Center for Computer Research in Music and Acoustics April 1986

Department of Music
Report No. STAN-M-34

SOURCE SEPARATION AND NOTE IDENTIFICATION IN POLYPHONIC MUSIC

by

Chris Chafe and David Jaffe

Research sponsored by

National Science Foundation
and
System Development Foundation

CCRMA
DEPARTMENT OF MUSIC
Stanford University
Stanford, California 94305

SOURCE SEPARATION AND NOTE IDENTIFICATION IN POLYPHONIC MUSIC

Chris Chafe and David Jaffe

Center for Computer Research in Music and Acoustics (CCR
Department of Music, Stanford University
Stanford, California 94305

ABSTRACT

Experiments in automatic music recognition at CCRMA have been in progress for five years. Digitized sound recordings of instrumental music are analyzed and transcribed by computer. The current effort is directed at polyphonic examples with a variety of instruments and musical styles. The paper discusses acoustic analysis issues in accurately transcribing polyphonic input.

The overall goal of the work is to provide a tool for the study of musical performance, for applications requiring tracking of live musicians, for manuscript work and for segmentation of digital audio recordings.

INTRODUCTION

In transcribing performed music, a system must extract every note played, identifying timings, pitch, dynamic information and other parameters. A capability for source discrimination is also required if a polyphonic musical texture is present, that is, if the signal results from multiple simultaneous sources.

Identification algorithms which are adequate for monophonic input [1,2,3] may not be effective with polyphonic textures. Balance between underdetection and overdetection may become difficult. Improved performance is possible by combining the strengths of alternative algorithms for each identification task. The use of acoustic knowledge and context generated at higher levels also assists in the recognition process. As signal history accumulates, a particular event need not be as visible for successful identification if it clearly fits into an already established context.

Quantifying *source coherence* has become the focus in understanding polyphonic input. Overlapping and interleaved sources are prone to yield spurious note interp
tions. By analyzing source coherence, the system can
the note hypotheses that it generates.

Knowledge of source acoustics can limit the numb
possible interpretations. This is particularly useful
disentangling spectral lines into their sources. As McA
[4] has suggested, constraint rules can be derived fo
allowable behavior of a given component as a member
source and may prove helpful in filling in obscured p
resulting from limitations in the available data.

SYSTEM OVERVIEW

The acoustic analysis is broken down into four steps

- **Spectral Transformation.** High resolution in
frequency domain is required to distinguish ne
components of pitches in close intervals [5].
Bounded-Q Frequency Transform (BQFT), is a p
constant-Q technique yielding better than semit
resolution in each octave of interest [6].

- **Event Detection.** To segment the signal, pertu
tion in different time-varying envelopes is detec
Total amplitude alone can provide correct att
detection of better than 95% for the piano. Rec
work in bowed string sound has motivated detect
of frequency domain events representing other ki
of musical articulations.

- **Generation of Early Context.** Musical cont
is used to complement the results of local segm
tation. Event timings first build a metrical g
Rhythmic patterns then suggest events detecti
that are weak or missing.

- **Periodicity Estimation and Source Tracki**
Stable segments between detected transitions are
amined. Chords are separated into source hypoth
and their individual partials are tracked in tin

This work was supported by the National Science Foundation under Contracts NSF MCS-8012476 and DCR-8214350. Xerox Corporation has provided computer equipment for the research.

Hypotheses are weighed for coherence and "good" notes are added to the note list, labelled with timing, pitch and dynamic information.

Event Detection

Periodicity estimation is more apt to be meaningful in the *stable islands* between transitions identified by segmentation. A number of event detection schemes are invoked in segmentation, to search for cues in:

- Amplitude envelope.
- Wide-band and narrow-band spectral change over time.
- Rhythmic expectation (discussed elsewhere [5,7]).

