Chapter 1

Introduction

This dissertation is an investigation into the use of case based reasoning for expressivity-aware tempo transformation of audio recorded performances of melodies. This specific task is illustrative of a wider application domain of science and technology that has emerged during the past decade, and the impact and importance of which is only recently being realized: content-based multimedia processing. Large scale access to the world wide web, and the omnipresence of computers have lead to a strong increase of image, video, and audio information available in digital format, and it is only realistic to assume that in the future the majority of information will be distributed in the form of multimedia. This shift toward non-text media asks for new ways of managing information. In parallel with the technology that aims at automatic semantic description of image, video, and audio information to make its content accessible, there is a need for technology that enables us to handle such information according to its content. This includes for example content based information retrieval, but also content based transformation, where previous information is reused for new purposes, often requiring non-trivial adaptations of the information to its new context.

Both in content extraction and manipulation progress has been made (see Aigrain [1999] for an overview). Nowadays high-quality audio time stretching algorithms exist (e.g. [Röbel, 2003; Bonada, 2000]), making pitch-invariant temporal expansion and compression of audio possible without significant loss in sound quality. Such algorithms perform *low level* content based transformation, i.e. by segmenting the audio signal into transient and stationary parts based on spectro-temporal content and stretching the audio selectively. The main goal of those algorithms is to maintain *sound* quality, rather than the *musical* quality of the audio (in the case of recorded musical performances). But as such, they can be used as tools to build higher level content based audio transformation applications. A recent example of this is an application that allows the user to change the swing-ratio of recorded musical performances [Gouyon et al., 2003]. Such audio applications can be valuable especially in the context of audio and video post-production, where recorded performances must commonly be tailored to fit specific requirements. For instance, for a recorded musical performance to accompany video, it must usually meet tight constraints imposed by the video with respect to the timing or the duration of the recording, often requiring a tempo transformation.

In order to realize a tempo transformation that maintains the musical quality of the musical performance, higher level content of the audio must be taken into account (as we will

argue in the next section). Content based transformation of music performances inevitably demands a thorough grip on musical expressivity, a vital aspect of any performed music. We use the term musical expressivity to refer to the deviations of the music as it is performed with respect to some norm, for example the score. This phenomenon is notoriously complex. Changes in the musical setting (for instance changes of tempo), often lead to subtle changes in expressivity that may nevertheless be indispensable to maintain a feeling of musical correctness. Through the use of case based reasoning as a state-of-the-art AI problem-solving methodology that has proved its merits in a variety of tasks, we try to realize tempo transformations of musical performances that are expressivity-aware. This means that apart from changing the rate at which the performance is being reproduced, changes to the expressive character of the performance are made to the extent that a human musician would change her way of performing to make the performance sound good at the new tempo.

The task and chosen approach raise a number of challenging topics. First there is the question of data modeling. It is an open issue how expressivity information can be extracted from the performance and appropriately represented, and what aspects of the melody should be explicitly described for content based manipulation of performed music. Secondly, from the case based reasoning perspective tempo-transformation of performed melodies is an interesting problem domain, since the problem and solution data are composite structures of temporal nature, and the domain expertise (expressively performing music) is almost entirely a tacit skill. Thirdly, since problem solving in case based reasoning is based on the reuse of previous problems, similarity measures for melodies and performances play a central role. This creates an overlap with the field of music information retrieval. Lastly, this research raises the question of how the quality of transformed performances can be evaluated. The evaluation of models for expressive music performance is an important unsettled issue, that deserves broad attention.

In the remainder of this chapter, we explain the problem of expressivity-aware tempo-transformation in more detail (section 1.1). Then we will outline the scope and the specific problems addressed in this dissertation (section 1.2). Finally, we give a brief overview of the structure of the dissertation (section 1.3).

1.1 Musical Expressivity and Tempo

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*). Even when musicians intentionally play in a mechanical manner, noticeable differences from the nominal performance occur [Seashore, 1938; Bengtsson and Gabrielsson, 1980]. Furthermore, different performances of the same piece, by the same performer, or even by different performers, have been observed to have a large number of commonalities [Henderson, 1937; Seashore, 1938]. Repp [1995b] showed that graduate piano students were capable just as well as professional piano players, of repeatedly producing highly similar performances of the same piece.

