# A Study of Synchronization of Audio Data with Symbolic Data

## Music254 Project Report
## Spring 2007
## SongHui Chon

**Abstract**

This paper provides an overview of the problem of audio and symbolic synchronization. This problem is not too difficult for human musicians, but is a different story for a computer. Two versions of problem, one realtime (score following) and the other non-realtime (score alignment) have been studied separately. One particular application of score following, automatic accompaniment, has also been presented separately. Voice accompaniment is mentioned also for its difficulty even among the automatic accompaniment problems. The two common techniques used for the synchronization problem – hidden Markov model (HMM) and dynamic time warping (DTW) have been also compared.

Chapter 1 provides a general overview of the problem. Then score alignment problem is described in chapter 2 and score following in chapter 3. The conclusion and future studies are presented in chapter 4.

## 1. Introduction

Music can be in different forms. It can be in a symbolic form, e.g. in the form of written score, which has information on pitch, duration, dynamics, etc. It also can be in a recorded form, more often in mp3 or wav formats these days. Humans can follow scores with not much trouble while listening to the same piece of music played from a recording. This is not as easy for a computer. The same is with accompaniment; human players can accompany quite successfully a soloist even when the soloist plays some wrong notes or in varying tempos, while a computer has a hard time to respond properly to such "deviations".

The score synchronization problem has attracted many research and many solutions have been proposed for different flavors of the problem. This paper studies the general problem of synchronization of audio data with symbolic data using computers. This

problem can be further divided into two problems – score alignment and score following. The score alignment problem, a non-realtime problem, is to synchronize audio data (recorded data in this case) with symbolic data, or audio data with audio data in two different recordings. In this problem, the computer does not have the same burden of having to respond in a very short time as in the real-time version.

The score following problem concerns realtime score synchronization between symbolic data (e.g. Score or MIDI) and the audio stream. A special example is the automatic accompaniment of a soloist by a computer. To do this, the computer needs to be able to "listen" to the soloist, 'detect' where they are in the specific piece of music at the given time, and "predict" what has to be done (e.g. playing the correct note), just like a human accompanist does. We will discuss the automatic accompaniment separately in section 3.1 of chapter 3, since it has the additional component of "playback" on top of the score following problem.

Various techniques are used to solve both problems. Among them are two very popular techniques – dynamic time warping (DTW) [Rabiner 1993] and hidden Markov model (HMM). We will look into these common techniques in following sections.

There are other research topics related to the synchronization problem, such as automatic transcription, performance style comparison and imitation, and query by humming. These are interesting problems by themselves, but will not be discussed in this paper.

## 2. Score Alignment

The score alignment problem is to associate events in a score with points in time axis of an audio signal [Orio 2001]. The main applications of score applications are [Orio 2001][Soulez 2003]

- Indexing of recorded media for content-based retrieval through segmentation
- Performance segmentation into note samples labeled and indexed to build a unit database
- Performance comparison for musicological research
- Construction of a new score describing the exact performance selected (including dynamics, mix information, lyrics, etc.)

The general procedure of a solution is [Soulez 2003]

1. Parse symbolic data (e.g. MIDI) into score events
2. Extract audio features from signal
3. Calculate local distances between score and performance
4. Compute optimal alignment path minimizing the global distance.

The score usually has some information on timing data such as the global tempo (for example andante, or a metronome marking) and the local tempo (e.g. ritardando). Still, there is much left to the performer's interpretation, which is why no two performance of the same piece of music are exactly the same in timing. Often they are of different length, therefore we need a way to match two different time lines (steps 3 and 4).

Two techniques are used for this time alignment – hidden Markov model (HMM) and dynamic time warping (DTW). Those two techniques are quite interchangeable for our score alignment problem [Durbin 1998].

### 2.1 Dynamic Time Warping

This method calculates local distances between two streams of data and chooses the path with the minimum overall distance with given constraints. Various features can be considered for the local distance calculation. [Orio 2001][Soulez 2003] use spectral features of the signal and attack/release note modeling.