The amplitude envelope of the signal $y(\cdot)$ is "surfed" with a moving linear regression of length L to find onsets with large slope increases [8].

$$\text{For each k, find}\{m_k, b_k\}\text{to minimize} \sum_{n=k}^{k+L-1} (y(n) - m_k n - b_k)^2$$

Reliability decreases when the technique is applied to multi-source textures where events are masked by overlapping sources, as well as for non-percussive instruments which can play a sequence of notes continously with little amplitude inflection. Highlighting spectral change supplies additional transition cues. For each of the M log magnitude channels y_m, a difference of the current sample and an average of the N previous samples is formed. This is converted to a "ratio error" by normalizing by the average sample with thresholding applied to ignore minor fluctuations. A pan-spectral ratio error signal $s(n)$ is formed by summing the ratio error signals of all the channels as follows:

$$s(n) = \sum_{m=1}^{M} \frac{y_m(n) - \hat{y}_m(n)}{\hat{y}_m(n)},$$

$$\hat{y}_m(n) = \frac{1}{N} \sum_{i=1}^{N} y_m(n-i).$$

Finally, the linear regression "surfboard" is run on the summed ratio error signal.

In cases where simple amplitude envelope detection blurs together multiple sources, division into spectral channels reveals individual sources. The advantage derived from the *adaptive* method of Eq. 2 is its sensitivity to spectral state. By dicing the spectrum more coarsely or finely different event types are revealed (Figs. 1 and 2).

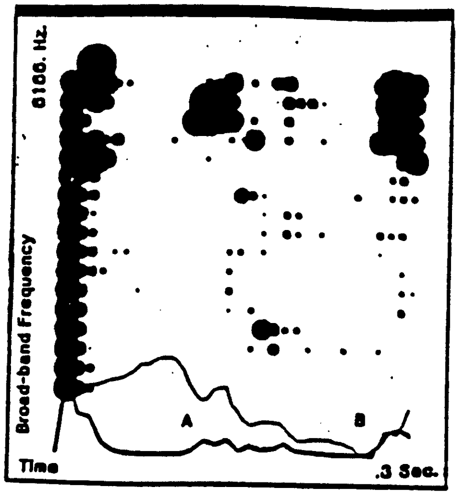

Figure 1. A pitch bowed on the cello is repeated at points A a
The first, a legato stroke, is not visible in the amplitude envelope
is evident in the *pan-spectral ratio error signal* (dark tr
The *broad-band* auditory transform in the spectogram-like dis
shows detail in spectral weighting, especially changes in n
components.

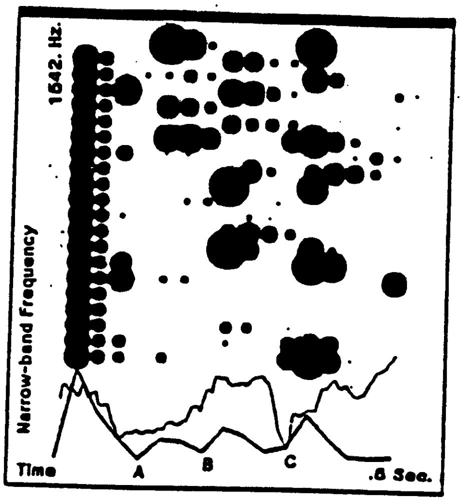

Figure 2. A *narrow-band* spectral transform reveals slurred
changes on the cello at points A and B followed by a bow and
change at C.

Some musical articulations appear to be detectable only one level of resolution. Therefore the capabilities several detectors which concurrently provide the perp tives of a variety of frequency and time resolutions a merged for stronger polyphonic event detection (Fig. 3).

Periodicity Estimation

Contaminating activity can bleed in from neighbori segments through the long time windows of the lowest o taves in the BQFT. To avoid this, segments from *stable lands* are excised from the original signal and then retrans

Periodicity estimation is accomplished using a "pian tuned" variant of Algorithm 2 from Amuedo [9]. The mod improve performance by taking advantage of knowledge

