Using Musical Knowledge to Extract Expressive Performance Information from Audio Recordings

Eric D. Scheirer
MIT Media Laboratory
E15-401C Cambridge, MA 02140
email: eds@media.mit.edu

December 14, 1995

Abstract

A computer system is described which performs polyphonic transcription of known solo piano music by using high-level musical information to guide a signal-processing system. This process, which we term *expressive performance extraction*, maps a digital audio representation of a musical performance to a MIDI representation of the same performance using the score of the music as a guide. Analysis of the accuracy of the system is presented, and its usefulness both as a tool for music-psychology researchers and as an example of a musical-knowledge-based signal-processing system is discussed.

1 Introduction

Traditionally, transcription systems (computer systems which can extract symbolic musical information from a digital-audio signal) have been built via signal processing from the bottom up. In this paper, we examine a method for performing a restricted form of transcription by using a high-level music-understanding system to inform and constrain a signal-processing algorithm.

The motivation for this work was to build a system which could extract performance parameters from digital audio recordings of known solo piano music accurately enough that they could be used for music-psychological analysis of expressively performed music. However, the system also stands as a example of a new general approach to the polyphonic transcription problem.

The parameters extracted are those which are controllable by the pianist: velocity (the force of the stroke), and attack and release timing. [8] suggests certain levels of timing accuracy in performance which can be understood as benchmarks for a system which is to extract note information accurately enough to allow the study of musical interpretation. For example, among expert pianists, the melody of a piece of music typically runs ahead of its accompaniment; for chords, where it is indicated that several notes are to be struck together, the melody note typically leads by anywhere from 10-15 ms to 50-75 ms, or even more, depending on the style of the music. Thus, if we are to be able to use such a system for examine timing relationships between melodies and harmony, it must be able to resolve timing differences between "simultaneous" notes to this level of accuracy or finer.

5 ms is generally taken as the threshold of perceptual difference (JND) for musical performance ([2]); if we wish to be able to reconstruct identical performances from the transcription data, the extracted timing data must be at least this accurate.

1.0.1 Score-based transcription; or, why cheating is good

It seems on the surface that using the score to aid transcription is "cheating", or worse, useless – what good is it to build a system which extracts information you already know? It is our contention that this is not the case; and, in fact, score-based transcription is an extremely useful restriction of the general transcription problem.

It is clear that the human music-cognition system is working with representations of music on many different levels which guide and shape the perception of a particular musical performance. Work such as Krumhansl's tonal hierarchy ([3]) and Narmour's multi-layered grouping rules ([5], [6]) show evidence for certain low- and mid-level cognitive representations for musical structure; and syntactic work such as Lerdahl and Jackendoffs' ([4]), while not as well-grounded experimentally, suggests a possible structure for higher levels of music cognition.

While the system described in this paper does not attempt to model the human music-cognition system per se (and, further, it is not at all clear how much transcription the human listener does, in the traditional sense of the word – see section 4.2, below), it seems to make sense to work towards multi-layered systems which treat musical information on a number of levels simultaneously. This idea is similar to those presented in Oppenheim and Nawab ([7]) regarding symbolic signal processing.

From this viewpoint, score-aided transcription can be seen as a step in the direction of building musical systems with layers of significance other than simply a signal-processing network. Systems built along the same lines with less restriction might be rule-based rather than score-based, or even attempt to model certain aspects of human music cognition. Such systems would then be able to deal with unknown as well as known music.

2 Architecture of the System

Figure 1 shows a schematic representation of the architecture of the current implementation of the computer program. [1]

Briefly, the structure is as follows: a initial score-processing pass determines predicted structural aspects of the music, such as which notes are struck in unison, which notes overlap, and so forth. In the main loop of the system, we do the following:

- Find releases and amplitudes for previously discovered onsets.
- Find the onset of the next pitch in the score.
- Re-examine the score, making new predictions about current local tempo in order to guess at the location in time of the next onset.

[1] All program code is currently in the Matlab matrix-processing language, and is available from the author via the Internet; email to `eds@media.mit.edu` with requests.

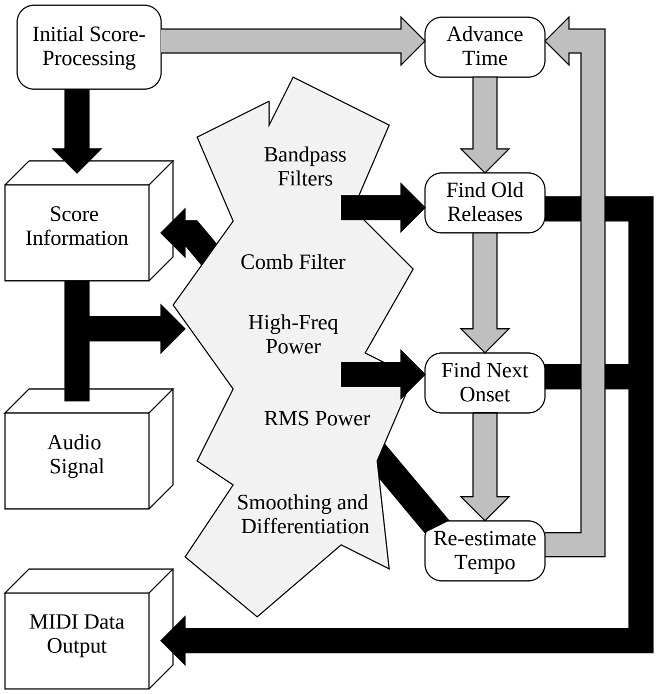

Figure 1: *Overview of system architecture, showing the main components (algorithms and data sources) of the system and how they interact.*

Once there are no more onsets left to locate, we locate the releases and measure the amplitudes of any unfinished notes, and write the data extracted from the audio file out as a MIDI (Musical Instrument Digital Interface) text file. It can be converted using standard utilities into a Standard Format MIDI file which can then be resynthesized using MIDI hardware or software.