Given that expressivity is a vital part of performed music, an important issue 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). Counter-evidence for this hypothesis has also been provided

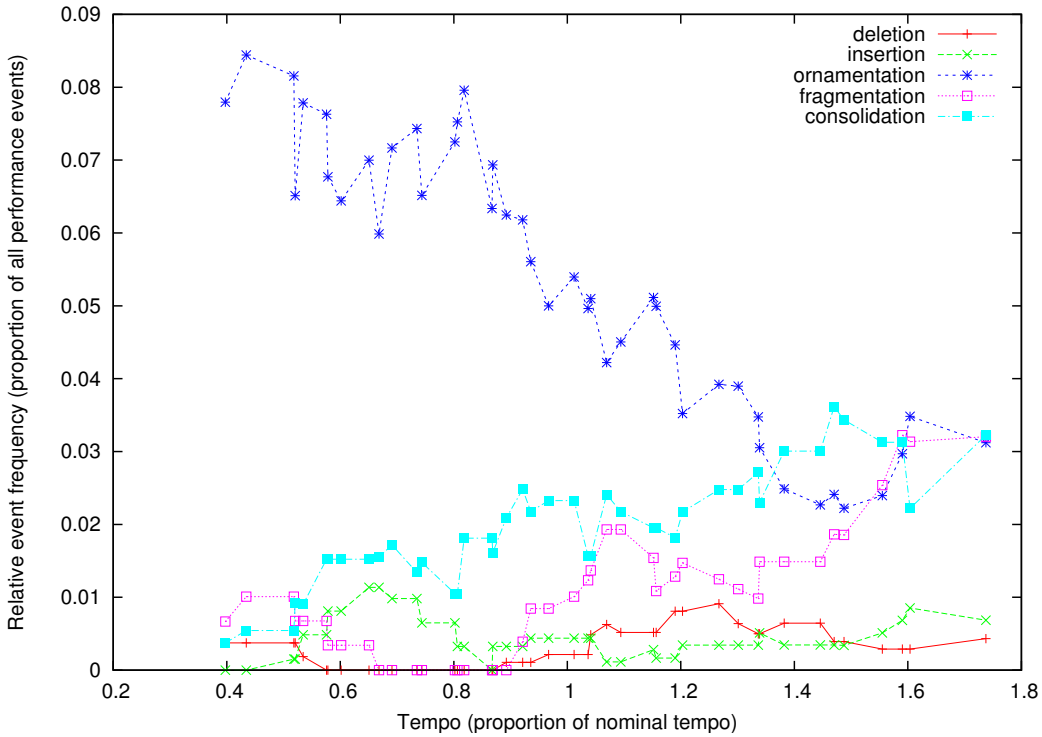

Figure 1.1: The frequency of occurrence of several kinds of performance events as a function of global performance tempo

however [Desain and Honing, 1994; Friberg and Sundström, 2002; 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, 2007].

A brief look at the corpus of recorded performances we will use in this study (details about the corpus are given in subsection 3.2) reveals indeed that the expressive content of the performances varies with tempo. Figure 1.1 shows the frequency of occurrence of various types of expressivity, such as *ornamentation* and *consolidation*, as a function of the nominal tempo of the performances (the tempo that is notated in the score). In subsection 4.3 we will introduce the various types of performance events as manifestations of musical expressivity in detail. Note that this figure shows the occurrence of discrete events, rather than continuous numerical aspects of expressivity such as timing, or dynamics deviations. The figure clearly shows that the occurrence of certain types of expressivity (such as ornamentation) decreases with increasing tempo, whereas the occurrence of others (consolidation most notably) increases with increasing tempo.

Figure 1.2 shows how various expressive parameters change systematically with tempo. The points in the figure represent comparisons of performances of the same phrase at differ-

