MUSICAL NOTE SEGMENTATION EMPLOYING COMBINED TIME AND FREQUENCY ANALYSES

Gordana Velikic, Edward L. Titlebaum, Mark F. Bocko
{gvel, tbaum, bocko}@ece.rochester.edu

Music Research Lab and ECE Department - University of Rochester, NY 14627 USA

ABSTRACT

In this paper we describe a musical stream note segmentation method that employs time-domain and frequency-domain analysis methods working in conjunction. The method has two demonstrated benefits: first, it leads to reliable results with very low probability of either missing or of falsely detecting notes, and second, it has temporal resolution on the order of 1 msec. We have applied the method to a variety of monophonic musical instrument recordings, including the clarinet, piano and violin, with results that vary from 95% to 100% accuracy.

1. INTRODUCTION

Segmentation of a musical audio stream into separate musical events (notes) typically is the first step in music recognition algorithms. Such algorithms have applications in areas such as "smart" musical accompaniment, automated music transcription and automated musical style analysis. It is in the latter context that the methods described in this paper have been developed, specifically for a computer-based system to provide real-time visual display of note timing for teaching in a musical ensemble setting and in our efforts to assess the influence of network latency on the performance musicality of remotely located musicians playing together over the Internet.

Music recognition systems characteristically combine "low-level" DSP based techniques with "high-level" decision making strategies, such as Bayesian methods, to achieve a parsing of the musical stream. High-level decision making strategies attempt to resolve any residual ambiguities in the low-level results, which may be due to either the limitations of the low-level methods or to ambiguity intrinsic to the musical stream. Thus, the effectiveness of an overall system is critically dependent upon the quality of the results from the low-level analysis and ideally, any remaining ambiguity in the low-level results would be only that which is intrinsic to the musical passage. The work described here presents a "low level" musical stream segmentation method that combines time-based and frequency-based analysis methods. This approach has several favorable features; first it enables precise note onset times to be determined (typically to 1 millisecond of resolution) and second, the combination of the time and frequency methods allows the resolution of note onset ambiguities that exist when a single analysis method is employed.

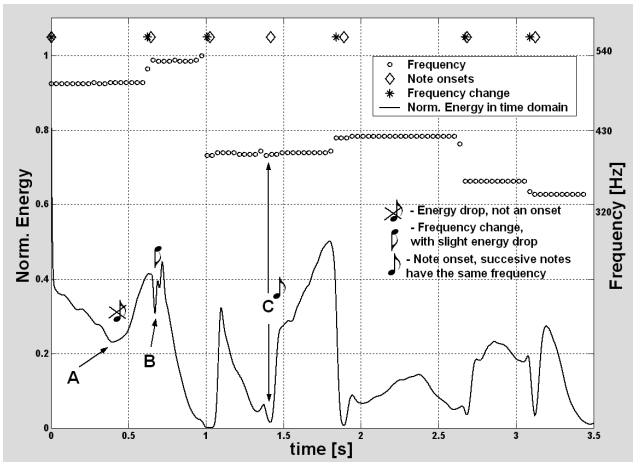

Figure1. The normalized signal energy and the frequency of the dominant spectral feature in the STFT are plotted versus time. Diamonds denote note onsets. The energy minimum at (A) may be mistaken for a note onset. This candidate may be vetoed by the frequency data, which indicates a continuous note. Time domain methods alone may miss event B (an example of slurred notes) in which there is a small change in energy but a significant frequency change. Finally, due to the relatively poor time resolution in the STFT, the transition between notes of the same frequency may be overlooked (event C). This onset is detected on the basis of a decrease in energy.

The greatest success at note segmentation has been achieved for piano recordings [1,2] since the percussive attack of the piano lends itself to energy-based techniques in either the time or frequency domains [3,4,5,6,7]. Wind instruments, such as the clarinet, or bowed stringed instruments, such as the violin, present a much more difficult challenge due to the less percussive nature of the note onsets, greater variations of pitch and the presence of vibrato and tremolo (defined in the present context as

amplitude and frequency modulation respectively). Therefore higher-level techniques that employ "learning" algorithms or predictive methods have been employed [8,9,10,11,12,13,14]. The best results for such systems have been achieved by employing training sets that have been segmented "by hand" [3].

