# MUSIC SCORE ALIGNMENT AND COMPUTER ACCOMPANIMENT

*By relating musical sound to musical notation, these systems generate tireless, expressive musical accompaniment to follow and sometimes learn from a live human performance.*

**By** Roger B. Dannenberg and Christopher Raphael

We are witnessing an explosion in the quantity and variety of music in computer-accessible form. There are primarily two kinds of music data: sampled audio files (such as those found on compact discs and scattered throughout the Web in various formats) and symbolic music representations, essentially listing notes with pitch, onset time, and duration for each note. Music audio is to symbolic music as speech audio is to text. In each case the audio representations capture the colorful expressive nuances of the performance, though anything other than a human listener will have difficulty "understanding" them. On the other hand, in

• Illustration by Robert Neubecker

both text and symbolic music, the high-level "words" are parsimoniously stored and easily recognized.

Here, we focus on a type of machine listening known as music score matching, score following, or score alignment. We seek correspondence between symbolic music representation and audio performance of the same music, identifying the onset times of all relevant musical "events" in the audio, usually notes. The two different versions of the matching problem are usually called "offline" and "online."

Offline matching uses the complete performance to estimate the correspondence between audio data and symbolic score. The offline problem allows one to "look into the future" while establishing correspondence. Thus an offline match can be viewed as an index into the performance, allowing random access to a recording. Such an index might allow a listener to begin at any location in a composition (such as the second quarter of measure 48) to link visual score representations with audio or to coordinate animation with prerecorded audio.

Offline score matching will soon enable many new and intriguing applications. For example, digital editing and post-processing of music often requires that the location of a particular note in an audio file be tuned, balanced, or tweaked in various ways. Score matching allows the audio engineer to automatically identify these locations, greatly simplifying the process. Another example involves musical sound synthesis, which generally relies on audio samples under a variety of conditions (such as pitch, dynamic level, and articulation). Score matching can be used to automate this arduous data-collection process. In a different direction, score matching also provides quantitative information about timing and tempo, aiding the study and understanding of musical expression. Score matching may also well be the key to studying many other musical attributes (such as dynamics, vibrato, and tone color) by providing note-based audio segmentation. Future synthesized music will greatly benefit from this inquiry.

Offline score matching's real-time cousin, sometimes called score following, processes audio data online as the signal is acquired, so "look ahead" is not possible. Score following aims to identify the musical events depicted in the score with high accuracy and low latency. The application dearest to our hearts is the accompaniment system, which generates a flexible musical accompaniment that follows a live soloist. "Hearing" a live player is a necessary ingredient in any solution. Other applications include the automatic coordination of audiovisual equipment with musical performance (such as opera supertitles and lighting effects), as well as real-time score-based audio enhancement (such as pitch correction and automatic page turning).

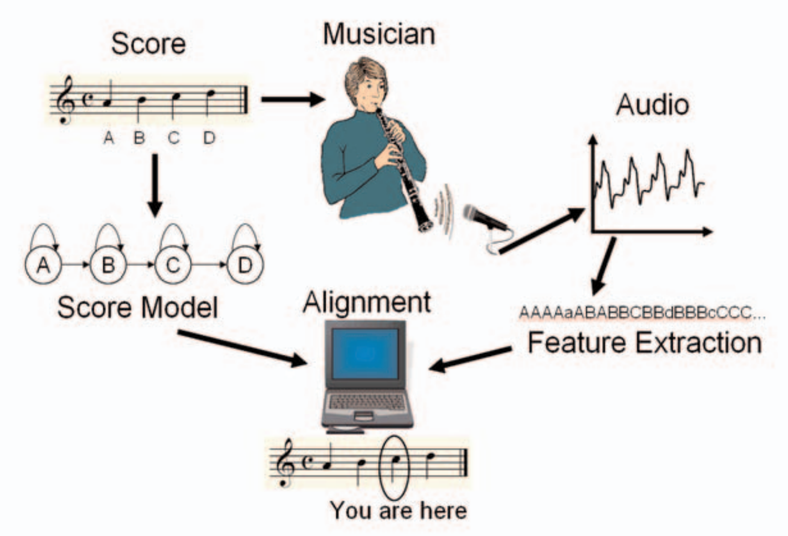

**Figure 1. Audio is analyzed ("feature extraction") to obtain a transcription of pitch estimates. These estimates correspond to a sequence of states from the score model derived from the symbolic score. An alignment process then relates the audio time position to the music score position.**

