Synthetic Rehearsal: Training the Synthetic Performer

Barry Vercoe & Miller Puckette
Experimental Music Studio
Media Lab, M.I.T.

1. ABSTRACT

Computer tracking of live instruments aims to understand the dynamics of live performance well enough to replace any member of an ensemble by a synthetic performer (computer model) so that the others cannot tell the difference. Modelling the performer experience has 3 major parts: *listen* (extract temporal and other cues from the other players), *perform* (organize a synchronized and suitably matched performance), and *learn* (remember enough of each experience to benefit future encounters).

While encouraging progress has been made on parts 1 and 2 (Vercoe, 1984), methods to date have relied exclusively on highly-sensitive live tracking. There has been no learning or training capacity to allow the benefits of previous experience. The synthetic perfórmer is essentially "sight reading every time" on the concert stage.

This paper outlines a method of integrating past and present experience into new strategies of rehearsal and improved performance. During a performance run, past or learned information is used to sensitize the perceptual and cognitive components, giving them a bias towards certain expected live input behavior. Eventual pitch contour matching and durational best fit are thus made easier and more robust. As before, rhythmic elasticity, and both pitch and rhythmic fault tolerance, are a necessary part of a practical performance system.

1. BACKGROUND AND OVERVIEW

During 1983-84, research was conducted jointly at the *Institut de Recherche et Coordination Acoustique/Musique* in Paris, France, by Barry Vercoe (MIT) and Lawrence Beauregard (flutist of the IRCAM Ensemble Intercontemporaine). The objective was to understand the dynamics of live ensemble performance well enough to replace any member of a group by a **synthetic performer** (i.e. a computer model) so that that the remaining live members could not tell the difference. This aimed to break clear of the "music-minus-one" syndrome that has characterized tape and instrument pieces of the past, and to recognize the machine's potential not as an amplifier of low-level switches and keys, but as an intelligent and musically informed collaborator in live performance.

The research was carried out using standard repertoire (flute sonatas by Handel and W.F.Bach) in which the flute part was played live and the responding harpsichord part was "performed" by a strictly synthetic accompanist. The motivation for the work was an IRCAM commission for Vercoe to compose a work, *Synapse*, for Larry Beauregard and the 4X real-time audio processor. Proof of how well the computer can behave as chamber music player in both traditional and contemporary contexts was then given in public demonstration at the 1984 International Computer Music Conference (Vercoe, 1984).

The live demonstration routinely showed a facile ability to track and remain in sync with wildly changing tempos. It also showed an ability to deal with heavy doses of live performer errors, sloppy rhythms and general distortion that would likely throw off a live accompanist. However, in these early implementations there was no data retention from one performance run to another: there was no performance "memory", and no facility for the synthetic performer to learn from past experience. The synthetic performer was essentially *sight reading* on the concert stage every time.

The aim of subsequent research has been to conceive and model a working representation of music **rehearsal** and music **learning**. This has meant rethinking the music abstractions so as to more adequately represent strongly structured scores, and to more easily model those elements of cognition and learning that depend heavily on structure. However, since many contemporary scores are only weakly structured (e.g. unmetered, or multi-branching with free decisions), it has also meant development of score following and learning methods that are not necessarily dependent on structure. This has led to a flexible score-following method in which dependence on temporal structure and on previous rehearsal information can be parametrically controlled.

2. THE ANATOMY OF PERFORMANCE

2.1 Organizing one's own Contribution

Live music performance comprises three major parts:
- Perception and cognition of external controls.
- Organizing one's own performance.
- Learning from the experience.

Although computer synthesis has traditionally specialized in organizing a performance, it provides a very poor model. Live music performance is decidedly more procedural, and many of its aspects remain unevaluated until the last possible moment in time. The chronological value of a 2-beat duration, for instance, is determinable only near the end of its performance. This means that the processes which control time-sensitive parameters (e.g. crescendo over several beats) must remain active throughout, and are susceptible to outside influences (changing tempo, etc.) at any time. In order to construct a synthetic performer, we must first understand the anatomical principles that drive real performers.

