CURRENT DIRECTIONS WITH MUSICAL PLUS ONE

Christopher Raphael
Indiana Univ., Bloomington
craphael@indiana.edu

ABSTRACT

We discuss the varieties of musical accompaniment systems and place our past efforts in this context. We present several new aspects of our ongoing work in this area. The basic system is presented in terms of the tasks of score following, modeling of musical timing, and the computational issues of the actual implementation. We describe some improvements in the probabilistic modeling of the audio data, as well as some ideas for more sophisticated modeling of musical timing. We present a set of recent pieces for live player and computer controlled pianos, written specifically for our accompaniment system. Our presentation will include a live demonstration of this work.

1 APPROACHES TO MUSICAL ACCOMPANIMENT SYSTEMS

Musical accompaniment systems are computer programs that serve as musical partners for live musicians. The types of possible interaction between live player and computer are widely varied, to some extent defying classification. Some approaches create sound by processing the live player's audio using simple analysis of the audio content itself, perhaps distorting, echoing, harmonizing, or commenting on the soloist's audio in largely predefined ways, [1], [2]. Other orientations are directed toward improvisatory music such as jazz, in which the computer follows and perhaps even composes a rendered musical part [3]. A third approach models the traditional "classical" concerto setting in which the computer's task is to perform a precomposed musical part in a way that follows a live soloist such as [4],[5]. There are a number of examples that blend these scenarios, while other approaches may be entirely outside this realm of possibilities.

Our work has focused on the latter "concerto-type" setting, as in a non-improvisatory composition for soloist and accompaniment — say a violin concerto. While the music has already been composed in this domain, the solo player may take great liberty with the performance, requiring the accompanying ensemble to be both flexible and responsive.

SMC 2009, July 23-25, Porto, Portugal

The motivation for this kind of accompaniment system is evident in the omitted for review (JSoM) at omitted for review where most of our recent experiments have been performed. For example, the JSoM contains about 200 student pianists while the regular orchestras perform two piano concerti every year using student soloists. With this in mind, it is clear that most of these aspiring pianists will never perform as orchestral soloist during their studies here. We believe this is truly unfortunate, as nearly all of these students have the necessary technical skills and musical depth to greatly benefit from the concerto experience. Our work in musical accompaniment systems strives to bring this rewarding experience to the music students, amateurs, and many others who would like to play as orchestral soloist, though, for whatever reason, don't have the opportunity.

Even within the realm of classical music, there are a number of ways to cast the accompaniment problem, requiring substantially different approaches. For instance, when accompanying early-stage musicians, the accompanist's role is not simply to follow the young soloist, but rather to encourage habits of accurate rhythm, steady tempo, while introducing musical ideas. In a sense, this is the hardest of all classical music accompaniment problems, since the accompanist must be expected to know *more* than the soloist, thus dictating when the accompanist must lead and when to follow. A coarse approximation to this accompanist role is to provide a rather rigid accompaniment that is not overly responsive to the soloist's interpretation (or errors); there are several commercial programs that take this approach. The notion of a pedagogical music system — one that follows and leads as appropriate — is largely undeveloped, possibly due to the difficulty of modeling the objectives. However, we see this area as fertile for lasting research contributions and hope that we, and others, will be able to contribute to this cause.

An entirely different scenario deals with music that evolves largely without a sense of rhythmic flow, such as in some compositions of Penderecki, Xenakis, Boulez, Cage, and Stockhausen, to name only a few. Such music is often notated in terms of seconds, rather than beats or measures, to emphasize the irrelevance of traditional pulse to the music's agenda. For works of this type involving soloist and accompaniment, the score indicates points of synchronicity, or time relations, between various anchor points in the solo

and accompaniment parts. Due to the lack of predictability of such music, a natural accompaniment approach is simply to wait until various solo events are detected, and then to *respond* to these events. This is the approach taken by the IRCAM score follower, with considerable success in a variety of pieces of this type [6],[7].

The third scenario, which includes our system [5],[8], treats works for soloist and accompaniment having a continuing musical pulse, including the overwhelming majority of "common practice" art music. This music is the primary focus of most of our performance-oriented music students, and is the music where our accompaniment system is most at home. Music containing a regular, though not rigid, pulse requires close synchronization between the solo and accompanying parts, as the overall result suffers greatly as this synchrony degrades. We will argue that this music cannot be performed effectively with the purely "responsive" approach as discussed above.

