TOWARDS IMPROVING ONSET DETECTION ACCURACY IN NON-PERCUSSIVE SOUNDS USING MULTIMODAL FUSION

Jordan Hochenbaum[1,2]

New Zealand School of Music[1]
PO Box 2332
Wellington 6140, New Zealand
hochenjord@myvuw.ac.nz

Ajay Kapur[1,2]

California Institute of the Arts[2]
24700 McBean Parkway
Valencia California, 91355
akapur@calarts.edu

ABSTRACT

In this paper we present a novel approach for improving onset detection using multimodal fusion. Multimodal onset fusion can potentially help improve onset detection for any instrument or sound, however, the technique is particularly suited for non-percussive sounds, or other sounds (or playing techniques) with weak transients.

1. INTRODUCTION AND MOTIVATION

Across all genre and styles, music can generally be thought of as an event-based phenomenon. Whether formal pitch relationships emerge in note-based music, or timbre-spaces evolve in non-note based forms, music (in one regard) can be thought of as sequences of events happening over some length of time. Just as performers and listeners experience a piece of music through the unfolding of events, determining when events occur within a musical performance is at the core of many music information retrieval, analysis, and musical human-computer interaction scenarios. Determining the location of events in musical analysis is typically referred to as *onset detection*, and in this paper we discuss a novel approach for improving the accuracy of onset detection algorithms.

There are many established approaches to detecting note onsets [1, 4–6, 8, 12]. For percussive sounds with fast attacks and high transient changes, algorithms in the time, frequency, magnitude, phase and complex domains have been established have proven to be accurate. The task of onset detection, however, becomes much more difficult when sounds are pitched or more complex, especially in instruments with slow or smeared attacks (such that of common stringed instruments in the orchestra).

The need for multimodal onset detection arose from collecting and analyzing data for long-term metrics tracking experiments. During initial observations of data collected from a performer improvising on a custom bowed instrument, we began to notice certain playing techniques in which onset detection algorithms could not accurately segment individual notes. As such, we began to explore other algorithms, and ultimately the multimodal approach presented in this paper.

Others have started to apply fusion techniques to the task of improving onset detection algorithms in recent years. Toh, et al. propose a machine learning based onset detection approach utilizing Gaussian Mixture Models (GMMs) to classify onset frames from non-onset frames [13]. In this work feature-level and decision-level fusion is investigated to improve classification results. Improving onset detection results using score-level fusion of peak-time and onset probability from multiple onset detection algorithms was explored by Degara, Pena, and Torres [3]. Degara and Pena have also since adapted their approach with an additional layer in which onset peaks are used to estimate rhythmic structure. The rhythmic structure is then fed-back into a second peak-fusion stage, incorporating knowledge about the rhythmic structure of the material into the final peak decisions.

While previous efforts have shown promising results, there is much room for improvement, especially when dealing with musical contexts that do not assume a fixed tempo, or that are aperiodic in music structure. Many onset detection algorithms also work well for particular sounds or instruments, but often do not generalize across the sonic spectrum of instruments easily. This is particularly true for pitched instruments, as demonstrated in the Music Information Retrieval Evaluation eXchange (MIREX) evaluations in recent years [9–11]. Added complexity also arises when trying to segment and correlate individual instruments from a single audio source or recording. These scenarios and others can be addressed by utilizing multimodal techniques that exploit the physical dimensionalities of direct sensors on the instruments or performers.

The reminder of the paper is as follows. In section 1.1 we first provide an overview of terms used in the paper, followed by a discussion on the strengths and weaknesses of performing onset detection on acoustic and sensor signals in 1.2. An overview of our system and fusion algorithm is provided in section 2, and finally, our thoughts and conclusions in section 3.

1.1. Definition of Terms

An onset is often defined as the single point at which a musical note or sound reaches its initial transient. To

further clarify what we refer to as the note onset, examine the waveform and envelope curve of a single snare drum hit shown in Figure 1. As shown in the diagram, the onset is the initial moment of the transient, whereas the attack is the interval at which the amplitude envelope of the sound increases. The transient is often thought of as the period of time at which the sound is excited (e.g. struck with a hammer or bow), before the resonating decay. It should be noted that it is often the case that an onset detection algorithm chooses a local maxima as the onset from within the detected onset-space during a final peak-picking processing stage. This corresponds with the peak of the attack phase depicted in Figure 1.