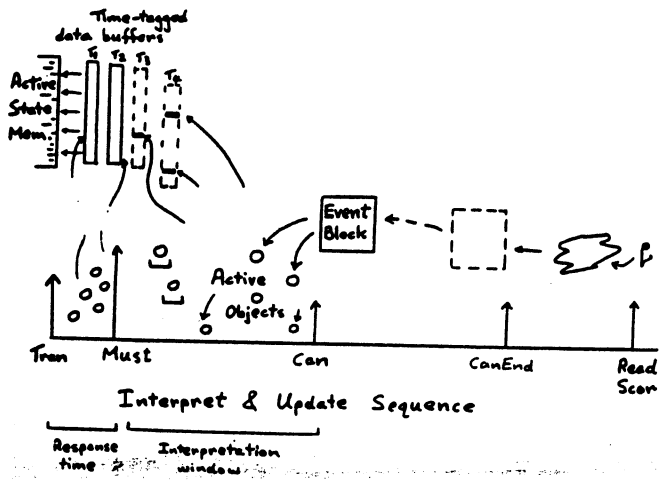

Figure 1.

The anatomical functions of a performance are depicted in Fig. 1. A musical note is initially understood as some future event, its distance away measurable only in beats. As the moment of performance draws near and the beat-defined event is drawn into active short-term memory, it passes a point where it can safely be given chronological definition ("can"). At this point the performer sets in motion the motor reflexes that commit him to the note's performance. The decision to perform may be delayed slightly, but not past a certain critical point ("must") which is the last chance to initiate the physical gesture that will deliver the note on time. The values of "can" and "must" in relation to the time of actual sound will depend respectively on the responsive sensitivity of the intended performance and on the physical inertia of the instrument and player concerned. Values of two-tenths and one-tenth of a second from actual sound are reasonable for a typical music situation.

The above scheme has been strictly represented in our synthetic performer system. The **target instrument** used in our studies has been either a digital model residing on a powerful audio processor (the IRCAM 4X), or a *Yamaha DX7* synthesizer under MIDI control. In both cases, data is prepared in advance by a host processor and left in time-tagged buffers for automatic transfer to active state memory. The **performer** is modelled as a set of ***control processes*** running on the host processor. These sense external conditions and generate instrument update data values. Whenever an event crosses the "can" boundary, it bursts into life by spawning a host of some 20 active objects representing the multiple facets of a single sensitively performed note. These active objects compete for computing resources of a single CPU. Since they must do so in the "can-must" time span, the system is sensitive to the notion of performer overload.

2.2 Listen-Perform Synchronization

Computer performance of music can easily demonstrate that strictly synchronous behavior lacks much of the information we routinely seek from live performance. It is as if the musical score acts as a carrier signal for other things we prefer to process. Much of this information derives from discrepancies between individual players. The degree of synchronization will vary as the live performer shifts the focus of his attention about. However, it is meaningful to ask: 'what are the tolerances involved, and what is the peak rate at which a skilled performer adjusts to incoming data?' The adjustment has an automatic throttle: upon sensing discrepancy, a performer will seek an aesthetic way of adapting, so as to preserve the integrity of his own line. How independent he can remain depends on what the score warrants and supports. We have here, in effect, a **loosely-coupled system** of performance and control, whose parameters depend on the topology of the score involved.

Listen-perform synchronization is based on two perceptions: currently observed tempo, and currently observed score position. Given an accurate observed tempo, and a comparison of the relative score positions, it is possible to determine by how much the synthetic performer should speed up or slow down. However, rhythmic time-shifts (either stylistic or in error) can distort both of these observations, and our initial synthetic performer implementation devoted considerable CPU attention to dealing with these aberrations. Many aberrant conditions can be described in the form of if-then productions, and much of our earlier success was due to this strategy. However, other quirks of interpretation can turn out to be quite localized and unsuited to such representation, even though they would likely occur at the same place in every performance of a given piece. These special moments catch a memoryless synthetic performer by surprise every time. Further progress on how skilled performers relate to one another is almost impossible without better models of score following, rehearsal and learning.

3. SCORE FOLLOWING

The problem of following a live player is complicated by the fact that neither the performer nor the sensors on his instrument are free from errors. As we move to the ultimately desirable situation in which we use only acoustic information from the performer, the frequency and magnitude of errors in the score tracker's input will increase. The success of the synthetic performer will be limited by our ability to reject errors while remaining sensitive to the useful information available.