It is important to note the relative simplicity of the signal-processing in the following descriptions of the algorithms. It is to be expected that more sophisticated filter techniques would lead to better results, but it is a mark in favor of the general attractiveness of the use of high-level information that good performance can be gained without them.

2.1 Onset Extraction

The onset extractor is the element of the system into which the most work has gone, and as a result is the most complex. It contains four main methods for filtering and parsing the signal, with parameters which are adjustable based on the information found in the signal and in the score-processing.

To find an onset, a time window within which the onset is likely to occur is passed from the tempo-estimator, and heuristics based on the score information are used to determine the type of digital signal processing which should be done, as follows.

If no other notes are struck at the same time, we can look for high-frequency energy (above 4000 Hz; this energy is noise from the hammer strike) in the signal during this time window, and an increase in the overall RMS power. If either of these occur in an area of positive derivative in the fundamental band of the pitch we're looking for, we can select the point of peak derivative of the high-frequency energy or RMS power as the onset time. This method leads to the highest accuracy (see the **Results** in the **Validation Experiment** section, below, for quantitative analysis of the accuracy).

If we cannot find a suitable peak in the high-frequency or RMS power, we use a comb filter based on the fundamental frequency of the target pitch. We look for the sharpest peak in the derivative of the RMS of the filtered signal, which will generally correspond to a point in the middle of the rise to peak power. We slide back in time to find the positive-going zero-crossing in the derivative, and take this as the onset. This introduces a timing bias into the extraction, but it can be corrected through comparsion to a "ground truth" signal to build a statistical model of the average bias by pitch.

In the case where multiple notes are struck simultaneously, we cannot use high-frequency or RMS power to select onset points, because there is no way to tell if the energy burst corresponds to a particular note in a chord, or to the chord as a whole. Instead, we build a multiple-bandpass filter, with pass regions selected to be those harmonics of the current pitch that do not have interference from another pitch – ie, that are not also overtones or fundamentals of another note that occurs at the same time. We filter the signal using this multiple-bandpass and take the RMS, and then use the derivative-based estimator described above.

The Q (ratio of center frequency to bandwidth) of the filters is selected depending on the expected proximity of other notes in pitch and time, ranging from 15 to 50. There is bias implicit in the higher group delay of narrowly tuned filters; this is a possible source of error in the algorithms which should be examined more closely.

2.2 Release Timing and Amplitude Measurement

Release timing and amplitude measurement is done with simpler techniques than the onset extraction. We build a multiple-bandpass filter based on the harmonics we know to be usable from the score-information, in a time window going from the previously-extracted onset time to a point 3 sec later in the signal. We look forward in time for the peak-power point in the filtered signal, and extract that as the amplitude. We then continue to look forward, for the point at which the power either drops below 10% of the peak power, or begins rising to another peak. This point is extracted as the onset. (See the **Discussion** section for criticisms of this method.)

2.3 Tempo Estimation

Tempo re-estimation is performed each time through the main program loop, to attempt to understand the local timing of the current performance and derive good guesses for the locations of the next few note onsets.

Currently, a regression line is calculated, matching predicted onset time for the last ten notes from the score against their extracted onset times. Predictions are then made by using the regression line to extrapolate timings for the next five notes. When the first note of a chord has been extracted, however, we choose its time as the prediction for the other notes in that chord.

This method is adequate for following the performance of the pieces used in the validation experiment. There are, of course, many other techniques for robust performance-following in the literature – [10], for example.

3 Validation Experiment

To analyze the accuracy of the timing and velocity information extracted by the system, a validation experiment was conducted using a Yamaha Disclavier MIDI-recording piano. This device has both a conventional upright piano mechanism, enabling it to be played as a standard acoustic piano, and a set of sensors which enable it to capture the timings (note on/off and pedal on/off) and velocities of the performance in MIDI format. The Disclavier also has solenoids which enable it to be used to play back prerecorded MIDI data like a player piano, but this capability was not used.

Scales and two excerpts of selections from the piano repertoire were performed on this instrument by an expert pianist; the performances were recorded in MIDI using the commercial sequencer *Studio Vision* by Opcode Software, and in audio using Schoeps microphones. The DAT recording of the audio was copied onto computer disk as a digital audio file; the timing-extraction system was used to extract the data from the digital audio stream, producing an analysis which was compared to the MIDI recording captured by the Disclavier.

It is assumed for the purposes of this experiment that the Disclavier measurements of timing are perfectly accurate; indeed, it is unclear what method could be used to evaluate this assumption in a robust fashion. One obvious test, that of re-synthesizing the MIDI recordings into audio, was conducted to confirm that the timings do not vary perceptually from the note timings in the audio, and this was in fact found to be the case.