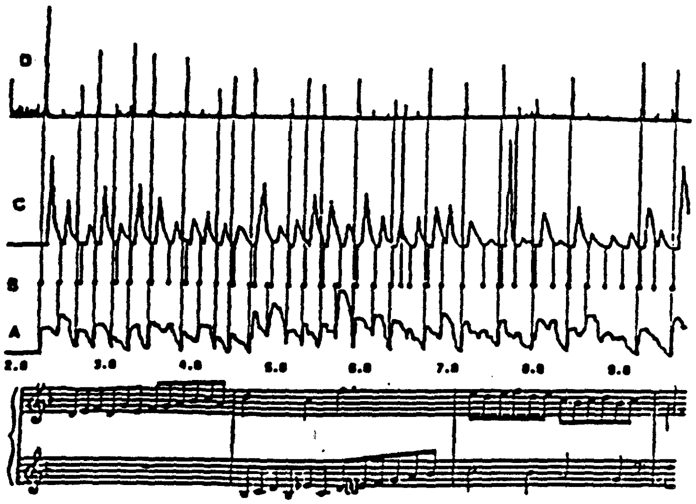

Figure 3. A piano performance of a Frescobaldi excerpt. Trace A, the detector on the amplitude envelope, misses some events caught by two detectors similar to those in the previous figures: Trace C (narrow-band, like figure 2), and trace D (broad-band, like figure 1). The merged set of event timings, B, is later pruned by pattern analysis and used to find stable segments for periodicity estimation.

some characteristics of the piano. The algorithm operates as follows:

- Generate pitch hypotheses.
- Quantize pitches to possible source piano pitches.
- Rank according to overtone presence.
- Rank according to total energy.

First, an instantaneous *best resolution* spectrum is obtained by taking the upper octave of each octave decimation in the BQFT. Pre-BQFT high frequency emphasis is removed and peaks representing significant sinusoidal components are identified. For each of these components, a maximum of N pitch hypotheses is generated by dividing the component frequency by $x_i = is^{i-1}$, where i goes from 1 to N and s is an approximation of the stretching resulting from the stiffness of the piano strings (nominally 1.0015, but varying with pitch register [10]). Piano tones nearly always produce significant energy at either the first or second partial. Therefore, N is set to 2.

Hypotheses lying in the cracks of the keys, and clusters of near-lying hypotheses (less than a semitone) are fixed into a single source at the closest tempered pitch. Any hypotheses which lie outside the range of the piano are rejected.

A maximum of M partial positions are generated by multiplying each hypothesis frequency by x_i. Use can be made of the fact that piano strings tend to have their energy concentrated in the lower partials. A score is created as $\sum_{i=1}^{M} \delta/\text{sqrt}(x_i)$, where $\delta = 1$ if the corresponding partial position is close to one of the originally identified peaks and $\delta = 0$ otherwise. Hypotheses with scores above some threshold are asserted.

Energy is summed for all partials of each hyp and those hypotheses whose energy sum falls t threshold are discarded. The algorithm returns th ing candidates as the set of fundamental periodicit in the signal.

Source Verification

The following *partial grouping clues* provide evi quantifying source coherence:

- Spectral evolution of transients.
- Modulation behavior.
- Resonance structure.

Partials of musical oscillators appear and decay ir fashion and have characteristic frequency skew pa points of excitation or transition. Those generat damped resonator with a single excitation (like string) start at approximately the same time and d ponentially. A group spread of 25 msec or less is a able definition of synchronicity.

Variation in frequency (vibrato) and amplitude (t are correlated in an ensemble of partials comprising On the multiple string notes of the piano, beating correlated AM across several partials. For self-su oscillators (after the initial transient), FM is locke all partials and coupled with varying amounts of A Such information may help to associate partials t time.

FM and AM coupling may also include time-varyin behavior (for example, a vowel formant structure varies in a coherent fashion for a given source. Re structures generally vary more slowly than the afc tioned modulations. Source identity is strengthe deducing the resonance structure and the constraint variation as a function of other note attributes (r loudness, etc.) [12].

SUMMARY

The digitized waveform is examined for acoustic (broadly defined), and these in turn are examined f tials, instrumentation and musical notes. The pa problem in polyphony is to group spectral componer sources in roughly the same fashion as does the ear.

Event detection combines cues from several domain outpt of an auditory transform [13] (synthesized from data) is processed in a crude approximation of sensory tation in the ear-brain system. The purpose is to c spectral cues available to the ear, ignoring *known* or *old information* while focussing attention on *new information*. Views of spectral data from multiple levels of resolution increase the number of cues that are recognizable.

Periodicity estimation techniques form the basis for suggesting possible sources within stable segments. Searches are constrained and reliability increasesd by building in knowledge of source acoustics. Other probes are needed that will bring to bare cues other than periodicity.