Most of the DSP based musical stream segmentation techniques are energy based, i.e., the energy of the signal, computed as a moving average in the time domain, is monitored for abrupt changes. Although this method may have good time resolution, typically it fails to detect a fraction of note onsets, such as slurred notes (notes of different pitches tied together) and grace notes (very short notes immediately preceding other notes) due to the small or rapid energy changes in such events. Energy-based techniques also may lead to false onset detections due to musical features such as vibrato and tremolo, see Figure 1. Thus the choice of the threshold in energy-based methods is critical and a major shortcoming of such methods is the strong dependence of the results on the choice of the threshold value. Finally, energy-based methods employ a moving average to smooth rapid fluctuations of the energy, which may have the unintended effect of shifting the inferred note onset times by an amount that is dependent upon the length of the averaging window and the detailed structure of the musical signal.

In musical events such as slurred and grace notes in which the energy does not change significantly there may be a significant frequency change, indicating the presence of a separate note. Furthermore, when vibrato is present the energy may fluctuate significantly without a change in frequency. Therefore, an optimal note segmentation method should combine evidence from both the time and the frequency domains.

In the spectrum of a musical tone normally the energy is distributed among a set of partials that are harmonically related to a fundamental frequency that determines the note's pitch. Often the largest concentration of energy may be in one of the harmonic overtones, even though the assigned pitch is that of the fundamental, and the distribution of energy among the harmonics may vary over the duration of the note. Such changes, which are described as changes in timbre, may be mistaken for changes in the note frequency by simply tracking the largest spectral component. Thus, we have included the following check for "harmonicity" as follows. If the frequency containing the greatest energy changes between frames then the frequency of the new maximum is compared to the frequency of the maximum in the previous frame to check if they are related harmonically. If so, then the difference may be ascribed to a change in timbre.

Good frequency resolution is critical to enable discrimination of tremolo and mistuned notes from the onsets of new notes. However, high frequency resolution comes at the price of reduced discrimination of energy changes due to the averaging effect of the long time frames required. Frequency interpolation may be achieved by zero padding the signal, however this requires greater processing time. To obtain pitch resolution better than a semi-tone over the full range of pitches we employed a frequency resolution of 2-4 Hz, which for our choice of STFT overlap gave a temporal resolution of 10-20 msec. If the note onset time were to serve simply as an indicator that a new note has begun then this time resolution would be sufficient. However, for the purposes of musical style analysis and assessment of the effects of latency on musical performance, greater time resolution was required.

Our goal is to accurately measure note onset times, which leads to the question - when, exactly, does the note begin? Each separate note has attack, sustain and release phases, the details of which depend on the musical instrument and the performance technique. However, in a musical stream each note may not be well defined. For example in very short duration notes, such as grace notes, the sustain portion may not be evident, with the rise time comprising the largest fraction of the note, see Figure 4. Furthermore, the energy may not drop to zero between notes or the duration of the attack may vary. Finally, the perceived point of attack depends upon the time resolution of the human auditory system. The question then becomes, to which point in the attack phase of the note should we assign the note onset? For simplicity we assign the note onset time to the minimum of the time domain energy at the beginning of a detected attack.

2. THE SEGMENTATION METHOD

Note onsets are detected by a change in the energy, the frequency or both. The main features of the method are described in Figure 2.

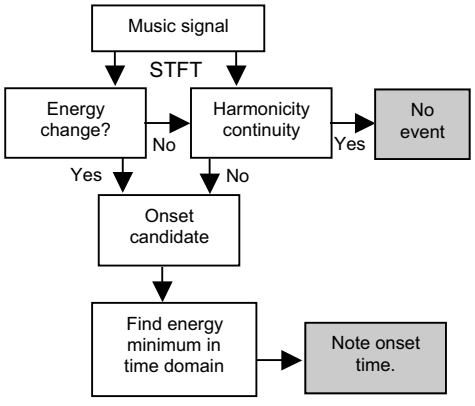

Figure 2. Block diagram of the onset detection method.