3.1 Performances

There were eight musical performances, totaling 1005 notes in all, that were used for the validation experiment. Three were scales: a chromatic scale, played in quarter notes at m.m. 120 (120 quarter notes per minute) going from the lowest note of the piano (A four octaves below middle C, approximately 30 Hz) to the highest (C three octaves above middle C, approximately 4000 Hz); a two-octave E-major scale played in quarter notes at m.m. 120; and a four-octave E-major scale played in eighth notes at m.m. 120. Each of the two E-major scales moved from the lowest note to the highest and back again three times.

Additionally, three performances of excerpts of each of two pieces, the G-minor fugue from Book I of Bach's *Well-Tempered Clavier*, and the first piece "Von fremden Ländren und Menschen" from Schumann's *Kinderszenen* Suite, op. 15, were recorded. The first line of the score for each of these examples is shown in fig 2. All three Bach performances were used in the data analysis; one of the *Kinderszenen* performances was judged by the participating pianist to be an poor performance, suffering from wrong notes and unmusical phrasing, and was therefore not considered.

These pieces were selected as examples to allow analysis of two rather different styles of piano performance: the Bach is a linearly-constructed work with overlapping, primarily horizontal lines, and the Schumann is vertically-oriented, with long notes and heavy use of the damper pedal.

3.2 Results

Figures 3 to 11 show selected results from the timing experiment. We will deal with each of the extracted parameters in turn: onset timings, release timings, and velocity measurements. In summary, the onset timing extraction is successful, and the release timing and amplitude measurement less so. However, statistical bounds on the bias and variance of each parameter can

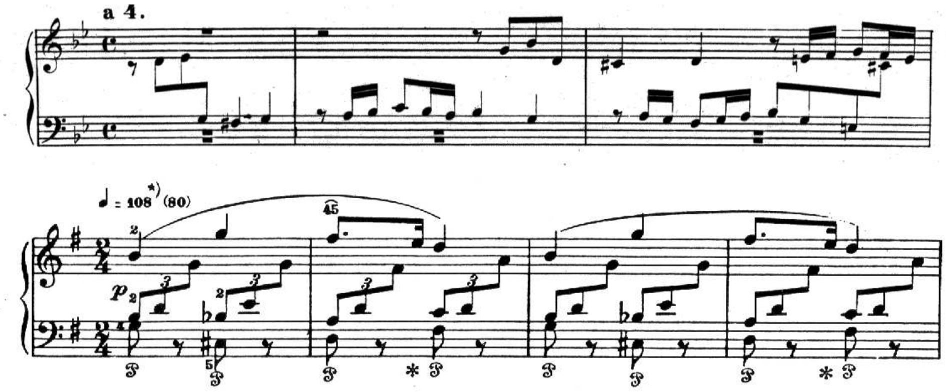

Figure 2: *Musical examples used: (top) Fugue 16, in G minor, from Book I of the* **Well-Tempered Clavier** *by J.S. Bach; (bottom) the first piece "Von fremden Ländred un Menschen" from Schumann's* **Kinderszenen** *suite*

be computed which allow us to work with the measurement to perform performance analysis of a musical signal in the stochastic sense.

3.2.1 Onset Timings

Foremost, we can see that the results for the onset timings are generally accurate. Fig. 3 shows a scatter-plot of the predicted onset time (onset time as recorded in the MIDI performance) vs extraction error (difference between predicted and extracted onset time) from one of the Schumann performances. The results for the other pieces are similar.

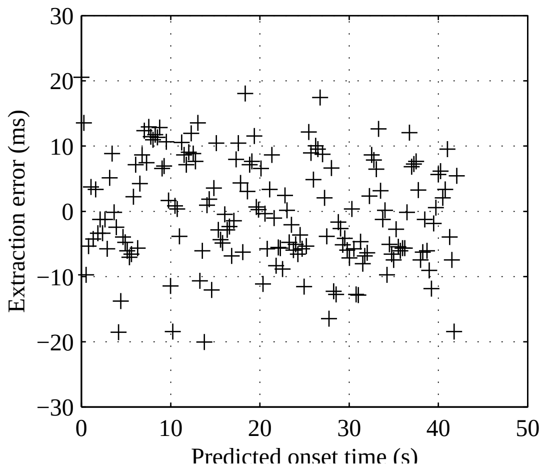

Figure 3: *Predicted vs extracted onset times. In general, onset extraction error is small.*

This is not nearly a strict enough test for our purposes, though. One possibility is to resynthesize the extracted performances and compare them qualitatively to the originals; or, for a quantitative comparison, we can examine the variances of the extracted timing deviations from the original.

Treating a piece as a whole, there is not useful information present in the mean of the onset timing deviations, as this largely depends on the differences in the start of the "clock time" for the audio vs MIDI recordings; measuring from the first onset in the extraction and the first attack in the MIDI simply biases the rest of the deviations by the error in the first extraction. In fact, the first extraction is often less accurate than those part-way through the performance, because there is not a tempo model built yet.

Thus, the global data shown below deals only with the variance of extraction error around the mean extraction "error". However, for results dealing with subsets of the data (ie, only monophonic pitches, or only pitches above a certain frequency), there are useful things to examine in the mean extraction error for the subset relative to the overall mean extraction error. We term this between-class difference in error the *bias* of the class.