Our system is known as omitted (MPO) due to its alleged improvement on the play-along accompaniment records from "Music Minus One" that inspired our work. We have been collaborating for several years with faculty and students in the JSoM on this traditional kind of concerto setting, in an ongoing effort to improve the performance of our system. What follows contains a description of some of these improvements not discussed elsewhere, as well as a number of illuminating examples and demonstrations. We will also discuss strengths and weakness of our rhythm model, while sketching possible improvements. We conclude with a presentation of our accompaniment system in new music, focusing on works by omitted for review, specifically written for our system.

2 OVERVIEW OF MUSIC PLUS ONE

2.1 Score Following

Score following is the task of computing an ongoing alignment between a symbolic music score and an audio performance of the score, as the audio data accumulates. Also known as on-line alignment, the problem is more difficult that its off-line cousin, since an on-line algorithm cannot consider future audio data in determining the times of audio events. Thus, one of the principal challenges of on-line alignment is the tradeoff between accuracy —- reporting the correct times of note events — and latency — the lag in time between the reporting time and and estimated note event time. As with all of the accompaniment systems discussed above, score following plays a crucial role in MPO. [9] gives a nice annotated bibliography of the many contributions to score following.

Our approach to score following is based on a hidden Markov model and is described in [10]. Perhaps one of the main virtues of the HMM-based score follower is the grounding it gives to navigating the accuracy-latency tradeoff. One of the worst things a score follower can do is report events before they have occurred. In addition to the sheer impossibility of producing accurate estimates in this case, the musical result often involves the accompanist arriving at a point of coincidence before the soloist does. When the accompanist "steps on" the soloist in this manner, the soloist must struggle to regain control of the performance, perhaps feeling desperate and irrelevant in the process. Since the consequences of false positives are so great, the score follower must be reasonably certain that a note event has already occurred before reporting its location. Through the probabilistic nature of the HMM, one can compute the *probability* that the currently pending note has passed. Once this has occurred, our score follower looks backward in time to find the most likely onset position for the note.

We will omit a detailed discussion of the innards of our score follower here, and content ourselves with a simple, obvious, and crucial observation: Before an audio event can be detected it must have sounded for some brief period of time. Thus any score follower must necessarily deliver its observations with latency. That is, while a note onset time may estimated correctly, the reporting of this time must come after the event has occurred.

This observation has important consequences for the musical accompaniment system: If coordination is to be achieved in a "responsive" way — by waiting until a solo event is detected and then playing the corresponding accompaniment note, the system will always be late. In theory, one may be able to construct a score follower whose latency is musically insignificant. However, this has not been possible in our experience with such latencies usually in the 60-90 ms. range. If all coincident accompaniment notes lag this far behind, the result is musically fatal.

Instead, we accept as a basic tenet that detection latency will be musically significant and base our coordination of parts on *prediction* rather than response. Thus, central to our approach is the recognition that score following alone is not enough to produce good musical accompaniment. In addition we need a means of predicting future musical events and scheduling them accordingly. In contrast, the IRCAM system's approach is responsive, playing events in direct response to observations of solo events. This system has been quite successful in music that does not have a sense of ongoing pulse — the IRCAM system was developed with this kind of music in mind. However, the extension of this work to other musical styles including the overwhelming majority of common practice art music and popular music, seems problematic. In contrast, with minor adaptations our approach is equally at home in pulseless music.

A video demonstrating our score following ability can be seen at `http://www.music.informatics.indiana.edu/papers/smc09`. In this video the rather eccentric performer ornaments wildly, makes extreme tempo changes,

plays wrong notes, and even repeats a measure, thus demonstrating the robustness of the system.

2.2 Modeling Musical Timing

As discussed above, our approach to accompaniment relies on the prediction of *future* musical events. We present here the model serving as the backbone for this process. We begin with three important traits we believe such a model must have.

1. Since our accompaniment must be constructed in real time, the computational demand of our model must be feasible in real time.

2. We anticipate training our prediction algorithm using a sequence of rehearsals in which the solo player demonstrates her interpretation, with all its variability. In order to benefit from these rehearsals our model must be automatically trainable. Thus, rehearsal will allow our system to more accurately anticipate the way future musical timing will unfold. This is certainly one of the objectives of human rehearsal, as well.