First, the STFT of the signal is computed employing an overlap to increase the time resolution. For each time frame of the STFT the harmonics containing the signal power are identified and the harmonicity (as defined above) is checked to look for continuity between

successive time frames. Discontinuity indicates a candidate note onset. At the same time the energy in each frame is found by summing the Power Spectral Density (PSD) computed by the STFT. A fixed energy threshold is applied to eliminate the very low energy portions of the signal corresponding to intervals of relative "silence" between notes. Then a "dynamic" energy threshold is applied, which searches for either rapid increase of the energy or a large gradual change in energy occurring over a longer time scale, such as would occur for notes with "soft" attacks. If a change of harmonicity is detected or if the dynamic energy threshold is reached, the event is marked as a candidate note onset.

The onset of a note is always preceded by an energy minimum. Furthermore, the energy computed in the time and frequency domains is the same (Parseval's Theorem). Thus, the final step in the procedure is to search for the absolute energy minimum (employing the time domain energy calculation) in the vicinity of each candidate onset time from the STFT based analysis. The energy minimum determined in this way indicates the note onset time with a resolution corresponding to the inverse of the sampling rate (about 0.2 msec in our case) - see Figure 3.

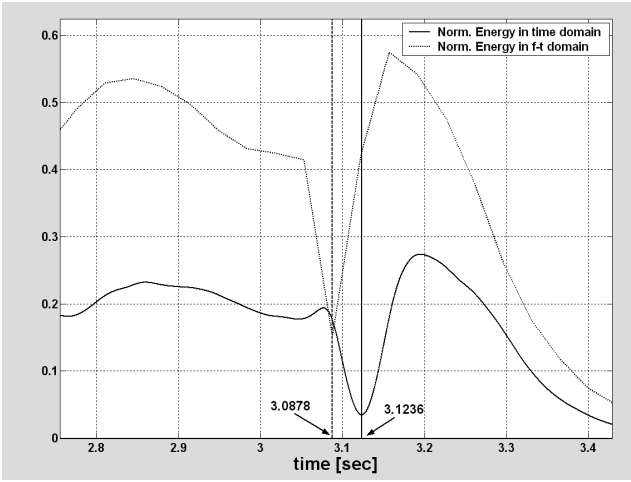

Figure 3. Timings for a note onset in both domains: there is a significant offset due to different resolutions.

Detection of note onsets by monitoring frequency changes alone may lead to false onset detection at the release of a note. Occasionally in the release phase, especially for wind instruments, there may be a slight change in frequency. When the STFT has poor frequency resolution (short time window), this may be interpreted as a new note. However, if the energy is monotonically decreasing such candidate events are discarded.

Rapid, steep amplitude changes, with continuous harmonicity are characteristic of bowed string instrument vibrato, which may lead to false candidate onsets. A temporal threshold applied only to this condition, vetoes such events. Although vibrato may have amplitude and frequency modulation components, narrow band AM and FM would appear similarly in a magnitude spectral plot, and thus our methods would not distinguish between them.

3. RESULTS

The method was applied to monophonic clarinet, piano and violin recordings of the same Mozart piece performed with various tempos and styles. The STFT was computed, using 85% overlap, for a FT length of 1024 (2048) samples and sampling rate of 4,410 samples/sec. This gave a time domain resolution of 0.2 msec and a frequency resolution of 4.3 (2.1) Hz, and a STFT time resolution of 17.5 (34.7) ms. The frequency resolution of 2-4Hz was sufficient to determine the pitch of the lowest pitched notes in the recordings to slightly better than a semi-tone.

In the clarinet and piano recordings the note onsets were detected with 100% accuracy, with no missed notes nor any false onset detections (197 and 153 note events were detected respectively). The accuracy of the violin recording was 95.5%. Out of 200 note events, 203 note candidates were detected, which 191 were correct note onsets with 11 false detections and 2 missed. In Figure 4 we show the results for the clarinet. Notice the correct detection of the grace note marked in the figure.

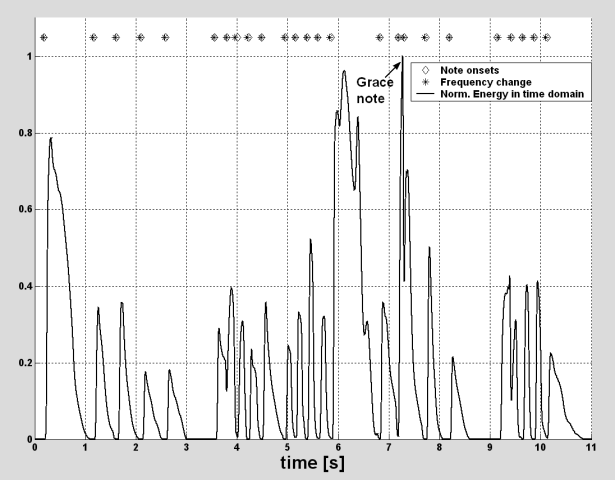

Figure 4. Segmentation results for the clarinet recording. Notice the detected grace note. On the same figure are examples of well separated notes (attack, sustain and release phases), and notes without a sustain phase.