[Dannenberg 2003] uses similarity matrix between the recorded audio data and the MIDI-generated audio data. They use chromagram to calculate similarity, since it is proven to be more efficient [Hu 2003] than other acoustic features such as MFCC [Logan 2000] and Pitch Histograms [Tzanetakis 2002].

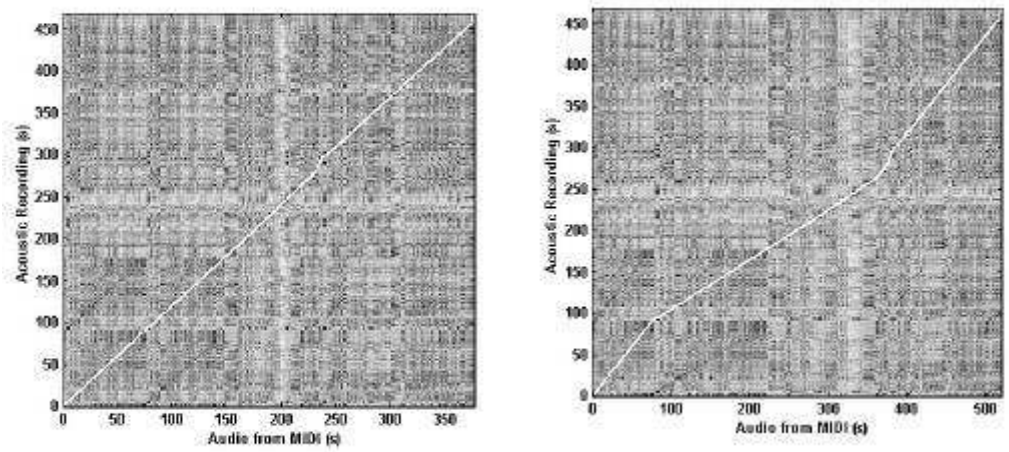


*Figure 1. Optimal alignment path of First movement of Beethoven Symphony No 5 (left) and that of the same piece artificially time-warped (right) (reproduction from* [Dannenberg 2003]*)*

In Figure 1, the x-axis is the time of MIDI-generated audio and y-axis the time in the audio recording. Notice different limits on the axes; the recorded audio data runs for a longer time than the MIDI-generated audio data. And yet as we can see from Figure 1, the alignment of score and audio data was successful even with artificially time-warped data (right).

In general, the optimal alignment path will be close to the diagonal of the similarity matrix. Therefore we can use some kind of pruning algorithm by eliminating the values far from the diagonal in the similarity matrix to keep the memory requirement manageable.

While DTW and HMM are completely interchangeable, DTW is simpler to implement and better for memory optimization than HMM [Orio 2001]. DTW also does not need training that HMM does.

### 2.2 Hidden Markov Model

Although HMM is more complicated to use than DTW, it can provide more general state transition possibilities. It also can be trained for a particular purpose, which is why it has been used in many researches including speech recognition and language processing.

For score alignment problem, the general algorithm using HMM is [Raphael 1999]:

1 Describe the likelihood of various segmentations by assigning probabilities using a

priori knowledge.

2 Develop a model that describes the likelihood of the given acoustic data using a hypothesized segmentation. A good training algorithm is necessary to learn efficient data model parameters with no supervision.

3 Calculate the globally optimal segmentation through dynamic programming, which minimizes segmentation errors.

Figure 2 shows the some examples of note models used in [Raphael 1999]. In the lower two models, the "articulation" state means the beginning of a note and is visited exactly once.