3. If our rehearsals are to be successful in guiding the system toward the desired musical end, the system must "sightread" (perform without rehearsal) reasonably well. Otherwise, the player will become distracted by the poor ensemble and not be able to play her part consistently with her convictions. Thus our model must be constructed around widely applicable musical assumptions, so it can perform reasonably well "out of the box."

Our model is expressed in terms of two sequences, $\{t_n\}$ and $\{s_n\}$ where t_n is the time, in seconds, at which the nth note begins and s_n is the tempo, in seconds per beat, for the nth note. The model is then

$$s_{n+1} = s_n + \sigma_n \quad (1)$$

$$t_{t+1} = t_n + l_n s_n + \tau_n \quad (2)$$

where l_n is the length of the nth note, in beats. With the $\{\sigma_n\}$ and $\{\tau_n\}$ variables set to 0, this model gives a literal and robotic musical performance. The introduction of these variables allow time-varying tempo through the σ's and elongation or compression of note lengths with the τ's. To complete the model we assume that

$$\begin{pmatrix} \sigma_n \\ \tau_n \end{pmatrix} \sim N(\mu_n, \Sigma_n)$$

where $N(\mu, \Sigma)$ denotes a joint normal distribution with mean μ and covariance Σ. Thus the $\{\mu_n\}$ vectors represent the *tendencies* of the performance — where the player tends to speed up ($\sigma_n < 0$), slow down ($\sigma_n > 0$), and stretch ($\tau_n > 0$), while the $\{\Sigma_n\}$ matrices capture the repeatability of these tendencies.

If the actual note observations generated by our score follower, $\{y_n\}$ are viewed as imperfect estimates of the *true* onset times,

$$y_n = t_n + \epsilon_n \quad (3)$$

where $\epsilon_n \sim N(0, \rho^2)$, and all of the $\{\sigma_n, \tau_n, \epsilon_n\}$ variables are modeled as *independent*, then the model is seen as a straightforward example of the Kalman filter. In this context, all of our desired traits are satisfied. We predict future evolution by first computing our knowledge of the current state given our observations, $p(s_n, t_n | y_1, \ldots, y_n)$. From this information we can predict future note onset times by applying our basic model to our current belief, thus allowing the system to sightread. Furthermore, using standard ideas from the Bayesian network literature, we can perform maximum likelihood estimation on the $\{\mu_n, \Sigma_n\}$ parameters, thus training our model from actual rehearsal data. Finally, the computational burden of these calculations is modest, at most, easily suiting the approach for real time.

Our system is concerned only with the scheduling of the currently pending accompaniment note. Every time new information becomes available, either in the form of a played accompaniment note or a detected solo note, we have new information about the pending note. Thus we reestimate the current state, predict the accompaniment location, and reschedule the note accordingly. If we consider the common situation involving a run of solo notes culminating in a point of coincidence between solo and accompaniment parts, we see that this time of coincidence will be rescheduled many times before its scheduled time finally occurs and the note is played. In this way, our system makes use of all information currently available, continually modifying its view of musical timing until it must finally act.

We have created a video to demonstrate this process, available at the aforementioned website. The video shows the estimated solo times from our score follower appearing as green marks on a spectrogram. Predictions of our accompaniment system are shown as analogous red marks. One can see the pending accompaniment event "jiggling" as new solo notes are estimated, until finally the time currently predicted time passes.

At this point there seems to be so much "good news" that one is loathe to make criticisms. However, long experience with this model in action has demonstrated a number of deficiencies, mostly perceived as a kind of musical naivete. We will discuss these and pose possible improvements in a later section.

2.3 Computational Approach

Our program consists of about 100,000 lines of C code with the graphical interface written in C++. The score follower

is implemented as a *thread* which continually polls to see if a new frame audio data is ready, with about 31 audio frames per second. When a new frame is available, the thread runs an iteration of the HMM forward algorithm. If the forward algorithm detects that the pending solo note has passed, the most likely onset frame is computed through the forward-backward algorithm, using all currently-available audio data. This most likely time is then modeled as a noisy estimate of the *true* solo time, (Eqn. 3) and the pending accompaniment note is rescheduled using the Kalman filter model.