Perhaps the most direct way to explain the challenges of score following is with a simple example. Suppose a monophonic (one note at a time) instrument is playing from a known score containing only a few notes, say, ABCD (see Figure 1). Every 50 milliseconds or so, we estimate a pitch value from the audio signal. Pitch estimates are not reliable, but there are many estimates per note, resulting in something like:

AAAAAAAAaABABABBBCBBCBcBBBBBdBBBddCDCCCCCcCCCCDDBDdDDcDBDDDDDD

where lower-case letters denote a lower octave from the one notated in the score.

Score matching must be able to segment this sequence into regions corresponding to the note sequence ABCD. One approach, shown in Figure 1, uses a simple state graph, labeled "Score Model," with four possible states: A, B, C, and D. At each step, the model remains in the same state or advances to the next state. Since the input in this example has 62 pitch estimates, the system must consider all possible paths of length 62 through the graph. Each path is scored according to how closely it matches the actual data sequence. For each symbol, there is no penalty for an exact match, a small penalty for an octave or neighboring pitch error, and a large penalty for anything else. Thus every sequence has a total penalty (the sum of all penalties encountered), and an optimal state sequence can be defined as the one that globally minimizes this penalty. (You may be familiar with dynamic programming, which finds a globally optimal solution without enumerating every path.) One result is an optimal state sequence—the score match—such as:

AAAAAAAAAAAABBBBBBBBBBBBBBBBBBBBCCCCCCCCCCCCCCDDDDDDDDDDDDDDDDD

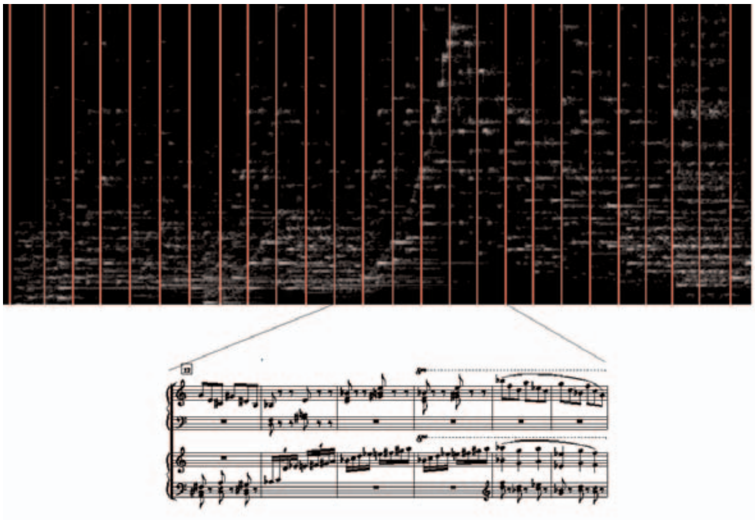

Figure 2. Spectrogram from the opening of "Mercury" from *The Planets* by Gustav Holst. The vertical lines indicate the positions of the bar lines in the music as determined through automatic score alignment. Clicks are added on the bar lines in the accompanying audio file.

While this is admittedly a toy example, its basic ideas can be extended and generalized to cover most of what would be encountered in a range of music. For instance, polyphonic music can be represented as a sequence of chords, where a new chord appears at every point at which a note begins or ends. The probability distribution on note length variation can be represented by allocating several graph states for each note and include transition probabilities.

A direct extension of the state graph idea to polyphonic music might replace the pitch estimates with sets of pitch estimates. The note sequence of the graph would be replaced by a chord sequence. Unfortunately, multi-pitch estimation is unreliable (and an active research area), so today's score-alignment systems use representations based on the distribution of frequencies in the signal, also known as the signal's spectrum. The spectrum is used because it is easily obtained from audio and reasonably predicted from the score.