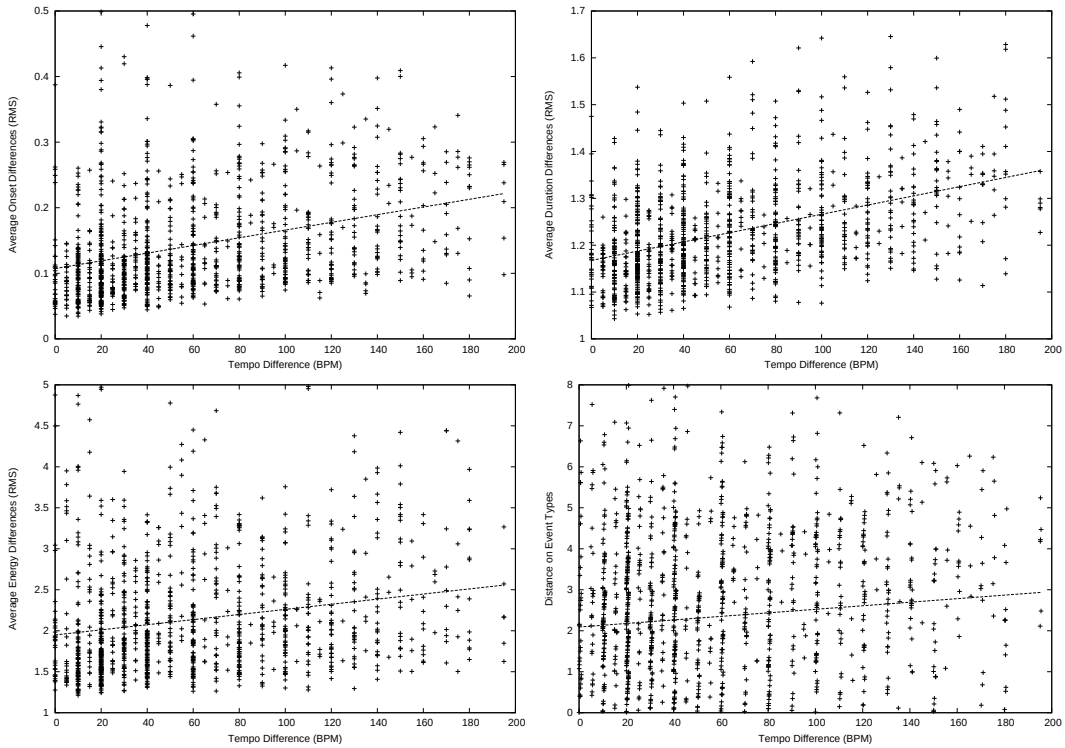

Figure 1.2: Dissimilarities between performances (of the same phrase) vs. their difference in tempo

ent tempos (tempos are specified in the number of *beats per minute*, or *BPM*). The x-axis shows the difference in tempo between the two performance. In the top left figure, the y-axis shows the *root mean square* (RMS) value of the pairwise absolute difference in note onset. The top right figure shows the RMS value of the pairwise duration difference (proportion). The bottom left figure shows the RMS value of the pairwise energy difference (proportion). The bottom right figure shows the distance value between the performances as sequences of performance events. The distance increases when the sequences contain different performance events. In all four expressive parameters, values tend to differ more when the tempos of the compared phrases increases. In some parameters the change as a function of tempo seems only small, but it must be kept in mind that the actual performances are a result of the combination of all parameters. The effects in the individual parameters are therefore cumulative.

The above observations amount to the belief that although in some circumstances relational invariance may hold for some aspects of expressivity, in general it cannot be assumed that all aspects of expressivity remain constant (or scale proportionally) when the tempo of the performance is changed. In other words, tempo transformation of musical performances involves more than *uniform time stretching* (*UTS*).

Throughout this dissertation, we will use the term UTS to refer to the scaling of the *temporal* aspects of a performance by a constant factor. For example, dynamics, and pitch will be left unchanged, and also no notes will be inserted or removed. Only the duration and onsets of notes the will be affected. Furthermore, we will use the term UTS in an abstract sense. Depending on the data under consideration it involves different methods to realize it. For example, it requires non-trivial signal-processing techniques to apply pitch-invariant UTS to the audio recording of the performance. In symbolic descriptions of the performance on the other hand, UTS consists in a multiplication of all temporal values by a constant. Note that this holds if the descriptions measure time in absolute units (e.g. seconds). When time is measured in score units (e.g. beats) UTS makes no sense, since changing the tempo of the performance only changes the translation of score time units to absolute units of time.

1.2 Problem Definition, Scope and Research Objectives

In this section we describe the main problem we address in this dissertation. We define the scope of the project with regard to the type of musical data and the level of processing. After that we will list the secondary objectives that derive from the main problem.

