease problems respectively. The other columns show data from the Wilcoxon signed-rank test. The p-values are the probability of incorrectly rejecting H_0 (that there is no difference between the *TempoExpress* and UTS results). This table also shows that for downward tempo transformations, the improvement of *TempoExpress* over UTS is small, but extremely significant ($p < .001$), whereas for upward tempo transformations UTS seems to be better, but the results are slightly less decisive ($p < .05$).

Conclusions

In this paper we have summarized our research results on a case-based reasoning approach to global tempo transformationsof music performances, focusing on saxophone recordings of jazz themes. We have addressed the problem of

how a performanceplayed at a particular tempo can be automatically rendered at another tempo preserving expressivity. Moreover, we have described the results of an extensive experimentation over a case-base of more than six thousand transformation problems. *TempoExpress* clearly performs better than UTS when the target problem is slower than the source tempo. When the target tempo is higher than the source tempo the improvement is less significant. Nevertheless, *TempoExpress* behaves as UTS except in transformations to very fast tempos. This result may be explained by a lack of example cases with fast tempos.

References

Arcos, J. L.; López de Mántaras, R.; and Serra, X. 1998. Saxex : a case-based reasoning system for generating expressive musical performances. *Journal of New Music Research* 27 (3):194–210.

Desain, P., and Honing, H. 1994. Does expressive timing in music performance scale proportionally with tempo? *Psychological Research* 56:285–292.

Friberg, A.; Colombo, V.; Frydén, L.; and Sundberg, J. 2000. Generating musical performances with Director Musices. *Computer Music Journal* 24(1):23–29.

Gabrielsson, A. 1987. Once again: The theme from Mozart's piano sonata in A major (K. 331). A comparison of five performances. In Gabrielsson, A., ed., *Action and perception in rhythm and music*. Stockholm: Royal Swedish Academy of Music. 81–103.

Gabrielsson, A. 1995. Expressive intention and performance. In Steinberg, R., ed., *Music and the Mind Machine*. Berlin: Springer-Verlag. 35–47.

Gómez, E.; Gouyon, F.; Herrera, P.; and Amatriain, X. 2003a. Using and enhancing the current MPEG-7 standard for a music content processing tool. In *Proceedings of Audio Engineering Society, 114th Convention*.

Gómez, E.; Grachten, M.; Amatriain, X.; and Arcos, J. L. 2003b. Melodic characterization of monophonic recordings for expressive tempo transformations. In *Proceedings of Stockholm Music Acoustics Conference 2003*.

Grachten, M.; Arcos, J. L.; and López de Mántaras, R. 2004. TempoExpress, a CBR approach to musical tempo transformations. In *Advances in Case-Based Reasoning. Proceedings of ECCBR*. Springer.

Grachten, M.; Arcos, J. L.; and López de Mántaras, R. 2005. Melody retrieval using the Implication/Realization model. MIREX `http://www.music-ir.org/evaluation/mirex-results/articles/similarity/grachten.pdf` .

Grachten, M.; Arcos, J. L.; and López de Mántaras, R. 2006. A case based approach to expressivity-aware tempo transformation. *Machine Learning*. In press.

Hazan, A.; Ramirez, R.; Maestre, E.; Perez, A.; and Pertusa, A. 2006. Modelling expressive performance: A regression tree approach based on strongly typed genetic programming. In *Proceedings on the 4th European Workshop on Evolutionary Music and Art*, 676–687.

Honing, H. 2002. Structure and interpretation of rhythm and timing. *Tijdschrift voor Muziektheorie* 7(3):227–232.

Honing, H. 2006. Is expressive timing relational invariant under tempo transformation? *Psychology of Music*. (in press).

Juslin, P. 2001. Communicating emotion in music performance: a review and a theoretical framework. In Juslin, P., and Sloboda, J., eds., *Music and emotion: theory and research*. New York: Oxford University Press. 309–337.

Levenshtein, V. I. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady* 10:707–710.

López de Mántaras, R., and Arcos, J. L. 2002. AI and music: From composition to expressive performance. *AI Magazine* 23(3):43–58.

Palmer, C. 1996. Anatomy of a performance: Sources of musical expression. *Music Perception* 13(3):433–453.

Plaza, E., and Arcos, J. L. 2002. Constructive adaptation. In Craw, S., and Preece, A., eds., *Advances in Case-Based Reasoning*, number 2416 in Lecture Notes in Artificial Intelligence. Springer-Verlag. 306–320.

Repp, B. H. 1994. Relational invariance of expressive microstructure across global tempo changes in music performance: An exploratory study. *Psychological Research* 56:285–292.

Sloboda, J. A. 1983. The communication of musical metre in piano performance. *Quarterly Journal of Experimental Psychology* 35A:377–396.

Sundberg, J.; Friberg, A.; and Frydén, L. 1991. Common secrets of musicians and listeners: an analysis-by-synthesis study of musical performance. In Howell, P.; West, R.; and Cross, I., eds., *Representing Musical Structure*, Cognitive Science series. Academic Press Ltd. chapter 5.

Suzuki, T. 2003. The second phase development of case based performance rendering system "Kagurame". In *Working Notes of the IJCAI-03 Rencon Workshop*, 23–31.

Temperley, D. 2001. *The Cognition of Basic Musical Structures*. Cambridge, Mass.: MIT Press.

Timmers, R.and Ashley, R.; Desain, P.; Honing, H.; and Windsor, L. 2002. Timing of ornaments in the theme of Beethoven's Paisiello Variations: Empirical data and a model. *Music Perception* 20(1):3–33.

Tobudic, A., and Widmer, G. 2004. Case-based relational learning of expressive phrasing in classical music. In *Proceedings of the 7th European Conference on Case-based Reasoning (ECCBR'04)*.

Widmer, G. 2000. Large-scale induction of expressive performance rules: First quantitative results. In *Proceedings of the International Computer Music Conference (ICMC2000)*.

Widmer, G. 2002. Machine discoveries: A few simple, robust local expression principles. *Journal of New Music Research* 31(1):37–50.