Our system can create the audio output using either MIDI, or resynthesizing the output audio from an accompaniment-only recording. This latter method is our preferred approach for traditional common practice art music, since it preserves much of the tonal quality and some of the performance intent of the original recording. We often use the Music Minus One recordings for this purpose. When using a recording, we resynthesize the audio using phase vocoding, thus allowing time warping in the original recording without any change of pitch.

A separate high-priority thread handles this audio output — while there is no great danger in delaying the processing of audio input, a delay in audio output can result in a "drop-out" with an associated click or gap in the audio output. This thread is time critical since we create the audio at the last possible moment allowing it to be influenced by the most current information from the audio analysis thread. Typically we buffer about .064 seconds of outgoing audio. This thread constructs each frame of audio according to the current vocoding "play rate," computed from the prediction model as the rate needed to arrive and the pending event at the predicted time.

While originally written for the Linux operating system, our preferred home, in recent years we have ported the system to Windows. Ideology aside, the target community of this work is actual practicing "classical" musicians, more familiar with Windows. No special-purpose hardware is needed to run the system.

3 MODELING THE ORCHESTRA'S CONTRIBUTION TO THE AUDIO

One of the often-touted virtues of the HMM is its trainability. That is, an HMM can use representative data to automatically improve its transition and output models, perhaps resulting in better performance. Though we continue to place faith in this trainable aspect of the HMM we have replaced a fully trained output model with a different model that performs significantly better, even without training.

This model computes the likelihood of an audio magnitude spectrum $q = (q_1, \ldots, q_K)$ given an assumption about the note or notes sounding in the solo part. In doing so, we construct a probability template $p = p_1, \ldots, p_K$ for the note or notes that may be sounding at a particular time. For a single note we have modeled p as a mixture of Gaussians centered at the harmonic frequencies of the note with decreasing mixture weights as harmonic number increases:

$$p_k = \sum_{h=1}^{H} w_h N(k; hf_0, (hf_0)^2\rho^2) \tag{4}$$

where $\sum_h w_h = 1$, f_0 is the fundamental frequency of the note, and $N(k; \mu, \sigma^2)$ is a discrete approximation to the normal density function. With this probability model in place, we view the actual audio magnitude spectrum as a random sample from the probability model. That is, we regard q_k as the number of observations at frequency k — q_k must be discretized for this to make sense. Then we have

$$p(q|p) = c(q) \prod_k p_k^{q_k}$$

where $c(q)$ is the multinomial constant. In the event that we are following a polyphonic instrument, we simply model p in Eqn. 4 with an additional sum over the collection of currently-sounding solo notes. This model has worked well in practice in a wide variety of situations and can be extended in some interesting ways, as follows.

The model above may describe reasonably well the audio signal that comes from the soloist, for purposes of note discrimination. However, our microphone will receive both this solo audio as well as the audio generated by our accompaniment system. When the accompaniment audio contains components that are confused with the solo audio, this can lead to the highly undesirable possibility of the accompaniment system *following itself* — in essence, chasing its own shadow. To a certain degree, the likelihood of this outcome can be diminished by "turning off" the score follower when the soloist should not be playing. We do this. However, there is still significant potential for shadow-chasing since the pitch content of the solo and accompaniment parts is often similar.

Our solution to this difficulty is to directly model the contribution of the accompaniment to the incoming audio signal we process. Since we *know* what the orchestra is playing, we add a component of this contribution to our probability model. More explicitly, if p_s is the solo template described above, and p_o is the known contribution of the orchestra to the currently analyzed audio frame, we create a probability model for the observed magnitude spectrum q by $p = \lambda p_s + (1 - \lambda)p_o$. This is the actual p we use in evaluating the data likelihood.

This addition creates significantly better results in many situations. The surprising difficulty in actually implementing the approach, however, is that there seems to be only weak agreement between the audio that our system plays and the accompaniment audio the comes in from the microphone. We can improve our model of p_o by various averaging tricks, thus modeling the room acoustics to some degree.

Doing so leads to a p_o estimate that seems to largely eliminate the undesirable shadow-chasing.