1.2.1 A System for High-level Content-Based Tempo Transformation

As mentioned at the beginning of this chapter, the primary objective of this research is to develop an expressivity-aware system for musical tempo transformations of audio recorded performances, that maintains not only the *sound quality* of the recording, but also *musical quality* of the performance. There are several ways to concretize the criteria for success of the system. For instance, we can say the tempo-transformation of a recorded performance is successful if:

- its expressive characteristics (statistically speaking) are in accordance with human performances at that tempo (more than with human performances at other tempos);
- it is preferred by human listeners over a tempo-transformed performance that was obtained by UTS;
- it is not recognized by human listeners as being a manipulated performance; they regard it as an original recording of a human performance;

Although there is not a strictly logical relation between the above criteria, we feel that in practice, each former criterion is implied by the subsequent criteria. That is, they are ordered from weak to strong. In subsection 2.4.6 we review evaluation paradigms for expressive music prediction models in general, and in subsection 6.3 we propose a hybrid evaluation approach that combines human judgment with a quantitative assessment of tempo-transformed performances.

1.2.2 Scope

In order to optimize the conditions for investigation of expressivity aware tempo transformation and to make the problem feasible as a project, the scope of research must be chosen appropriately. We work with a corpus of recorded performances that has some rather specific characteristics, that make it suitable for this research. Firstly, the recordings are performances of jazz standards, taken from The Real Book [2004]. Jazz is a good genre for studying expressivity because of its emphasis on liberal and expressive performance rather than precise reproductions of notated melodies. Secondly, the performances are played by a saxophone, an instrument that offers a very broad range of sound-qualities that allow a skilled performer to perform expressively in an elaborate manner. Thirdly, the melodies are monophonic, which relieves the need for voice separation, and lets us focus directly on the expressivity in the performance. Finally, the type of expressivity in the recording is what we call 'natural': the melodies are performed as the performer (a professional musician) thinks they should without any explicit intentions of expressing a particular mood or affection. See section 3.2 for more details on the musical corpus.

In addition to the focus on a specific type of musical data, we focus on a particular level of data processing. As explained before, we will work within the paradigm that separates content extraction from content analysis/manipulation, as opposed to for example purely signal processing approaches to audio-transformation[1]. The research presented in this dissertation exclusively addresses content analysis/manipulation. For the content extraction (audio analysis) and reconstruction of audio from content descriptions (audio re-synthesis), we rely on an external system for melodic content extraction from audio, developed by Gómez et al. [2003b,a].

1.2.3 Representation of Expressivity

One of the secondary objectives is the development of a suitable representation scheme for the expressivity in the data that we work with. This will require a study of the expressive phenomena encountered in the musical corpus.

1.2.4 Melodic Retrieval Mechanisms

The case based approach to the tempo transformation problem implies the need for a retrieval mechanism for cases that is based on the melodic material contained in the cases. We will investigate and compare several approaches to melodic similarity.

1.2.5 Performance Model Evaluation

In order to evaluate the proposed tempo transformation approach we need an evaluation methodology that assesses the quality of tempo transformed performances. We will discuss common evaluation paradigms.

[1] This two-level paradigm has been argued for by e.g. Scheirer [1995]

1.3 Outline of the Dissertation

In chapter 2, we provide the relevant background of the work presented in this dissertation. We discuss the principal concepts involved, notably musical expressivity, performance models and their evaluation, melodic similarity, and case based reasoning. Along with this survey we will review related research, methodologies, and notable findings.

In chapter 3 we propose a system architecture for the tempo transformation system. We define the terms that will be used, and summarize the goal and functionality of the different components. Furthermore, we detail the musical corpus used for experimentation, we present various representation schemes of the musical data, and discuss their value for the current application.

In chapter 4 we explain in detail the methods we have developed to process the input data, in order to form cases that contain not just the raw input data, but provide higher level knowledge-enriched data descriptions that interpret and interrelate the input data, and are used at various stages throughout the case based reasoning process.

Chapter 5 is devoted to the problem solving process, that consists in applying case based reasoning to construct a tempo transformed performance of the input phrase.

Chapter 6 describes a set of experiments that validate the different components of the system, notably knowledge acquisition (performance annotation), and abstract melody representation, and melody retrieval. The final section of the chapter describes an experimental validation of the system as a whole.

In chapter 7 we summarize the research presented in the dissertation. We list the main contributions of our work, and identify future work.

The appendices respectively contain abbreviations/notational conventions, annotated scores of the phrases used as musical corpus to form the case base, a list of songs used for a melodic similarity comparison experiment, and a list of publications by the author.