One can use some basic knowledge of musical acoustics to predict spectra from chords. Each note in the chord produces a spectrum that can be approximated by a set of harmonics. A collection of pitches is modeled as a superposition of these one-note models. One can then base a penalty function on the distance between predicted and observed spectra, or on a probability model for the data spectrum, given the hypothesized notes of the chord. Another way to predict spectra is to play the score through a music synthesizer and compute spectra from the resulting audio. This method also opens the possibility of audio-to-audio alignment of two different performances.

A popular technique for comparing spectra from scores to spectra in music audio is to reduce the detailed spectrum to a 12-element vector, where each element represents all the energy in the spectrum associated with one of the 12 pitch classes (such as C, C-sharp, and D). The Euclidean distance between these so-called "chroma vectors" is a robust distance metric for score alignment [6].

Armed with these essential ideas, researchers have devised score matchers that handle realistic audio data; for example Figure 2 includes the spectrogram of an orchestra performance in which the recognized measure locations are indicated with vertical lines. Recognition of this audio was achieved by augmenting the notion of "state" discussed earlier to include time-varying tempo [8]. The corresponding audio file can be heard at xavier.informatics.indiana.edu/~craphael/acm with clicks added to mark the measures.

## Computer Accompaniment

Computer accompaniment represents the greatest success to date of online score following, and accompaniment systems show promise in several areas. The first is the traditional soloist-accompaniment scenario in which a live player wishes to play with a flexible, sensitive (and tireless) accompaniment. While human accompanists provide stiff competition in this traditional domain, accompaniment systems manage to beat their human counterparts in several ways. For example, accompaniment systems have nearly unlimited technical facility, allowing the coordination of arbitrarily fast notes and complex rhythms. Such virtuosity has been exploited in recent music composed explicitly for accompaniment systems by Jan Beran of the Universität Konstanz and Nick Collins of Cambridge University (xavier.informatics.indiana.edu/~craphael/music_plus_one/). Perhaps accompaniment systems will someday also find a place in jazz and other improvisatory music.

Most computer accompaniment systems are organized along the lines of Figure 3, including four fairly independent components: audio feature extraction; score following; accompaniment control; and music synthesis. Note that the input includes machine-readable scores for both the performer and the accompaniment, while the output is real-time musical accompaniment. The feature-extraction component performs signal processing on the incoming audio to generate a stream of data (such as pitch estimates and spectral frames). The score following component matches this sequence to the score to derive an evolving estimate of the current position of the soloist—the real-time version of the score alignment in Figure 1. The score follower provides a running commentary on the soloist's performance (such as "you are here"), providing estimates of solo event times with variable latency.

Accompaniment control is the brains of the accompaniment system, integrating the information in the score, musical constraints, and perhaps knowledge distilled from past performances to determine when to play accompaniment notes. The fourth component—music synthesis—generates the actual accompaniment sound, either by means of a note-based sound synthesizer or by resynthesizing prerecorded audio on-the-fly.

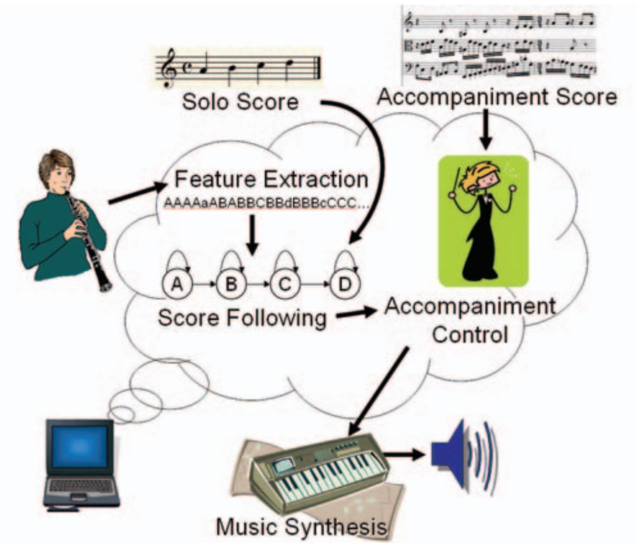

Figure 3. Structure of a computer accompaniment system.