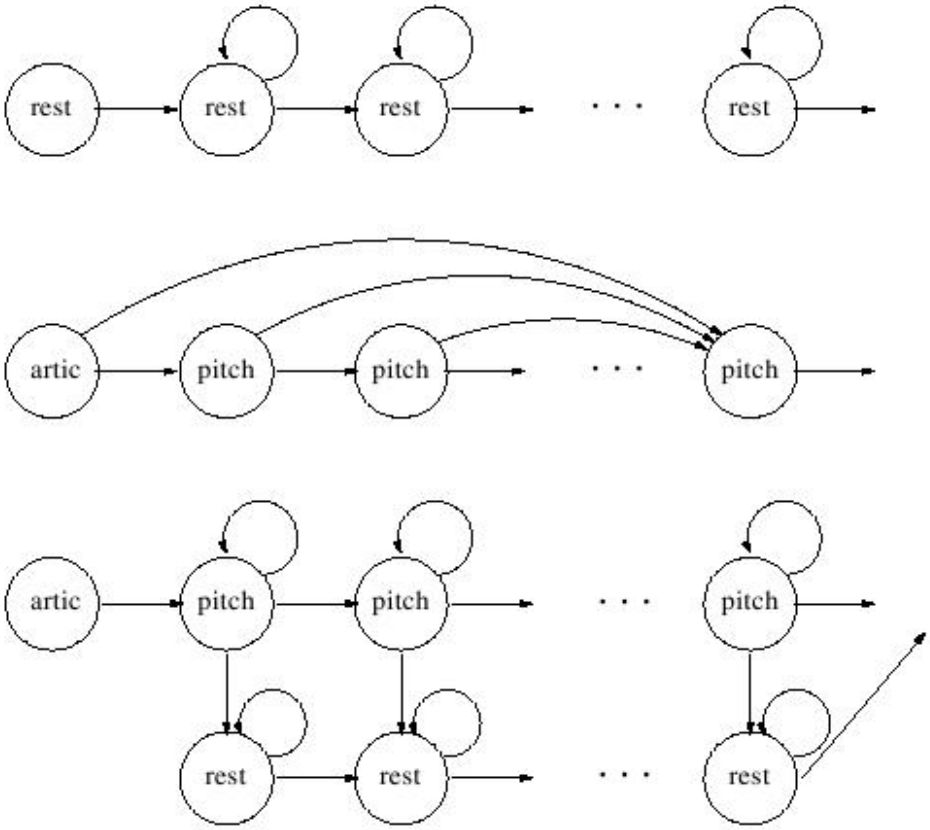


*Figure 2. Top: A long rest model. Middle: A short note model. Bottom: A note model with optional rest at end. (reproduction from* [Raphael 1999]*)*

[Raphael 1999] used HMM on monophonic instruments, while [Cont 2006] considered hierarchical HMM for polyphonic music. In both cases, the pitch and the duration information are assumed to be independent variables, which may not be a fair assumption. This assumption "allows the much greater freedom in the position of

note onset times and, in practice, puts more weight on the data model." However, with recorded audio data that has much detail (by placing mics closely, for example), deemphasizing the timing data may still yield better result [Raphael 2006].

## 3. Score Following Problem

The score following is a realtime version of the audio and symbolic data synchronization problem, hence there is the burden of low latency that does not exist for the score alignment problem. Still, because of the realtime nature of the problem, it has many popular applications such as virtual score (e.g. automatic page turning), automatic subtitles at an opera and automatic accompaniment. The automatic accompaniment will be described in a separate section, since it has the "playback" problem on top of the score following problem.

Since the score following deals with realtime audio stream data, the computer has to be able to extract necessary information from the audio input with low latency. There have been many successful note-based algorithms to estimate pitch for monophonic data, based on autocorrelation or spectral characteristics. However these techniques will not be feasible for polyphonic music in general. Instead, the idea of "compound events" (e.g. chords) are used for polyphonic signal. It does make sense to use chord-based techniques for polyphonic music, but at the same time it creates a bigger challenge of how to group notes.