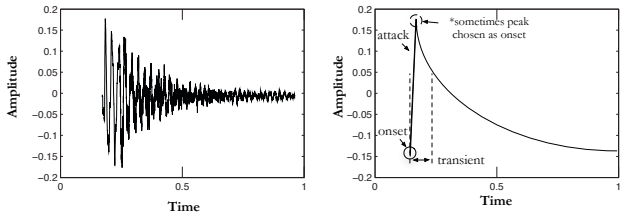

Figure 1 – Snare drum waveform (left) and envelope representation (right) of the note *onset* (circle), *attack* (bold), and *transient* (dashed)

1.2. Detection: Strengths and Weaknesses

There are many strengths and weaknesses that contribute to the overall success of audio and sensor based onset detection algorithms. The first strength of audio-based onset detection is that it is non-invasive for the performer. It is also very common to bring audio (either from a microphone or direct line input) into the computer and most modern machines provide built-in microphones and line inputs. This makes audio-based approaches applicable to a wide audience without the need of special hardware. In contrast sensors have often added wires that obstruct performance, they can alter the feel and playability of the instrument, or restrict normal body movement. For example, putting sensors on the frog of a bow could change its weight, hindering performance. In recent years however, sensors have not only become much more affordable, but also significantly smaller. Through engaging in communication with musicians during our experimental trials, we found that with careful consideration of placement and design, the invasiveness of instrumental sensor systems can be minimized such that the musicians do not notice that they are there. In fact, embeddable sensors like accelerometers and gyroscopes are already finding their way into consumer products beyond our cellphones, as demonstrated by the emerging field in wearable technology. The technologies are also beginning to appear into commercial musical instruments, and wireless sensing instrument bows already exist, such as the K-Bow from Keith McMillen instruments[1].

Another consideration between audio onsets and sensor onsets has to do with what information the onsets are actually providing. In the acoustic domain researchers have not only explored the physical onset times but the closely related perceptual onset (when the listener first hears the sound), as well as the perceptual attack time (when the sounds rhythmic emphasis is heard) [2, 14]. These distinctions are very important to make depending on the task, and when dealing with non-percussive notes, such as a slow-bowed stroke on a cello or violin (who's rhythmically perceived onset may be much later than the initial onset). This exposes a weakness in audio-based onset detection—which has trouble with non-percussive, slow, or smeared attacks. In this paper we propose a technique that uses sensor onset detection, which is capable of detecting slow, non-percussive onsets very well. This does not come without certain considerations, as described in greater detail in this section.

In the sensor domain the onset and surrounding data is often providing a trajectory of physical motion, which can sometimes even be correlated with the perceptual attack time. Sometimes however, the physical onset from a sensor might not directly align with the acoustic output or perceptual attack time, and so careful co-operation between onset-fusion is necessary. In learning contexts, this trajectory can provide a highly nuanced pipe into information about the player's physical performance. The data can directly correlate with style, skill level, the physical attributes of the performance, and ultimately the acoustic sound produced.

As shown later in this section, the differences in the information provided from separate modalities can actually be used strengthen our beliefs in the information from either modality individually. This helps overcome weaknesses in the modalities, such as the fact that a sensor by itself may not have any musical context (e.g. gesturing a bow in-air without actually playing on the strings). Combining information from both modalities can be used to provide the musical or other missing context from one modality for the other.

Additionally, while audio onset detection has proven to work very well for non-pitched and percussive sounds, they have increased difficulty with complex and pitched sounds. This can often be addressed with sensor onset detection, as the sensor may not be affected by (and do not necessarily have any concept of) pitch.

Lastly, many musical recordings and performances are outside of the practice room, and contain multiple instruments. This reality makes onset detection increasingly difficult as there is the additional task of segmenting instruments from either single stream, or from bleed in an individual stream, as well as ambient noise and

[1] http://www.keithmcmillen.com/

interference (e.g. clapping, coughing, door shutting, etc). As there is a great deal of overlap in the typical ranges of sound produced by traditional instruments, polyphonic sound separation is an extremely difficult task. Physical sensors however are naïve to other instruments and sensors other than themselves, and are normally not affected by other factors in the ambient environment. Thus they provide (in these ways) an ideally homogenous signal from which to determine, or strengthen onset predictions.

	Strengths	Weaknesses
Audio	• Non-invasive • Onset time can be close to perceptual attack time • No special hardware	• Algorithms have trouble with pitched and complex sounds • Algorithms have trouble with slow / smeared attacks • Ambient noise / interference • Source segmentation / non-homogenous recording
Sensor	• Very sensitive physical measurements and trajectories • Can detect onsets from slow / smeared attacks • Not negatively affected by pitched or complex sounds • Resistant to factors in the environment / ambience • Typically mono-sources, no separation necessary	• Can sometimes be invasive • No musical context • Onsets may or may not be related to the acoustic / auditory onsets

Figure 2 – Strengths and Weaknesses of Audio and Sensor Onset Detection

2. SYSTEM DESIGN AND IMPLEMENTATION

In designing this system, a primary goal was to create a fusion algorithm that could operate independently of any one particular onset detection algorithm. In this way, the fusion algorithm was designed such that it is provided two onset streams (one for onsets detected from the acoustic or audio stream of the instrument, and one from the sensors), without bias or dependence on a particular detection algorithm. The onset algorithms can be tuned both to the task and individual modalities (perhaps one onset detection function works best for a particular sensor stream vs. the audio stream). This also ensures compatibility with future onset functions that have yet to be developed. We propose multimodal onset fusion as a post-processing ("late-fusion") step. It does not replace, rather our approach improves, the robustness and accuracy of the chosen onset detection algorithm(s).