4 BETTER MODELING OF MUSICAL TIMING

We have already discussed the strengths of the musical timing model of Eqns. 1-2, however, it would be disingenuous to claim there are no weaknesses. Clearly our model must allow for a range of possible musical performances, since we know we will encounter variation in practice. Since we do not know the nature of this variation, we have overparametrized the model, allowing for way too much flexibility — and perhaps not the right kind. Surely the player will not make a change to the tempo *and* apply tempo-independent note length variation on every note. However, our model allows such a performance (and accommodates it reasonably well). We propose a couple of possible variations on the basic rhythm model here.

Our first observation concerns the $\{\tau_n\}$ variables of the model, which represent changes in note length not naturally expressed through tempo. The prime musical example would be the *agogic* accent, in which one stresses a note by lengthening it, though keeping the same basic tempo in subsequent notes. This is a common expressive device in playing passages of fast running fast notes, to highlight important metric positions, harmonic changes, dissonances, etc. While this example of note lengthening is familiar in a variety of musical styles, we don't believe the same holds for *shortening* of note length. Of course, there are musical examples where the conceptual rhythm may differ from that explicit notation, such as the double-dotting of a French overture, or the swing of jazz. But these are examples where "stolen" time is given back elsewhere, unlike the case of $\tau_n < 0$. We expect that the musical plausibility of our model is improved by removing this possibility.

Our second observation is that tempo changes and note length variation introduced by the player is sparse — most notes are rendered without any such deviation, while it may not be musically meaningful to have both agogic accent and tempo change in the same position. Phrased in terms of our model, most of the $\{\sigma_n, \tau_n\}$ variables are 0 and we should not allow $\sigma_n \neq 0$ *and* $\tau_n \neq 0$ for fixed n.

We propose the following model to capture these notions. We let $x_1, x_2, \ldots$ be a hidden discrete process where n continues to index the notes of the piece. We assume $x_n \in \{1, 2, 3, 4\}$, with the following interpretations:

$$\begin{aligned} x_n = 1 &\iff \sigma_n = \tau_n = 0 \\ x_n = 2 &\iff \tau_n = 0 \\ x_n = 3 &\iff \sigma_n = 0, \tau_n \sim N(\mu_3, \rho_3^2) \\ x_n = 4 &\iff \sigma_n = 0, \tau_n \sim N(\mu_4, \rho_4^2) \end{aligned}$$

That is,

1. When $x_n = 1$ we arrive at note n exactly in tempo.
2. When $x_n = 2$ the tempo may change between notes $n-1$ and n, but there is no additional note length variation.
3. When $x_n = 3$ we have have an agogic accent and no tempo variation. This is the case of a small agogic accent where the parameters $\mu_3 > 0$ and ρ_3^2 are chosen to ensure that a negative value is highly unlikely.
4. When $x_n = 4$ we have have a similar situation, but now account for the longer agogic accent. Thus $\mu_4 > \mu_3$ with ρ_4^2 also chosen to make negative values of τ_n rare.

Of course these 4 cases are not equally likely, thus we model the probabilities of $p(x_n = i)$ to reflect that $i = 1$ is, *a priori*, the most likely, with reasonable choices for the other 3 cases. It may even be reasonable to model the $x_1, x_2, \ldots,$ process as a Markov chain allowing for some small degree of memory in the choice of expressive actions.

The model is now a Switching Kalman filter [11]. For the Switching Kalman filter, the exact computation of the filtered distribution: $p(x_n, s_n, t_n | y_1, \ldots, y_n)$ is not tractable due to the large number of paths $x_1, \ldots, x_n$ that must be marginalized over, in accounting for all of the ways we can arrive at state (x_n, s_n, t_n). However, there are numerous ways to approximate this calculation, using various approximation schemes. In addition, such models are also amenable to automatic training using ideas analogous to those employed with Kalman Filters and HMMs. Here we train the $p(\sigma_n, \tau_n)$ parameters, as before, and additionally train the $p(x_n)$ probabilities. Thus we learn the *qualitative* behavior of the soloist through the $p(x_n)$ probabilities, which tell us where various kinds of actions are likely to occur, as well as the quantitative description learned through the $p(\sigma_n, \tau_n)$ parameters. Experiments are currently underway with such a model.