Errors, both from the performer and from sensors, can be divided into qualitative errors (missing, extra, or wrong notes) and quantitative ones (early and late notes.) To reject the former it is sufficient to identify them; then we can ignore notes which do not correspond to the score, and use the timing information which is present even in wrong notes. Our best protection against quantitative error is to reduce it through averaging.

Our main desire is to extract reliable estimates for tempo and current beat, updating them as new events arrive. This amounts to finding our location in the score as a function φ of time, which best fits the observed data. To do so we need to define a measure of how closely a theory about the current tempo and score position fits the incoming data.

Suppose that a part of the score is given by $\{(s_1,x_1), \cdots (s_n,x_n)\}$, where (s,x) denotes the event, "object x arrived at time s." Suppose that we detect a performer as playing the sequence, $\{(t_1,y_1), \cdots (t_m,y_m)\}$. We want a *theory* of the form, $\{(a_i,b_i)\}$, $a_1<a_2<\cdots$, $b_1<b_2<\cdots$, which is taken to mean, "The event (t_{b_i},y_{b_i}) played by the performer corresponds to the event (s_{a_i},x_{a_i}) in the score."

We assign to such a theory a *cost*, which measures its deviation from what we imagine to be a good performance. First we charge for combinatorial errors: p_m for each note in the score which is missing from the performance, p_e for each extra note in the performance, and p_w for each wrong note (i.e. whenever $x_{a_i} \neq y_{b_i}$.) To this we add the cost associated to metrical errors. We choose nonnegative weights $\{w_i\}$ and take the least-squares fit of a line through the points (s_i,t_i), weighted by w_i. This line, $t=ps+q$ say, represents an assumption of a locally constant tempo of p real seconds per score second. The metrical cost of the theory is a constant p_t times the deviation of the (s_i,t_i) from the line $t=ps+q$; we measure this deviation as $\sqrt{\frac{\lambda_1}{\lambda_2}}$, where λ_1,λ_2 are the minor and major moments of $\{s_i,t_i,w_i\}$. We find that we will want to drop some points out of this linear fit, which means that we are ignoring the times at which the performer plays some notes. To do this to a note which is right we charge another constant p_{tr} and if it is wrong (in which case we are more willing to ignore its timing) we charge p_{tw}. The cost of the theory is the sum of all the above charges.

The problem of tracking the performance is that of finding the least costly theory. We then use the values of p and q associated to that theory either to predict what the performer will do next or to schedule the performance of an accompaniment.

Since tempo is not a global constant but is only nearly constant over short stretches of time, we actually wish to form a new theory associated to each note the performer plays. We choose the weights w_i to emphasize recent events over ones further in the past, thus keeping our tempo measurements local. We thus constantly update our estimates for tempo and current beat on the basis of fresh data.

Finding the absolute optimum theory for a note in a score of nontrivial size is probably intractable, but we have found an algorithm which works well in practice. For each pair (i,j) we find the best theory we can fitting the first i events in the performance to the first j events in the score. We choose the cheapest of four theories. First, the pair (i,j) might be a point of true correspondence between the score and the performance. The cost of this theory is the combinatorial cost of the best $(i-1,j-1)$ theory plus p_w if the current note is wrong plus the cost of a linear fit of (i,j) with the other correspondences in the $(i-1,j-1)$ theory. Second, (i,j) might be a correspondence whose time we wish to ignore; in this case we leave it out of the linear fit but add a charge of p_{tr} or p_{tw} depending on whether the note is right or wrong. Third, the ith note of the performance might be an extra note; hence we add p_e to the cost of our best $(i-1,j)$ theory; and finally the jth note of the score could be omitted, costing p_m plus the cost of the best $(i,j-1)$ theory.

This algorithm is a fusion of the purely combinatorial one of Dannenberg (1984) with our own methods for following both pitches and rhythms (Vercoe, 1984). As such, it is able to combine hints from the two regimes to keep it in the closest possible synchrony with the performer. Via the mechanism of the parameters p_m, p_e, p_w, p_t and p_l, we may tune the algorithm to the specific tendencies of a performer/sensor pair. For example, if it is known that we almost never get extra notes, we will improve performance by increasing p_e, for the algorithm will rarely lie that a note is missing then. Our current setting is $p_e=p_m=p_w=1$, $p_{tr}=.4$, $p_{tw}=.15$, and $p_l=10$.