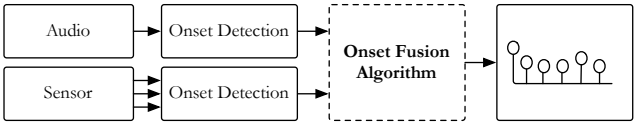

Figure 3 – General Overview of Multimodal Onset Fusion

2.1.1. Fusion Algorithm

A fusion function is provided two onset streams, one from the audio output and one for the sensor(s). First, the algorithm searches for audio onsets residing within a window (threshold) of each accelerometer onset. A typical window size is 30ms – 60ms and is an adjustable parameter called *width,* which effects the sensitivity of the algorithm. If one or more audio onsets are detected within the window of a sensor onset, our belief increases the best (closest in time) audio onset is considered a true onset; the onset is then added to the final output fusion onset list.

If a sensor onset is detected, however, no audio onset is found within the window (width), we have a 50/50 belief that the sensor onset is an actual note onset. To give a musical context to the sensor onset, the audio window is split into multiple frames, and spectral-flux and RMS are calculated between successive frames. The max flux and RMS values are then evaluated against a threshold parameter called *burst* to determine if there is significant (relative) spectral and energy change in the framed audio-window. Because onsets are typically characterized by a sudden burst of energy, if there is enough novelty in the flux and RMS values (crosses the burst threshold), our belief in the onset increases and the sensor onset time is added to the fusion onset list. The burst threshold is a dynamic value that is calculated as a percentage above the average spectral-flux and average RMS from the audio-window. By default burst is set to equal 20% increase in the average flux value, and a 10% increase in the average RMS from the current audio selection. Increasing or decreasing the burst threshold decreases or increases the spectral flux and RMS, changing the algorithms sensitivity.

OnsetAlgorithmPseucodeCode.cpp

OnsetAlgorithmPseucodeCode.cpp › No Selection

```
1  //  OnsetAlgorithmPseudocodeCode.cpp
2  //  2012 Jordan Hochenbaum
3
4  for each sensor onset
5  {
6      if audioOnset >= window-width and <= window+width
7      {
8          add closest audioOnset to fusion onset list;
9      }
10
11     else
12     {
13         calculate spectral flux over framed audio-window;
14         calculate rms over framed audio-window;
15
16         if max flux and rms > burst (thresholds)
17         {
18             add the onset to the fusion onset list;
19         }
20     }
21 }
22
```

Figure 4 – Multimodal Onset Detection Pseudocode

3. CONCLUSION AND FUTURE WORK

This ability to correctly differentiate musical events is at the core of all music related tasks. In this paper we have proposed a technique to improve the computers ability to detect note onsets. Our approach is a late multimodal fusion step, making it compatible with nearly any onset detection algorithm. Additionally, this allows the onset

detection algorithms to be selected and tuned to individual modalities, making the technique extremely acute to the particular scenario.

In our preliminary trials, we were able to successfully fuse not onsets from a performer bowing a custom string hyperinstrument. When playing in a tremolo style, we were not able to achieve satisfactory results using traditional audio based on detection alone. Performing onset detection from the gesture data provided from an accelerometer on the frog of his bow, we were able to detect most bow events, however, there were many false positives (as the bow provides a continuous stream, whether or not he was actually playing the instrument, or when he moved in between strokes). Using our fusion algorithm, we were able to synergize the accuracy of the sensor onset detection, with the musical context provided by the audio onset detection. We are currently conducting controlled experiments to quantify the results. In the future, results could be further improved by tailoring the fusion parameters more specifically to the data being analyzed. There are many ways to do this, both by hand and dynamically, and we hope to explore these in the future. For example, dynamic range compression (DRC) during a pre-processing phase could help generalize certain parameters by reducing the amount of variance in the dynamic range of the input data, which changes from day to day and recording to recording. Additionally, there is a considerable room to experiment with the onset detection function currently used, not only in terms of adjustable parameters, but also in using different onset detection functions that are tailored to exhibit better performance for a specific modality.

4. REFERENCES

[1] Bello, J. et al. 2005. A Tutorial on Onset Detection in Music Signals. *Speech and Audio Processing, IEEE Transactions on*. 13, 5 (2005), 1035–1047.

[2] Collins, N. 2006. Investigating Computational Models of Perceptual Attack Time. (2006).