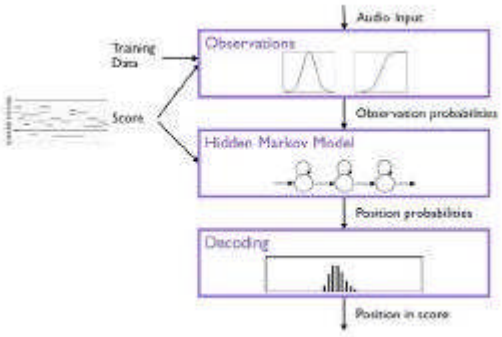


*Figure 3. Overview of a score following system (reproduction of* [Schwarz 2006]*)*

According to [Dannenberg 2003], the general procedure of score following is:

1 Convert symbolic data to audio data using a synthesizer, then to a spectral

format (such as chromagram, pitch histogram, MFCC, etc.)

2 Convert the given audio data (of performance) into the same spectral format

3 Align both spectral formats.

Score following can be implemented with either DTW or HMM (for step 3), just like in score alignment problem. Figure 3 shows an example system implemented with HMM (reproduction of [Schwarz 2006]). The HMM block can be replaced with a DTW block when applicable.

### 3.1 Automatic Accompaniment

The automatic accompaniment problem is a score following problem combined with realtime playback. In general, an automatic accompaniment system will

- Listen and analyze acoustic signal
- Anticipate using Bayesian belief network, and
- Synthesize output using a decision making system.

The "anticipation" part is innately human. Human players expect in advance what has to be played, according to the previous audio input, and accompany in response to the input, even when there is a wrong or missing note or a tempo change. Human players will get familiar with a soloist's playing style from rehearsals therefore be able to respond better. A computer will have a learning phase that will be analogous to rehearsals.

Then when the output needs to be generated for playback, it can be done in two ways. The easier option will be to synthesize audio output using MIDI, though it will not sound very "realistic". The other option is to use some techniques on sampled audio, such as phase vocoding [Raphael 2003-1&2] or synchronous overlap add (SOLA), which will be able to respond to local tempo changes without making any audible artifacts. The second case can be very useful especially with orchestral accompaniment [Raphael 2003-1&2], since it is hard to get a chance to play with a real orchestra when one studies the solo part of a concerto. There are pre-recorded systems such as Music Minus One that are available for this very purpose, but then the soloist needs to respond the dynamics of the orchestra, not the other way around. The automatic accompaniment system creates a more realistic rehearsal environment for the soloist.

There is a subproblem in automatic accompaniment that needs to be specifically mentioned, which is the vocal accompaniment problem [Puckette 1995][Grubb 1997]. This is different from the problems involving other instruments because of the natural characteristics of the voice. Voice is quite a challenge to process in audio format, because 1) it naturally has vibrato, a small change in pitch usually less than in a semitone, and 2) it may start without a specific note onset. The first characteristic makes it very hard to estimate a sung pitch and the second to detect the start of a note. Vowel detection algorithms are sometimes used for better performance.

[Puckette 1995] distinguished the instantaneous pitch from the "steady-state" pitch. The first has very little delay and therefore is useful for note onset detection. The latter is used for estimating the pitch of the sung note. A stochastic method was used in [Grubb 1997] using information such as recent tempo estimations, features extracted from the performance and elapsed time. Either formally or empirically, it estimates the probabilities that describe data.

## 4. Conclusion and Future Work

We have reviewed the problem of audio and symbolic data synchronization from a general point. We considered both the realtime version (i.e. Score following) and non-realtime version (i.e. Score alignment) and reviewed two popular techniques in solving both problems – DTW and HMM. For the score following problem, we also looked in to the automatic accompaniment.

Even though score and audio synchronization is considered largely solved in music information, there are still quite a lot of challenging problems that are yet to be resolved. There is a need for development of more reliable detection and tracking algorithms. A study of instruments that may not have strong onsets (e.g. Voice or strings) is also necessary. Development of a refined system structure will be helpful that is more modular and less inter-dependent. Lastly, but certainly not leastly, a thorough comparison of commonly used techniques such as HMM and DTW is due next from a systematic perspective.

