Automatic Capture for Spectrum-Based Instrument Models*

Roger B. Dannenberg and Minoru Matsunaga
School of Computer Science, Carnegie Mellon University
{rbd, minoru}@cs.cmu.edu, http://www.cs.cmu.edu/~rbd

Abstract: Our goal is to automate the analysis of recorded acoustic performances in order to study the relationship between scores and performance. An automated system segments a recorded performance into individual notes. These are then analyzed to determine pitch and amplitude envelopes. Spectral data is also measured. The technique consists of two stages. First, a rough estimation stage performs pitch detection based on MQ analysis. Second, an accurate estimation stage uses period-synchronous analysis. The data will ultimately be used by a machine learning process to build instrument and performance models. Experiments with trumpet tones are described.

1. Introduction

To produce realistic approximations of acoustic instrument tones, it is important to use appropriate time-varying control functions. However, control functions are complex. Even simple control functions use a handful of parameters which depend upon musical context. (Dannenberg, Pellerin, and Derenyi 1998) Because of their complexity, it seems that the best way to obtain good control functions is to derive and study examples from acoustic performances. Our previous work was based on analyses that involved many manual steps. For example, performances were segmented into individual notes using graphical editors. These isolated tones were sorted and processed by hand to obtain envelopes for study.

Consequently, the number of envelopes available for study was limited. To explore performance with greater precision, we need much more data from a wide range of musical styles. We are working on an automated analysis system to assist in extracting envelopes for study. Simply generating more envelope data will not do much good if the envelope data must then be analyzed by hand. Therefore, we are also working on machine learning techniques for processing and generalizing performance models from the data. In this paper, however, we will only consider the analysis aspects of the overall problem.

In addition to control envelopes, our synthesis method (described below) requires a spectral database: a table of spectra indexed by amplitude and fundamental frequency. In our previous work, regions of tones to be analyzed were selected by hand, and the resulting data was then organized into a table. By combining pitch and amplitude detection, we can automate the process of creating the spectral database.

To summarize, the goal of this work is to automatically extract example control functions and spectra from digital recordings of acoustic instruments. The extracted information is used to develop better models of control and to simplify the gathering of spectra needed to characterize an acoustic instrument.

1.1 Combined Spectrum Interpolation Synthesis

Combined Spectral Interpolation Synthesis (CSIS), described by Derenyi and Dannenberg (1998), produces digital audio output from symbolic musical score data. CSIS has two sub-parts.

The first, called the *performance model*, produces time-varying amplitude and frequency control curves as an intermediate representation. The performance model is based on an examination of measured amplitude and frequency contours and how they relate to symbolic score data. For example, a quarter note followed by a tongued attack will have a much different envelope than the same quarter note slurred to the next note.

Previously, notes were carefully selected and analyzed by hand, and it was difficult to generalize from a small set of examples. The resulting performance model, although good, had limitations.

The performance model drives the second part, called the *instrument model*, which produces the final audio output. The instrument model is based on the idea of generating the appropriate spectrum for a given amplitude and fundamental frequency. To build an instrument model, we must measure the spectrum at many different frequencies and amplitudes. This is also an area where automation can help.

*Originally published as: Dannenberg, Roger B. and Minoru Matsunaga. 1999. "Automatic Capture for Spectrum-Based Instruments." In *Proceedings of the International Computer Music Conference 1999.* San Francisco: International Computer Music Association, pp.145-148.

1.2 The Problem

The goal is to obtain control envelopes from a performance of real music as opposed to isolated tones. We need to relate envelope features to symbolic score features, and for that we want the envelopes corresponding to individual notes. Thus, one of the first requirements is to segment a performance into individual notes. Once a note is identified, it is straightforward to determine its amplitude and frequency envelopes. In addition, we need to measure spectra for each pitch at several different amplitudes. This is also straightforward once notes are identified and separated. The most difficult aspect of this problem is to locate transitions between notes accurately. When performances are resynthesized, small errors in analysis lead to audible artifacts, so accuracy is important. (Although our goal is not pure resynthesis, it seems reasonable to assume that if the data cannot pass the resynthesis test, it will lead to problems elsewhere.) Identifying a note from a recorded performance is difficult for several reasons:

i. Determining the start or stop time of a note is different depending on whether the transition is to another note or to silence.