4. CONCLUSIONS

The challenge of assessing transmission latency effects in distributed musical performance led us to develop a computer tool for time-precise note segmentation. Normal time-frequency representations, such as the STFT, have limitations for this application, as there is always the time-

frequency resolution trade off. Therefore we developed a technique in which we rely on both time and frequency domain analyses and combine the benefits of both domains. The STFT is used to coarsely locate the time of the note onsets and the frequency domain information from the STFT is employed in the onset identification task. Candidate note onsets are then examined in the time-domain to "zoom in" on the onset events. This allows note onsets to be measured to the resolution of the inverse sampling frequency. This technique for high time resolution audio segmentation may be applied to music, speech or other similar signals.

An effective music recognition system must combine multiple levels of analysis. In this paper we described a low-level, DSP-based, method that combines separate frequency and time domain analyses to unambiguously detect note onsets and to determine note onset times to milliseconds of accuracy. Although remaining ambiguities may be resolved at higher levels of analysis that employ expert knowledge or learning algorithms, this job is made easier by performing the most complete and reliable analysis possible at the low level.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation under award number IIS-0112689.

REFERENCES

[1] S. Dixon, "On the Computer Recognition of Solo Piano Music", Australasian Computer Music Conference, Brisbane, Australia, pp. 31 – 37, 2000.

[2] M. Marolt, A. Kavcic, M. Privosnik, "Neural Networks for Note Onset Detection in Piano Music", Proceedings of ICMC, Stocholm, Sweden, 2002.

[3] K. Jensen and D. Murphy, "Segmenting Melodies Into Notes", Proceedings of the DSAGM, Copenhagen, Denmark, 2001.

[4] C. Tait, W. Findlay, "Wavelet Analysis For Onset Detection", International Computer Music Conference, Hong Kong, 1996.

[5] A. Klapurri, "Sound Onset Detection by Applying Psyhoacustic Knowledge", Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing

[6] L.R. Rabiner, "On The Use Of Autocorrelation Analyses For Pitch Detection", IEEE Trans. ASSP, Vol. ASSP-25, No. 1, pp. 23-33, February, 1977.

[7] C. R. Janowski Jr., T. F. Quatieri and D. A. Reynolds, "Measuring Fine Structure In Speech: Application To Speaker Identification", Proc. ICASSP-95, pp. 325-328, May 1995.

[8] Klapuri A., "Automatic Transcription of Music", MSc. Thesis, Tampere University of Technology, 1998.

[9] P. de la Cuadra, A. Master, C. Sapp, "Efficient Pitch Detection Techniques for Interactive Music", Proceedings of ICMC 2001, International Computer Music Conference, La Habana, Cuba, September 2001

[10] J. P. Bello, G. Monti, M. L. R. Sandler, "An Implementation of Automatic Transcription of Monophonic Music with a Blackboard System", Proceedings of the Irish Signals and Systems conference (ISSC 2000), Dublin, Ireland, June 2000.

[11] P. J. Walmsley, S. J. Godsill, and P. J. W. Rayner. "Bayesian modelling of harmonic signals for polyphonic music tracking", Cambridge Music Processing Colloquium, Cambridge, UK, September 30 1999.

[12] W. Fong, S. Godsill, "Sequential Monte Carlo Simulation Of Dynamical Models With Slowly Varying Parameters: Application To Audio", Proc. ICASSP-02, pp. 1605 - 1608, May 2002.

[13] M. Davy, S. Godsill, "Detection Of Abrupt Spectral Changes Using Support Vector Machines An Application To Audio Signal Segmentation", Proc. ICASSP-02, pp. 1313-1316 , May 2002.

[14] S. Godsill, M. Davy, "Bayesian Harmonic Models For Musical Pitch Estimation And Analysis", Proc. ICASSP-02, pp. 1769 - 1772, May 2002.