[3] Degara-Quintela, Norberto et al. 2009. A Comparison of Score-Level Fusion Rules For Onset Detection In Music Signals. *10th International Society for Music Information Retrieval Conference* (2009).

[4] Dixon, S. 2007. Evaluation of the Audio Beat Tracking System BeatRoot. *Journal of New Music Research*. 36, 1 (Mar. 2007), 39–50.

[5] Dixon, S. 2006. Onset Detection Revisited. *Proceedings of the 9th International Conference on Digital Audio Effects (DAFX'06)* (Montreal, Canada, 2006).

[6] Goto, M. and Muraoka, Y. 1999. Real-time beat tracking for drumless audio signals: chord change detection for musical decisions. *Speech Commun.* 27, 3-4 (1999), 311–335.

[7] Lartillot, O. 2011. MIRtoolbox 1.3.4 User's Manual. Finnish Centre of Excellence in Interdisciplinary Music Research, University of Jyväskylä, Finland.

[8] Lartillot, Olivier et al. 2008. A Unifying Framework for Onset Detection, Tempo Estimation, and Pulse Clarity Prediction. *11th International Conference on Digital Audio Effects* (Espoo, Finland, Sep. 2008).

[9] MIREX Audio Onset Detection Evaluation Results: 2006. *http://www.music-ir.org/mirex/wiki/2006:Audio_Onset_Detection_Results*.

[10] MIREX Audio Onset Detection Evaluation Results: 2007. *http://www.music-ir.org/mirex/wiki/2007:Audio_Onset_Detection_Results*.

[11] MIREX Audio Onset Detection Evaluation Results: 2009. *http://www.music-ir.org/mirex/wiki/2009:Audio_Onset_Detection_Results*.

[12] Scheirer, E. 1998. Tempo and beat analysis of acoustic musical signals. *The Journal of the Acoustical Society of America*. 103, 1 (1998), 588–601.

[13] Toh, C. et al. 2008. Multiple-Feature Fusion Based Onset Detection for Solo Singing Voice. *Proceedings of the 9th International Society for Music Information Retrieval Conference (ISMIR'08)* (2008).

[14] Wright 2008. The Shape of an Instant: Measuring and Modeling Perceptual Attack Time with Probability Density Functions. March (2008), 202.

[1] Bello, J. et al. 2005. A Tutorial on Onset Detection in Music Signals. *Speech and Audio Processing, IEEE Transactions on*. 13, 5 (2005), 1035–1047.

[2] Collins, N. 2006. Investigating Computational Models of Perceptual Attack Time. (2006).

[3] Degara-Quintela, Norberto et al. 2009. A Comparison of Score-Level Fusion Rules For Onset Detection In Music Signals. *10th International Society for Music Information Retrieval Conference* (2009).

[4] Dixon, S. 2007. Evaluation of the Audio Beat Tracking System BeatRoot. *Journal of New Music Research*. 36, 1 (Mar. 2007), 39–50.

[5] Dixon, S. 2006. Onset Detection Revisited. *Proceedings of the 9th International Conference on Digital Audio Effects (DAFX'06)* (Montreal, Canada, 2006).

[6] Goto, M. and Muraoka, Y. 1999. Real-time beat tracking for drumless audio signals: chord change detection for musical decisions. *Speech Commun.* 27, 3-4 (1999), 311–335.

[7] Lartillot, O. 2011. MIRtoolbox 1.3.4 User's Manual. Finnish Centre of Excellence in Interdisciplinary Music Research, University of Jyväskylä, Finland.

[8] Lartillot, Olivier et al. 2008. A Unifying Framework for Onset Detection, Tempo Estimation, and Pulse

Clarity Prediction. *11th International Conference on Digital Audio Effects* (Espoo, Finland, Sep. 2008).

[9] MIREX Audio Onset Detection Evaluation Results: 2006. *http://www.music-ir.org/mirex/wiki/2006:Audio_Onset_Detection_Results*.

[10] MIREX Audio Onset Detection Evaluation Results: 2007. *http://www.music-ir.org/mirex/wiki/2007:Audio_Onset_Detection_Results*.

[11] MIREX Audio Onset Detection Evaluation Results: 2009. *http://www.music-ir.org/mirex/wiki/2009:Audio_Onset_Detection_Results*.

[12] Scheirer, E. 1998. Tempo and beat analysis of acoustic musical signals. *The Journal of the Acoustical Society of America*. 103, 1 (1998), 588–601.

[13] Toh, C. et al. 2008. Multiple-Feature Fusion Based Onset Detection for Solo Singing Voice. *Proceedings of the 9th International Society for Music Information Retrieval Conference (ISMIR'08)* (2008).

[14] Wright 2008. The Shape of an Instant: Measuring and Modeling Perceptual Attack Time with Probability Density Functions. March (2008), 202.