ii. At transient points such as the attack and release, the waveform is not periodic enough to determine the frequency reliably, so changes in pitch are not always evident.

iii. Frame-based analysis techniques such as RMS amplitude or the STFT tell that something happened within a certain frame, but they do not give precise timing.

iv. Sometimes, the performance is "noisy" because the performer fails to drive the instrument quickly into a stable oscillation.

v. Sometimes, analyzers output meaningless or noisy, ambiguous results.

Note that the problem could be much more difficult if we were to analyze performances involving extended techniques – esentially any performances where the instrument is not producing a stable oscillation – as this would introduce much more energy in the form of irregular transients and noise. Also, we are currently working only with trumpet and trombone tones, so there is no possibility of polyphony as in string instruments.

2. Related Work

Many researchers have analyzed music signals for transcription and music analysis. Space prohibits a detailed review of transcription systems, but in general, these systems are not so concerned with the precise boundaries of notes as they are with the identification of notes and the quantization of note times into rhythmic values. Foster, Schloss, and Rockmore (1982) describe some signal processing techniques for identifying note transitions that relate to ours: namely, tracking features backward in time from well-defined regions, looking for sharp transitions in the features indicating an attack. Their work addresses overlapping notes in keyboard and vibraphone music, a difficult problem that is not considered here.

Work by De Poli, Roda, and Vidolin (1998) is closely related to ours, and, like our project, analyzes envelopes to study performance nuance. This paper notes the difficulty of designating precise starting and ending times for notes, and the authors resort to a simple threshold crossing technique. This works well for studying properties of envelopes, but we found it is not accurate enough for detailed envelope models and resynthesis using CSIS.

3. Procedure

Our analysis procedure consists of two stages. The first stage is a rough estimation stage using MQ (McAulay-Quatieri) analysis from Beauchamp's (1993) SNDAN program. The second stage is for accurate estimation using period-synchronous analysis. The first stage calls upon SNDAN to estimate pitch, and then applies the following steps:

i. Quantize each fundamental frequency to the nearest semitone.

ii. Smooth the quantized frequency curve.

iii. Determine the start and end times of continuous frequency segments.

iv. Eliminate any segment that is too short.

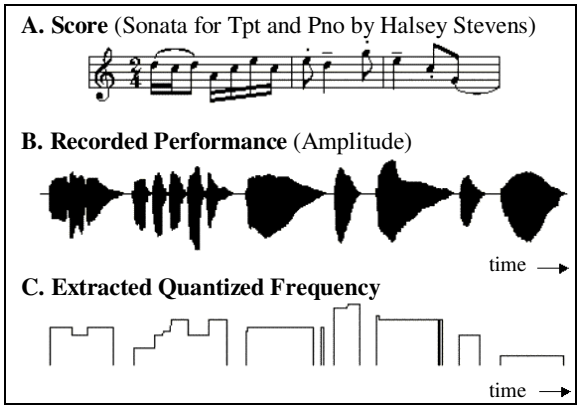

Figure 1. Rough estimation stage.

Figure 1 shows a score, a recorded performance, and the extracted and quantized frequency.

In the second stage, the fundamental frequency and estimated start and end times, which are output by the first stage, are used as initial values. The process is based on the idea of visually following a waveform in a waveform editor and works as follows:

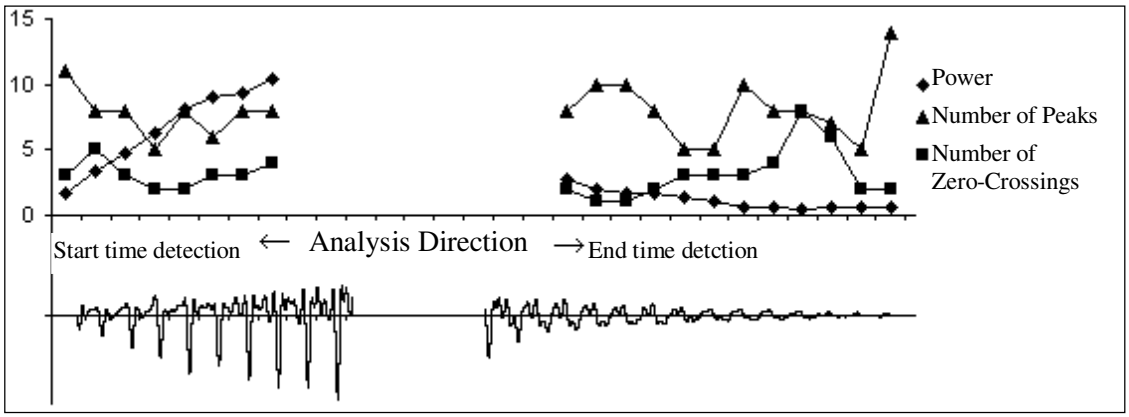

Figure 2. Relationship between recorded performance and feature parameters.

i. Determine an early steady portion of the note by determining where the amplitude reaches 60% of the maximum.

ii. Calculate an initial period in the steady portion of each note using an autocorrelation function. The fundamental frequency estimate from the first stage is used to determine which peak of the autocorrelation function represents the fundamental, assumed to be constant.

iii. Calculate power, number of zero-crossings, and number of peaks per fundamental period, starting from the steady portion and searching backward in time.

iv. If there is no silence between notes, the period which has the minimum power value will indicate a reasonably accurate start time. If there is silence between the notes, determine the start time by looking for no zero-crossings or a 20% increase in the number of peaks from the previous or the initial period.

v. To determine the end, pick a steady portion of the note where the amplitude falls to 20% of the maximum. Repeat the previous steps (iii. & iv.), but search forward in time instead of backward.

Figure 2 shows a recorded performance and extracted feature parameters for detection with silence before and after the note. Note how the number of peaks increases sharply on the far left and far right. These increases indicate the beginning and end of the note, respectively. If the zero crossing count goes to zero, that is also interpreted as a beginning or end.

In addition to note segmentation, we need to analyze spectra. A recording is made of a chromatic scale with separation between each note. These notes are easily segmented using the procedure just described. Then, each note is analyzed using period-synchronous Fourier transforms. RMS amplitude is computed period-by-period from the short-term spectra. The data is then scanned to select a spectrum for each amplitude of interest. The selected tuples of (pitch, amplitude, spectrum) data are saved for synthesis.

4. Experiments

We applied this analysis system to a recorded performance of a trumpet playing the first 29 measures of "Sonata for Trumpet and Piano," by Halsey Stevens (1959). There are 81 notes ranging from 16^{th} notes to a half note tied with an 8^{th}, at a rate 116 to 120 beats per minute. Pitches range from f3 to a5, and there are 7 rests. Articulations include staccato, slurred, and legato. Only one recorded performance was fed to the capturing system. The system has also been used to capture trumpet and trombone spectra necessary for CSI synthesis.

5. Results

In our task, we must segment the notes with very accurate transition times. Although there is not always a well-defined boundary, we would like the error to be less than a few periods, or about 5ms. Compared to hand labeled data, the automated system identified attack positions with an average error of 3.7ms, and releases with an average error of 15.1ms. These numbers are good, but not always good enough for our requirements.

In our experiment, 53 notes (65.3%) were captured without problem. 13 meaningless segments, 2 notes with shorter attacks, 8 notes with longer attacks, 16 notes with shorter releases and 2 notes with longer releases were also captured. 13 meaningless segments are because of frequency analysis errors. 11 of the shorter releases are due to the noisy release between notes (as shown in Figure 3A.), and 2 of the longer releases are due to the misleading shape of the amplitude envelop (as shown in Figure 3B.). 4 shorter releases are due to frequency analysis errors.

6. Future Work

The experiment shows promising results, but it is not reliable enough for a completely automated system.