Sources that have been uncovered in the signal are tracked through time for verification. The present effort is aimed at gathering features which provide a foundation for establishing *source coherence* in observed data.

The gathering of musical representations of the signal proceeds in parallel. As processing moves in time, chunks of data "bubble-up" from lower processes. Higher levels build streams of abstractions at lower data rates [3, 14]. As musical features are recognized, they are used to fine-tune the results of the acoustic analysis stage [15].

For a simplified picture of how such feedback improves analytic performance, consider the role of the "cognitive flywheel" [16] in hearing music. The ear picks out particular instruments (up to a point) within complex musical fabrics. Various features allow this discermination: dissonant pitches, registral placement, quality, characteristic rhythm, etc. Attention is focussed not only on individual sonic events but on longer chunks of time comprised of patterns of events as well. Aggregates of events become easily condensed into single entities such as "thematic repeat." Once the template for a larger chunk is established, attending to it is less a function of receiving all sonic cues than of distinguishing some sufficient subset. The current state of the music can be described by some number of currently active, partially filled templates and, perhaps, some new ones in the process of formation.

Events in a short term acoustic picture of the music can offer an idea of which instruments are currently sounding. However, the problem arises as to whether an acoustic event can be identified from its apparent features (perhaps incomplete) or whether it is better to deduce its identity by its place in a chunk of events describe by a well established template. The latter is likely to be the stronger choice. Instead of requiring a large number of principal acoustic features be noticed, advantage is made of information which has been gathered over multiple occurences of an event.

REFERENCES

[1] S. Foster et al. "Toward an Intelligent Editor of Digital Audio: Signal Processing Methods," *Computer Music Journal* vol. 6, no. 1, 1982.

[2] S. Foster, "A Pitch Synchronous Segmenter for Musical Signals," ICASSP, Paris FR, 1982.

[3] C. Chafe et al. "Toward an Intelligent Editor of Digital Audio: Recognition of Musical Constructs," *Computer Music Journal* vol. 6, no. 1, 1982.

[4] S. McAdams, personal communication.

[5] C. Chafe et al. "Techniques for Note Identification in Polyphonic Music," *Proc. Inter. Computer Music Conf., 1985*

[6] B. Mont-Reynaud. "The Bounded-Q Approach to Time-Varying Spectral Analysis," *Department of Music Technical Report STAN-M-28*

[7] B. Mont-Reynaud and M. Goldstein, "On Finding Rhythmic and Melodic Patterns in Musical Lines," *Proc. Inter. Computer Music Conf.* Vancouver BC, 1985.

[8] A. Schloss, "On the Automatic Transcription of Percussi Music," Ph.D. Thesis, Dept. of Speech and Hearing, Stanford University, June 1985. *Department of Music Technical Report STAN-M-27.*

[9] J. Amuedo, "Periodicity Estimation by Hypothesis-Directed Search," ICASSP, Tampa FL, 1985.

[10] J.O. Smith, personal communication.

[11] S. McAdams, "Spectral Fusion, Spectral Parsing and the Formation of Auditory Images," Ph.D. Thesis, Dept. of Speech and Hearing, Stanford University, June 1985. *Department of Music Technical Report STAN-M-22.*

[12] X. Rodet, "Time-domain Formant-wave-function Synthesi in J.C. Simon (ed.), *Spoken Language Generation and Understanding* Reidel:Dordrecht, 1980.

[13] A. Stautner, "Analysis and Synthesis of Music Using the Auditory Transform," M.S. Thesis, Dept. of E.E. and C.S., MIT, May 1983.

[14] B. Mont-Reynaud et al. "Intelligent Systems for the Analysis of Digitized Acoustic Signals, Final Report," *Department of Music Technical Report STAN-M-15.* 1984.

[15] B. Mont-Reynaud " Problem-Solving Strategies in a Music Transcription System ," *Proc. IJCAI, 1985.*

[16] H.P. Nii. "Signal-to-Symbol Transformation: HASP/SIAP Case Study," Stanford University, April 1982. *Heuristic Programming Project Rep. HPP-82-6.*