Fig. 4 shows the standard deviation of onset timing extraction error for each of the eight pieces used (in order, the chromatic scale, the two-octave E major scale, the four-octave E major scale, the three performances of the Bach, and the two performances of the Schumann). We can see that the standard deviation varies from about 10 ms to about 30 ms with the complexity of the piece. Note that the second performance of the Schumann excerpt has an exceptionally high variance. This is because the tempo subsystem mis-predicted the final (rather extreme) *ritardando* in the performance, and as a result, the last five notes were found in drastically incorrect places. If we throw out these outliers as shown, the variance for this performance improves from 116 ms to 22 ms.

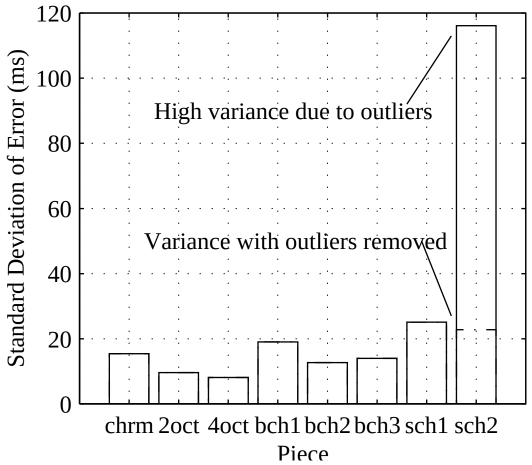

Figure 4: *Onset error standard deviation for each performance. Except for one piece with "outlier" errors, there is little difference between system performance for differing musical styles.*

Fig. 5 shows histograms of the deviation from mean extraction error for a scale, a Bach performance, and a Schumann performance. For each case, we can see that the distribution of deviations is roughly Gaussian or "normal" in shape. This is an important feature, because if we can make assumptions of normality, we can easily build stochastic estimators and immediately know their characteristics. See the **Discussion** section for more on this topic.

We can also collect data across pieces and group it together in other ways to examine possible systematic biases in the algorithms used. Fig. 6 shows the bias (mean) and standard deviation of onset timing extraction error collected by octave. We see that there is a slight trend for high

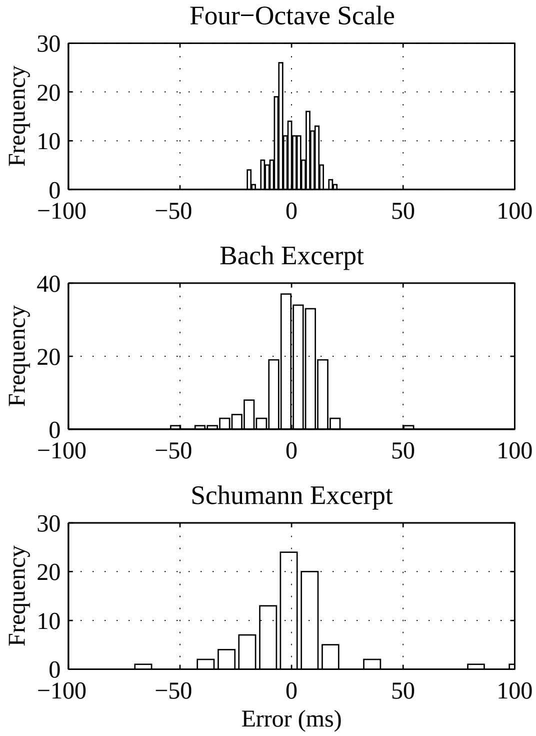

Figure 5: *Onset error standard deviation for three performances. The errors are distributed roughly normally, which implies that a Gaussian noise model is appropriate for stochastic characterization of the output of the system.*

pitches to be extracted later, relative to the correct timing, than lower pitches. Understanding this bias is important if we wish to construct stochastic estimators for the original performance. Note that this is not a balanced data set; the point in the center-of-piano octave represents about 100 times more data than the points in the extreme registers.

Similarly, Fig. 7 shows the bias and standard deviation of onset timing extraction error collected by the *method* used to extract the onset. As discussed in the **Algorithms** section, different methods are used to extract different pitches, depending upon the characteristics of the high-level score information, and upon the heuristic information extracted by the signal processing networks.

In Fig. 7, the "method used" is as follows:

- [A] No notes were struck in unison with the extracted note, and there is sufficient high-frequency energy corresponding with positive derivative in the fundamental bin to locate the note.

- [B] No notes were struck in unison with the extracted note. High frequency energy could not be used to locate the note, but RMS power evidence was used.

- [C] No notes were struck in unison with the extracted note; but there was not sufficient high frequency or RMS evidence to locate the note. The comb-filter and derivative method was used. These are, in general, represent "hard cases", where the audio signal is very complex.

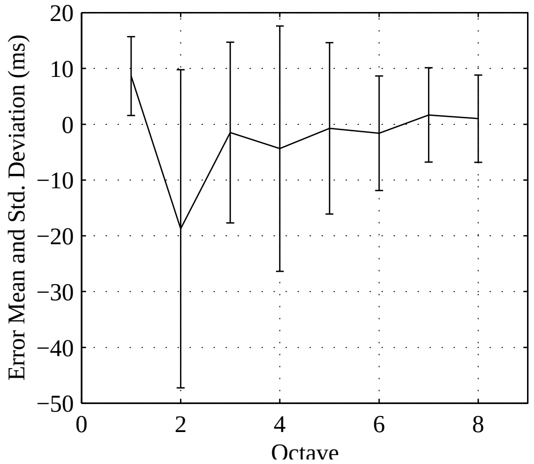

Figure 6: *Onset error mean and standard deviation by octave. Note bias of error by pitch, and increased accuracy (lower variance) for higher pitches*

- [D] There were notes struck in unison with the extracted note, so high-frequency and RMS power methods could not be used. The allowable overtones and derivative method were used.