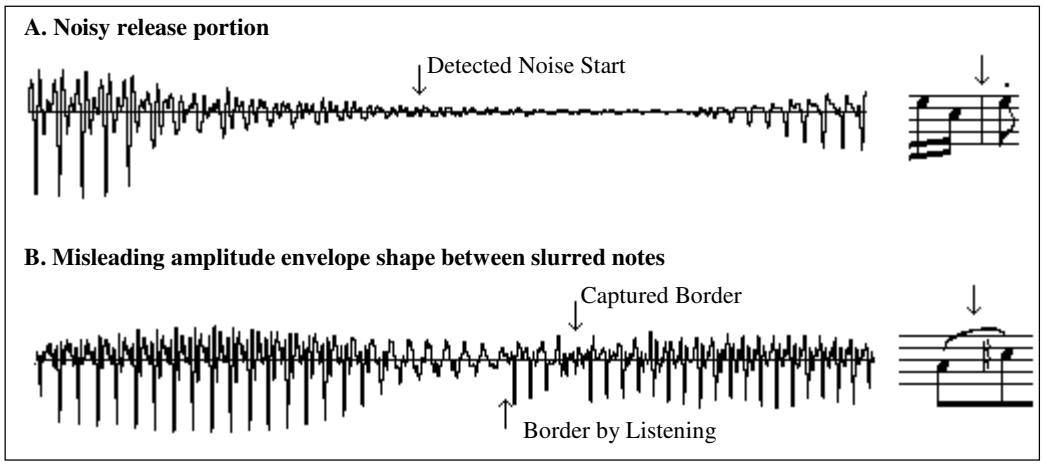

Figure 3. Typical Errors

To improve the accuracy, adoption of more advanced analysis and pattern matching techniques, consideration of score parameters, and parameter tuning for specific instruments will help. To build the performance model of CSIS, the captured notes need to be labeled with score parameters, and fed to a proper machine learning process.

This technique can be applied not only to CSIS, but also to any interesting behavior of notes in a musical context. For example, one might study intonation trends among different instruments or articulation differences between two performers. Using automation to parse recordings simplifies the use of actual performances as opposed to isolated tones. In addition, automated segmentation may make it possible to collect enough data to relate analysis features to properties of the notated score. For example, one might study whether a performer's intonation is biased from equal temperament toward harmonic ratios, something that would require an examination of intonation in a musical context.

7. Summary and Conclusions

We have constructed a system to extract notes from recorded music performances. RMS amplitude, the number of peaks, and the number of zero-crossings are used as features to determine the start and end times of notes. Once extracted, notes can be analyzed to obtain example envelopes, spectra, and other interesting properties. Our techniques segmented about two-thirds of the input data as well as can be done manually. This is adequate for some applications, but further improvements and/or interactive analysis systems will be required for more demanding situations.

References

Beauchamp, J. 1993. "Unix Workstation Software for Analysis, Graphics, Modification, and Synthesis of Musical Sounds." Audio Engineering Society Preprint, No. 3479 (Berlin Convention, March).

Dannenberg, R.B., H. Pellerin, and I. Derenyi.1998. "A Study of Trumpet Envelopes." In *Proceedings of the International Computer Music Conference*. ICMA. pp. 57-61.

De Poli, G. A. Roda, and A. Vidolin. 1998. "Note-by-Note Analysis of the Influence of Expressive Intentions and Musical Structure in Violin Performance." *Journal of New Music Research,* **27**(3) (September), pp. 293-321.

Derenyi, I. and R. B. Dannenberg. 1998. "Synthesizing Trumpet Performances." In *Proceedings of the International Computer Music Conference*. San Francisco: ICMA, pp.490-496.

Foster, S., W. A. Schloss, and A. J. Rockmore. 1982. "Toward an Intelligent Editor of Digital Audio: Signal Processing Methods." *Computer Music Journal,* **6**(1) (spring), pp. 42-51.

Stevens, H. 1959. "Sonata for Trumpet and Piano," (musical score) New York. C. F. Peters Corporation.