In this way we can follow a performer. That is to say, if the performer plays his score without too many mistakes and without rapidly changing his tempo, we can approximate his instantaneous tempo and score position at any time. But we will still be sensitive to two elements of the performance which do not convey information about tempo and which we should be able to reject. First, a human performer realizes music in part by giving it a rhythmic microstructure; we should not try to follow this as a changing tempo from note to note but should learn to expect it and account for it when we track the performer. Second, there is the truly nonrepeatable microstructure. Even though we believe that this is musically essential we do not wish it to fool our following. The natural way to do this is to use memory of past performances constructively.

4. REHEARSAL, MEMORY AND LEARNING

4.1 Storing the Experience

Without information from previous performances, the Synthetic Performer's view of another performer's part will remain a prescribed contextual record. The score will be only minimally structured, conveying only the barest of syntactic features and nothing of performance or semantic interpretation. Every live performance heard, then, is at odds with this score, and requires the full services of facile though transient interpreters to maintain a reliable sense of score position. There is no learning from experience. The condition is typified by consecutive performance runs that are identically surprised by the same idiosyncratic input.

The first step towards admitting rehearsal as a determinant of subsequent expectations is to save away as much rehearsal data as might be useful. In the temporal domain, the information is of two major kinds: rhythmic aberration, and tempo warping. The two are not distinct, and one can easily be mistaken for the other in small localized contexts. The perceptual trick is to constantly gauge the effective tempo and current score position, then to regard the remaining distortions as rhythmic. As each incoming note is recognized in the score, the beat fraction by which the event is earlier or later than expected is written into its performance record. This amount is strictly a difference index with respect to the currently believed score position. There is no attempt to amend it on the fly, even in the face of changing opinion.

4.2 Post-Performance Memory Massaging

The effect of rehearsal on musical memory is to permit construction of new semantic concepts that will help delineate the score in its next performance. We can use the rehearsal data to improve future tempo detection, for example, by gathering statistics about how the performance of each note in the score tends to relate to the extracted tempo. After each performance, we take the event list of performer action times and incorporate this experience into the score by doing an unrehearsed, non-real-time tempo analysis of the entire event list, thus giving a score position as a function φ of real time. Since we are not calculating this in real time, we can look into the future to get a more accurate estimate of φ. For each event (s_i,x_i) in the score we collect the sample mean m_i and standard deviation σ_i of the measured discrepancies in score position between the observed note and the expected one, or $s_i-\varphi(t_j)$, where (i,j) is a correspondence between the real score and the observed one.

The first correction we now make is for m_i. We do tempo matching not to the events (s_i,x_i) but to the mean-corrected events (s_i-m_i,x_i). In this way we prevent that part of the microstructure of performance which repeats from rehearsal to rehearsal from skewing the tempo observations.

We use the sample standard deviations σ_i to guess the relative strength of the correlation between the time the note occurs and the actual beat it lands on. When we use the event (s_i,x_i) as part of a tempo determination we weight it by σ_i^{-1}, since that yields weighted averages of minimum variance.

In this way, we have used rehearsal memory to aid the score follower in precisely the language it can best understand. Since the follower sees its problem as one of finding a best fit between theory and observation, the most useful information we can give it is an estimate of how much on the average the data fails to meet the theory and at what points the data is more or less important to match. Similar memory massaging strategies can be used for other information tracks.

5. CONCLUSION

We have described a method of tracking live performers in such as way that they may be accompanied by synthetic performers that learn from rehearsals. Continual semantic reinterpretation of a piece under rehearsal is a complex, data intensive task, and a fully representative interpretation requires numerous rehearsals before the gathered statistics reliably anticipate live performer behavior. When that point is reached, however, the synthetic performer can be described as a well-rehearsed musical colleague, capable of robust yet sensitive collaborative performance.

Eventually it should be possible to develop a synthetic performer that would not require priming with a written score of what initially to listen for. Chamber music players typically perform from single part-books, building their sense of the full score strictly from the experience of rehearsal. In that this appears to be a prime route by which those players inform their overall performance, we would eventually like to understand a little of how that works.

6. REFERENCES

Dannenberg, Roger (1984). Tracking a Live Performer. *ICMC Proceedings*, 1984.

Vercoe, Barry (1984). The Synthetic Performer in the Context of Live Performance. *ICMC Proceedings*, 1984.