We can see that there is a bias introduced by using method C, and relatively little by other methods. In addition, it is clear that the use of the high-frequency energy or RMS power heuristics, when possible, leads to significantly lower variance than the filtering-differentiation methods.

3.2.2 Release Timings

The scatter-plot of predicted release timing is shown in fig 8. As can be seen, there is similarly high correlation between predicted and extracted values as in the onset data. We can also observe a time relation in the data – this is due to the bias of release timing by pitch.

We can additionally plot predicted *duration* vs extracted; we see that there is not nearly as much obvious correlation, although the $r = .3163$ value is still highly significant statistically. This is shown in fig 9.

3.2.3 Amplitude/Velocity

A scatter-plot of predicted velocity against extracted log amplitude relative to the maximum extracted amplitude is shown in fig 10. As with duration, there is a high degree of correlation in the data, with $r = .3821$, although obviously not as much as with the onset timing extraction.

We can correct for the unit conversion between abstract MIDI "velocity" units in the predicted data and extracted log amplitude energy values by calculating the regression line of best fit to the fig 10 scatter-plot – $y = 7.89 - 79.4x$ – and using it to re-scale the extracted values.

When we treat the amplitude data in this manner, we see that once again, the noise from extraction error is quite well represented as a Gaussian distribution (fig 11), with standard deviation of error equal to 13 units on the MIDI velocity scale.

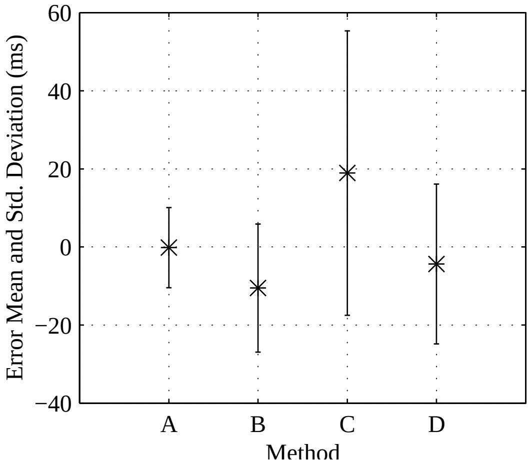

Figure 7: *Onset error mean and standard deviation by extraction method. See text for key to methods. As expected, method A (high-pass energy) is the most accurate and least biased.*

4 Discussion

There are a number of different levels on which this work should be evaluated: as a tool for music-psychology research, as an example of a system which performs musical transcription, and as an example of a multi-layered system which attempts to integrate evidence from a number of different information sources to understand a sound signal. We will consider each of these in turn, and then discuss ways in which the current system could be improved.

4.1 Stochastic Analysis of Music Performance

Part of the value of the sort of variance-of-error study conducted in the **Results** section is that we can treat extracted data as a stochastic estimator (cf, for example, [9]) for the actual performance, and make firm enough assumptions about the distribution of the estimation that we can obtain usable results.

It is clear that some aspects of expressive music performance can be readily analyzed within the constraints of the variance in extraction discussed above. For example, tempo is largely carried by onset information, and varies only slowly, and only over relatively long time-scales, on the order of seconds. Even the worst-case performance, with standard deviation of extraction error about 30 ms, is quite sufficient to get a good estimate of "instantaneous tempo" at various points during a performance.

For example, assume that two quarter notes are extracted with onsets 1.2 seconds apart, say at $t_1 = 0$ and $t_2 = 1.2$ for the sake of argument. We can assume, then, that these extractions are taken from Gaussian probability distribution functions (pdf's) with standard deviations of .02 seconds, and calculate the pdf of the inter-onset time $t_2 - t_1$ as Gaussian with mean 1.2 seconds and standard deviation .0283 seconds, giving us 95% probability that the actual tempo is in the interval [47.75, 52.48].

We can similarly recreate other sorts of analyses such as those found in [8] or [1] by treating

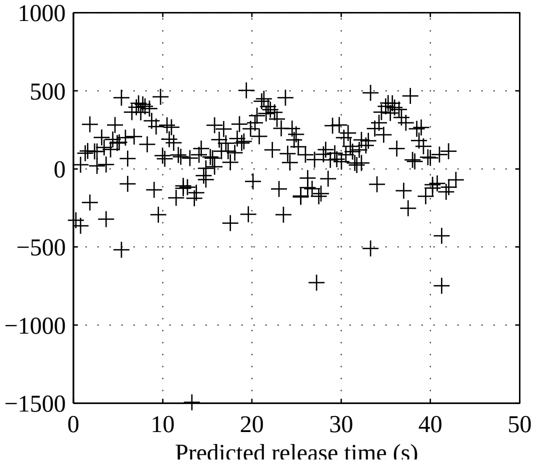

Figure 8: *Predicted vs extracted release time. Bias over time corresponds to bias by pitch.*