An accompaniment system (or person) that tries to achieve synchrony by waiting to hear a note and then responding is always late, since all musical events are detected with latency. A better approach coordinates musical parts by extrapolating the future evolution of the music based on the detected note onsets, as well as on other information. The essential musical extrapolation challenge requires the fusion of several knowledge sources, including musical score, output of the score follower, and past training examples from the soloist and human-played accompaniment. Early accompaniment systems in the 1980s used hand-coded rules to guide this extrapolation [1, 3]. More recently, belief networks have begun to address the challenge of automatic learning and extrapolation from incomplete information [9].

The earliest computer accompaniment systems matched note sequences rather than spectral data, reducing data rates and computation [1, 3, 12]. Hidden Markov models are commonly used with spectral representations of audio and learn from rehearsal data to follow more accurately and respond with less latency [6, 10, 11]. Another approach—the Probabilistic Vocalist Tracker [4]—avoids the problem of discrete states by representing score position as a continuous probability density function (for a demonstration, go to www.cs.cmu.edu/~music/video/CANGIO.MOV). Whereas computer accompaniment emphasizes responsiveness and musical accompaniment, the main goal of offline score matching is accurate alignment—obtainable through short audio frames and specialized features for matching [7].

### Score Following Applications

Score following and computer accompaniment have enabled composers to exploit the virtuosic performance capabilities of computers and apply live digital audio signal processing to the sounds of live performers. Computer accompaniment is also being used in several notable music education systems:

**The Piano Tutor.** Score following is the basis for an intelligent piano tutoring system called the Piano Tutor [2] that models the interactions between students and their human teachers. The system has a repertoire of about 100 songs it can select for students. As they perform the music, real-time score following tracks their performance and automatically turns pages by updating the display. After the performance, the system uses score alignment to identify all skipped notes and extra notes. A student's tempo is automatically estimated, and timing errors are identified. Based on this analysis and guided by an expert system, a voice may say "Watch out, you missed a note here," while a box is drawn around the problem area in the music notation. Score following allows the Piano Tutor to evaluate relatively unconstrained piano performances, resulting in a simpler interface and (we think) a more enjoyable experience for students (www.cs.cmu.edu/~music/video/pianotutor.mov).

**SmartMusic.** SmartMusic (smartmusic.com) is a commercial software product for ordinary PCs. Using computer accompaniment as a tool for music education, it includes accompaniments for more than 30,000 pieces of music for wind instruments, vocalists, and beginning string players. Traditional music students practice alone but often lack the skills and motivation to practice effectively. SmartMusic challenges them to master pieces from beginning to end. When practicing with accompaniment, they hear their music in the context of the rhythm and harmony missing from solo parts. The accompaniment also provides a pitch reference so students are better able to hear when they are out of tune and when they must take corrective action. Although only small-scale studies of students practicing with accompaniment have been done (music.utsa.edu/tdml/conf-VI/VI-Repp/VI-Repp.html), they do indicate that those who practice with accompaniment improve their performance skill and confidence, feel greater motivation, and increase practice time.

**Music Plus One.** Music Plus One is a system primarily oriented toward Western art music in which the accompaniment is synthesized from prerecorded audio using a phase vocoder, thus allowing synchronization with the soloist while retaining the rich sonic palette and musicality of the original performance. The system performs score following with a hidden Markov model, coordinating parts with a belief network, including the accompaniment control component. This network predicts future accompaniment note onsets based on the score follower, the score itself, and examples of past solo and human-played accompaniment. These predictions create a trail of breadcrumbs the phase vocoder links together seamlessly (see xavier.informatics.indiana.edu/~craphael/music_plus_one for an example).

## Conclusion

The improving proficiency of offline and online score matching algorithms have made possible many new applications in digital audio editing, audio database construction, real-time performance enhancement, and accompaniment systems. While music is varied enough to violate almost any assumption, the related algorithms perform reliably even in established and challenging music domains while promising many new possibilities.