## References

1. [Arifi 2004] Vlora Arifi, Michael Clausen, Frank Kurth, Meinard Muller:

Automatic synchronization of musical data: a mathematical approach, Computing in Musicology 13, 2003-04, pp. 9-33

2. [Cont 2004] Arshia Cont, Diemo Schwarz, Norbert Schnell: Training IRCAM's score follower
3. [Cont 2006] Arshia Cont: Real time audio to score alignment for polyphonic music instruments using sparse non-negative constraints and hierarchical HMMs, In IEEE ICASSP, 2006
4. [Dannenberg 1993] Roger Dannenberg: Music Understanding by Computer, IAKTA/LIST International workshop on Knowledge Technology in the Arts Proceedings, pp. 41-56, 1993
5. [Dannenberg 2001] Roger Dannenberg: Music information retrieval as music understanding, ISMIR 2001 Invited Address
6. [Dannenberg 2003] Roger Dannenberg and Ning Hu: Polyphonic audio matching for score following and intelligent audio editors, In Proc. Of ICMC, pp.27-33, 2003
7. [Durbin 1998] Richard Durbin, Sean R. Eddy, Anders Krogh, Graeme Mitchison: Biological sequence analysis: Probabilistic models of proteins and nucleic acids, Cambridge University Press, 1998
8. [Grubb 1997] Lorin Grubb, Roger Dannenberg: A stochastic method of tracking a vocal performer, In Proc. Of ICMC, 1997, pp. 301-308
9. [Hu 2003] Ning Hu, Roger Dannenberg and George Tzanetakis: Polyphonic audio matching and alignment for music retrieval, in IEEE workshop on Applications of signal processing to audio and acoustics, 2003, pp. 185-142
10. [Logan 2000] Beth Logan: Mel frequency cepstral coefficients for music modeling, In First international symposium on music information retrieval, 2000
11. [Orio 2001-1] Nicola Orio, Diemo Schwarz: Alignment of monophonic and polyphonic music to a score, In Proc. Of ICMC, 2001
12. [Orio 2001-2] Nicola Orio, Francois Dechelle: Score following using spectral analysis and hidden markov models, In Proc. Of ICMC, 2001
13. [Orio 2003] Nicola Orio, Serge Lemouton, Diemo Schwarz: Score following: state of the art and new developments, In Proc. Of NIME-03, pp. 36-41
14. [Puckette 1995] Miller Puckette: Score following using the sung voice, In Proc. Of ICMC, 1995
15. [Rabiner 1993] Lawrence Rabiner and Biing-Hwang Juang: Fundamentals of speech recognition, Englewood Cliffs, NJ: Prentice Hall, 1993

16. [Raphael 1999] Christopher Raphael: Automatic Segmentation of Acoustic Musical Signals Using Hidden Markov Models, IEEE Trans. On PAMI, 21(4): 360-370, 1999
17. [Raphael 2003-1] Christopher Raphael: Orchestra in a box: a system for real-time musical accompaniment, IJCAI 2003
18. [Raphael 2003-2] Christopher Raphael: Orchestra musical accompaniment from synthesized audio, In Proc. Of ICMC, 2003
19. [Raphael 2004] Christopher Raphael: Musical accompaniment systems, Chance Magazine volt 17:4, pp.17-22, 2004
20. [Raphael 2006] Christopher Raphael: Aligning music audio with symbolic scores using a hybrid graphical model, Machine Learning (2006) 65: 389-409
21. [Rehmeyer 2007] Julie Rehmeyer: The machine's got Rhythm, Science News, April 21, 2007 Vol.171, pp. 248-250
22. [Schwarz 2006] Diemo Schwarz, Arshia Cont, Nicolla Orio: Score Following at IRCAM
23. [Soulez 2003] Ferreol Soulez, Xavier Rodet, Diemo Schwarz: Improving Polyphonic and Poly-Instrumental Music to Score Alignment, ISMIR 2003
24. [Tzanetakis 2002] George Tzanetakis, Andrey Ermolinskyi and Perry Cook: Pitch histograms in audio and symbolic music information retrieval, ISMIR 2002