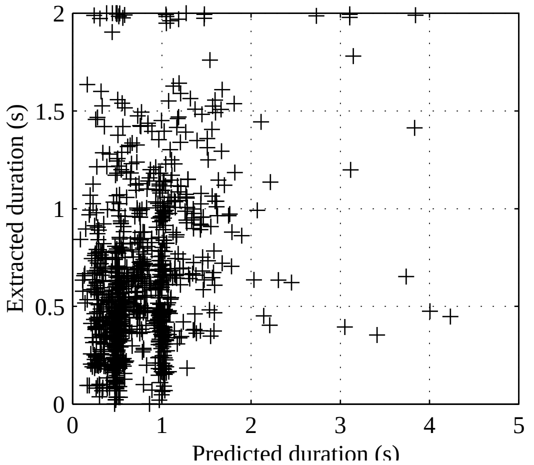

Figure 9: *Predicted vs extracted duration. The data fits a linear model rather well (see text).*

the timing variables as random Gaussian variables rather than known values.[2]

Depending on which question we want to answer, though, the answers may be less satisfactory for small timing details. For example, an important characteristic of expressive performance of polyphonic music is the way in which a melody part "runs ahead" or "lags behind" the accompaniment. To examine this question, we wish to determine the posterior probability that a particular note in a chord has been struck last, given the extracted onset timings.

Consider a two-note dyad, where the score indicates the notes are to be struck simultaneously; the onsets have been extracted as 1.000 and 1.015 sec, respectively. We can calculate the probabilities that the notes were actually struck within the 5 msec window of perceptual simultaneity, or that the earlier or later was, in fact, struck first. To do this calculation, we build a Bayesian estimator of the time lag, and use error functions; we find that the probability that the earlier

[2]It is arguable that they should have been treated this way in the cited work to begin with, since there is bound to be sensor noise coming into play.

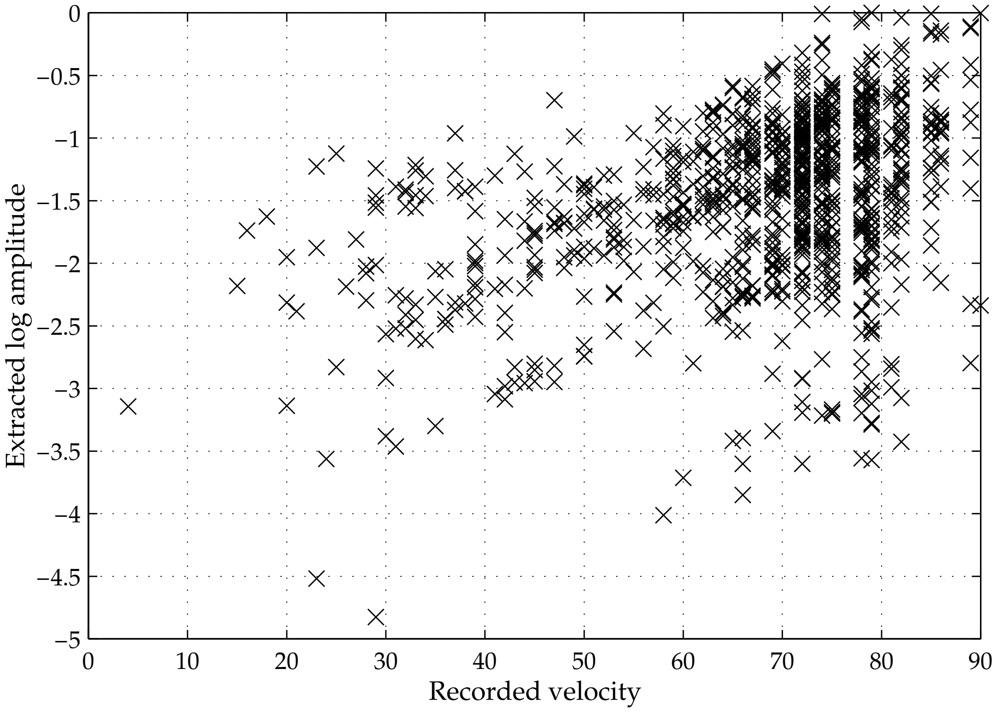

Figure 10: *Predicted MIDI velocity vs extracted amplitude. This data can be used to calcluate the proper scaling of output velocities.*

extraction was actually struck first is 0.6643, that and that the later extraction was actually first is .2858, assuming that the standard deviation is the worst-case of 25 msec.

4.2 Polyphonic Transcription

It is clear that using this sort of layered method with the score enables polyphonic transcription with more accuracy than previously-existing systems. When the extracted MIDI is resynthesized, the resulting performance is clearly the same piece performed in the "same style"; it is not indistinguishable from the original performance, due to errors, but many of the important aspects of the original performance are certainly captured.

The system has not been exhaustively tested on a wide variety of musical styles. The Bach example has four-voice polyphony in the score, which ends up being six- or eight-voice polyphony at points due to overlap in the performance. The Schumann has heavy use of the damper pedal, and so has sections where as many as nine notes are sustaining at once. The only musical cases that are not represented among the example performances analyzed above are very dense two-handed chords, with six or eight notes struck at once, very rapid playing, and extreme use of *rubato* in impressionistic performance.

It is anticipated that any of these situations could be dealt with in the current architecture, although the tempo-follower would have to be made more robust in order to handle performance which are not well-modeled by linear tempo segments. This is generally a solvable problem, though – see [10] for an example.