While many musicians use accompaniment systems to make practice more enjoyable and instructive, broader acceptance of the technology requires researchers to overcome a number of challenges, particularly how to capture a deeper understanding of musical aesthetics. This requires expertise in several areas, including musicology and computer science, that traditionally have had only limited awareness of one another.

Computer accompaniment is bound to find commercial applications touching the lives of millions of musicians and music lovers. Could musicians be displaced by these machines? We doubt this will be the predominant effect. On the contrary, accompaniment systems will make music more accessible to more people, fostering a greater appreciation and love for music making. The ultimate effect will be increased demand for "all-human" music, from listeners and participants alike.

## References

1. Baird, B., Blevins, D., and Zahler, N. Artificial intelligence and music: Implementing an interactive computer performer. *Computer Music Journal 17,* 2 (Summer 1993), 73–79.
2. Dannenberg, R., Sanchez, M., Joseph, A., Capell, P., Joseph, R., and Saul, R. A computer-based multimedia tutor for beginning piano students. *Interface Journal of New Music Research 19,* 2–3 (1993), 155–173.
3. Dannenberg, R. Real-time scheduling and computer accompaniment. In *Current Research in Computer Music,* M. Mathews and J. Pierce, Eds. MIT Press, Cambridge, MA, 1989, 225–261.
4. Grubb, L. and Dannenberg, R. A stochastic method of tracking a vocal performer. In *Proceedings of the 1997 International Computer Music Conference* (Thessaloniki, Greece, Sept. 25–30). International Computer Music Association, San Francisco, 1997, 301–308.
5. Hu, N., Dannenberg, R., and Tzanetakis, G. Polyphonic audio matching and alignment for music retrieval. In *Proceedings of the 2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (New Paltz, NY, Oct. 19–22). IEEE Computer Society Press, Piscataway, NJ, 2003, 185–188.
6. Loscos, A., Cano, P., and Bonada, J. Score-performance matching using hidden Markov models. In *Proceedings of the 1999 International Computer Music Conference* (Beijing, Oct. 22–27). International Computer Music Association, San Francisco, 1999, 441–444.
7. Orio, N. and Schwarz, D. Alignment of monophonic and polyphonic music to a score. In *Proceedings of the 2001 International Computer Music Conference* (Havana, Sept. 17–22). International Computer Music Association, San Francisco, 2001, 155–158.
8. Raphael, C. A hybrid graphical model for aligning polyphonic audio with musical scores. In *Proceedings of the Fifth International Conference on Music Information Retrieval* (Barcelona, Spain, Oct. 10–14). Audiovisual Institute Pompeu Fabra University, Barcelona, Spain, 2004, 387–394.
9. Raphael, C. A Bayesian network for real-time musical accompaniment. In *Proceedings of Neural Information Processing Systems (NIPS) 14* (Vancouver, B.C., Dec. 3–8). MIT Press, Cambridge, MA, 2001.
10. Raphael, C. Automatic segmentation of acoustic musical signals using hidden Markov models. *IEEE Transactions on PAMI 21,* 4 (Apr. 1999), 360–370.
11. Schwarz, D., Cont, A., and Schnell, N. From Boulez to ballads: Training IRCAM's score follower. In *Proceedings of the 2005 International Computer Music Conference* (Barcelona, Spain, Sept. 5–9). International Computer Music Association, San Francisco, 2005.
12. Vercoe, B. The synthetic performer in the context of live performance. In *Proceedings of the International Computer Music Conference 1984* (Paris, Oct. 19–23, 1984). International Computer Music Association, San Francisco, 1985, 199–200.

**Roger B. Dannenberg** (roger.dannenberg@cs.cmu.edu) is an associate research professor of computer science and art at Carnegie Mellon University, Pittsburgh, PA.

**Christopher Raphael** (craphael@indiana.edu) is an associate professor in the School of Informatics and an adjunct professor in the School of Music at Indiana University, Bloomington, IN.