5 NEW MUSIC WITH ACCOMPANIMENT SYSTEM

Our work with accompaniment systems has mostly focused on common practice music for soloist and orchestra, however, we believe the accompaniment system is by no means limited to this domain. There has been a long tradition of compositions for live soloist and accompanying electronica, with many possible techniques for coordinating parts. In some of these, the live player is completely responsible for synchronization, by either following a tape or playing along with a click track. In others, a human plays the role of the "conductor," cueing electronic or computer parts at the appropriate times. There have also been some examples in which the computer genuinely *follows* the live player, but

with some of the best results in music not relying on regular pulse, such as with IRCAM's score follower mentioned above. We believe that the notion of pulse is in no way limited to common practice music, as exemplified by the vast collection of contemporary music that employs metered rhythm. Thus we believe our accompaniment system may create possibilities for new music, perhaps not playable by any other means, whose composition is of genuine interest to living composers.

Recently we have recorded two such new works for oboe and computer-controlled pianos written specifically for our accompaniment system by Swiss composer name omitted for review: *Mist Covered Mountains* and *Winter*. While the pieces use traditional rhythmic notation and sometimes have a highly rhythmic feel, they require a level of pianistic virtuosity and ease with complex polyrhythms posing nearly superhuman demands on the pianist. This is fitting, since the piano part(s) were not intended to be played by humans.

One of the main challenges for the oboist is in understanding the rhythmic relationship between the parts; the score notates all rhythm precisely, though there is an aleatoric feel to large sections. While a good deal of score study was necessary to accomplish this, quite a bit of rote memorization was also necessary, accomplished through regular listening over a period of several months. Perhaps the author's original understanding of this music was something like the young student's knowledge of the "Pledge of allegiance" — knowing the sequence of syllables, but perhaps not the meaning of the words. However, the music began to make sense after passing through this stage. The accompaniment system was a significant aid in *learning* these pieces, since it came to our rehearsals already understanding the complex rhythmic relations and reinforced these through repetition and automatic adaptation to the soloist's errors.

The music was recorded in a studio, recording the live oboe while listening to a MIDI performance of the pianos through headphones, as controlled by the accompaniment system. The MIDI piano performance was as captured and later used to control a Bösendorfer reproducing piano. The resulting piano audio was then mixed with the original oboe. Recordings of sections of these pieces are available at the web page mentioned earlier. Though the merit of these pieces does not lie in their reliance on new technology, it seems nearly impossible to perform these pieces with anything other than an accompaniment system.

6 REFERENCES

[1] Lippe C., "Real-time Interaction Among Composers, Performers, and Computer Systems", *Information Processing Society of Japan, SIG Notes*, Vol. 2002, Num. 123, pp 1-6, 2002.

[2] Rowe R., *Interactive Music Systems*, MIT Press, Cambridge, MA, 1993.

[3] Dannenberg R.and Mont-Reynaud B., " Following an Improvisation in Real Time" *Proc. of the 1987 International Computer Music Conference*, pp. 241-248, 1987.

[4] R. Dannenberg, H. Mukaino "New Techniques for Enhanced Quality of Computer Accompaniment" *Proc. of the International Computer Music Conference* 243–249, Köln, 1988.

[5] omitted for review, "A Bayesian Network for Real-Time Musical Accompaniment" *Advances in Neural Information Processing Systems, NIPS 14*, MIT Press, 2002.

[6] Cont A., Schwarz D., Schnell N., "Training IRCAM's score follower", *Proceedings of IEEE Int. Conf. on Acoustics Speech and Signal Processing (ICASSP)*, Philadelphia, USA 2005.

[7] Cont A., Schwarz D., Schnell N., "From Boulez to Ballads: training IRCAM's score follower", *Proceedings of the Int. Computer Music Conf. (ICMC)*, Barcelona, Spain, 2005.

[8] omitted for review, "Music Plus One: a System for Expressive and Flexible Musical Accompaniment" *Proc. Int. Comp. Music Conf., 2001* Havana, Cuba, 2001.

[9] Schwarz D., "Score following commented bibliography", *http://www.ircam.fr/equipes/temps-reel/suivi/bibliography.html*, 2003.

[10] omitted for review, " Segmentation of Acoustic Musical Signals Using Hidden Markov Models", *IEEE Trans. on Pattern Analysis and Machine Intelligence* 21(4), 1999.

[11] Bar-Shalom Y., *Estimation and Tracking: Principles, Techniques, and Software*, Artech House, Boston, Massachusetts, 1993.