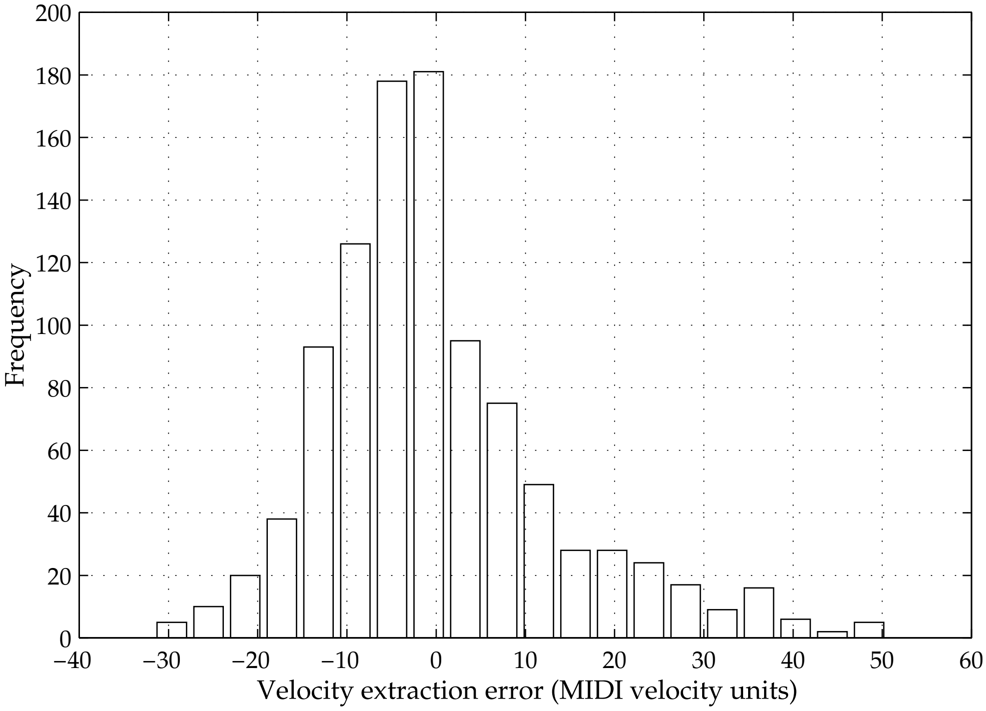

Figure 11: *Histogram of rescaled velocity extraction error. Again, note generally normal distribution.*

A larger issue regarding the problem of general polyphonic transcription is the goal and motivations underlying them. Why is there so much interest in building transcription systems?

We submit that it is for several reasons. Obviously, having a working transcription system would be a valuable tool to musicians of all sorts – from music psychologists to composers (who could use such a tool to produce "scores" for analysis of works of which they had only recordings) to architects of computer music systems (who could use it as the front-end to a more extensive music-intelligence or interactive music system).

Another reason that so much effort has been invested in the construction of transcription systems is that on the surface, it seems as though it "should be" possible to build them, because the necessary information "must be" present in the acoustic signal. While this feeling seems to underlie much of the work in this area, it is so far drastically unjustified.

This point relates to a final reason, which is based on a hypothesis of the human music cognition system – that human listeners are doing something rather like transcription internally as part of the listening process. Stated another way, there is an implicit assumption that the musical score is a good approximation to the mid-level representation for cognitive processing of music in the brain.

It is not at all clear at this point that this hypothesis is, in fact, correct. It may well be the case that in certain contexts (for example, densely orchestrated harmonic structures), only a schematic representation is maintained by the listener, and the individual notes are not perceived at all. Since it is exactly this case that existing transcription systems have the most difficulty with, perhaps we should consider building transcription systems with other goals in mind than recreating the original score of the music.

In particular, building systems which can build *human-like* representations from an audio stream would be adequate to supplant transcription for at least two of the three above-mentioned goals. For music psychologists, it is obviously only necessary to be able to extract parameters from music at a level similar to that possible by human performers; and for architects of interactive music systems, while it might be desirable to extract more detailed information, it is clearly (since humans are quite good at music performance with the ears they have) not necessary.

Even for composers, or jazz musicians who wish to produce fixed records of performances to study on paper, a representation which contains only the information a human perceives in the music (say, a melodic improvisation, a bass line, and a harmonic/rhythmic structure) would be quite an asset, even if it didn't tell the complete story.

4.3 Evidence-Integration Systems

The evidence integration aspects of the system are the most novel, and at the same time, the least satisfying. It is very difficult to build architectures which allow the use of data from many sources simultaneously; the one for this system is perhaps not as sophisticated as it could be. For example, the current system does not have the ability to use knowledge discovered in the attack (other than the timing) to help extract the release. Similarly, it would be quite useful to be able to examine the locations of competing onsets and decays in the extraction of parameters for a note with overlapping notes.

At the same time, though, the success of the system in its current state is promising with regard to the construction of future systems with more complex architectures.

4.4 Future Improvements to System

There are many directions which contain ample room for improving the system. Obviously, more work is needed on the release- and amplitude- detecting algorithms. It is expected that more accurate amplitude information could be extracted with relatively little difficulty; the results here should be considered preliminary only, as little effort has currently gone into extracting amplitudes.

Release timings are another matter; they seem to be the case where the most sophisticated processing is required in a system of this sort. Fig. 12 shows the major difficulty. When a note (for example, the C4 in fig 12) is struck after but overlapping a note which has the fundamental corresponding to an overtone (the C5), the release of the upper note becomes "buried" in the onset of the lower. It does not seem that the current methods for extracting release timings are capable of dealing with this problem, and that instead, some method based on timbre-modeling would have to be used.

It would improve the robustness of the system greatly to have a measure of whether the peak extracted from the signal for a particular note has a "reasonable" shape for a note peak. Such a measure would allow more careful search and tempo-tracking, and also enable the system to recover from errors, both its own and those made by the pianist.

Such a heuristic would also be a valuable step in the process of weaning a system such as this one away from total reliance upon the score. It is desirable, obviously, even for a score-based system to have some capability of looking for and making sense of notes that are not present in

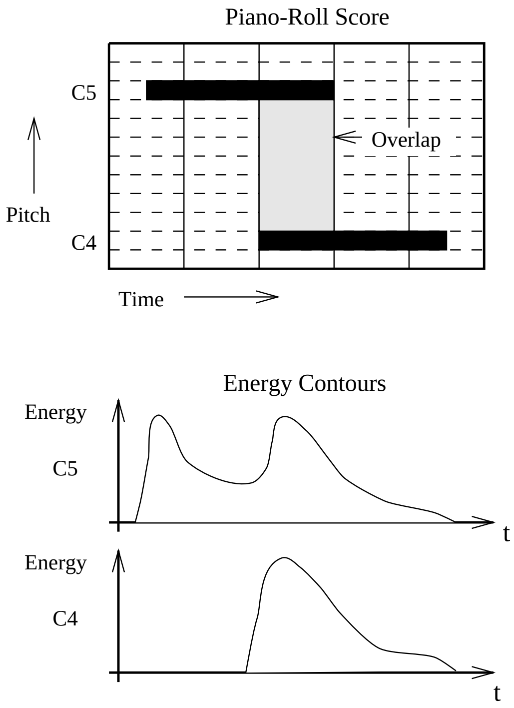

Figure 12: *A release gets buried by overlapping energy from a note an octave below, due to harmonic energy interference.*

the score. At the least, this would allow us to deal with ornaments such as trills and mordents, which do not have a fixed representation.

There are other methods possible for doing the signal-processing than those actually being used. One class of algorithms which might be significantly useful, particularly with regard to the abovementioned "goodness of fit" measure, is those algorithms which attempt to classify *shapes* of signals or filtered signals, rather than only examining the signal at a single point in time. For example, we might record training data on a piano, and use an eigenspace method to attempt to cluster together portions of the bandpass-filtered signal corresponding to attacks and releases.

Ultimately, it remains an open question whether a system such as this one can be expanded into a full-fledged transcription system which can deal with unknown music. Certainly, the “artificial intelligence” component, for understanding and making predictions about the musical signal, would be enormously complex in such a system. Work is currently in progress on a “blackboard system” architecture (see, eg, [7]) for investigation of these issues. An initial system being built using this architecture will attempt to transcribe “unknown but restricted” music – the set of four-part Bach chorales will be used – by development of a sophisticated rule-based system to sit on top of the signal processing.

5 Conclusion

Results here have shown that certain limited aspects of polyphonic transcription can be accomplished through the method of “guess and confirm” given enough a priori knowledge about the contents of a musical signal.

The resulting system is accurate enough to be useful as a tool for investigating some, but not all, aspects of expressive musical performance. The uncertainty introduced into note timings as part of the extraction is small enough to allow accurate tempo estimation and perhaps certain sorts of studies of phrasing. The system in its current form is probably not yet accurate enough to investigate subtle questions of melodic-harmonic timing.

Possible future work includes:

- Increasing the accuracy of the algorithms used in this system.
- Exploring the use of other algorithms or timbre models to augment the algorithms already in place.
- Building heuristics to determine whether a candidate note in a signal is likely to actually be a note.
- Building systems which can extract from instruments other than pianos, with more complex envelope shapes.
- Building rule-based or music-cognitive systems to replace the role of the score in this system.

6 Acknowledgments

Thanks to Barry Vercoe, Michael Hawley, and John Stautner for their advice during the progress of this research. Thanks also to Charles Tang for the piano performances used in the validation experiment, and to Teresa Marrin for providing the initial suggestion which led to this research. As always, the graduate students in the Machine Listening Group of the Media Lab have been helpful, insightful, provocative, and ultimately essential in support of the production of this paper and the research on which it is based.

References

[1] Jeff Bilmes. Timing is of the essence: Perceptual and computational techniques for representing, learning, and reproducing expressive timing in percussive rhythm. Master's thesis, MIT Media Laboratory, 1993.

[2] Stephen Handel. *Listening*. MIT Press, Cambridge, MA, 1989.

[3] Carol Krumhansl. *Cognitive Foundations of Musical Pitch*. Oxford University Press, Oxford, 1991.

[4] Fred Lerdahl and Ray Jackendoff. *A Generative Theory of Tonal Music*. MIT Press, Cambridge, MA, 1983.

[5] Eugene Narmour. *The Analysis and Cognition of Basic Melodic Structures*. University of Chicago Press, Chicago, 1990.

[6] Eugene Narmour. *The Analysis and Cognition of Melodic Complexity*. University of Chicago Press, Chicago, 1993.

[7] Alan Oppenheim and S. Hamid Nawab. *Symbolic and Knowledge-Based Signal Processing*. Prentice-Hall, Inc, 1992.

[8] Caroline Palmer. *Timing in Skilled Music Performance*. PhD thesis, Cornell University, 1989.

[9] Athanasios Papoulis. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, New York, NY, third edition, 1991.

[10] Barry Vercoe. The synthetic performer in the context of live performance. In *Proc. Int. Computer Music Conf.*, 1984.