Studying the Timing of Discrete Musical Air Gestures

Luke Dahl

Computer Music Journal, Volume 39, Number 2, Summer 2015, pp. 47-66 (Article)

Published by The MIT Press

Luke Dahl
Composition and Computer Technologies
McIntire Department of Music
University of Virginia
P.O. Box 400176, Charlottesville, Virginia
22904, USA
lukedahl@virginia.edu

Studying the Timing of Discrete Musical Air Gestures

Abstract: Motion-sensing technologies enable musical interfaces where a performer controls sound by moving his or her body "in the air," without touching a physical object. These interfaces work well when the movement and resulting sound are smooth and continuous, but it has proven difficult to design air instruments that trigger discrete sounds with precision that feels natural to performers and allows them to play rhythmically complex music.

This article presents a study of "air drumming" gestures. Participants performed drumming-like gestures in time to simple recorded rhythms. These movements were recorded and examined to look for aspects of the movement that correspond to the timing of the sounds. The goal is to understand what we do with our bodies when we gesture in the air to trigger a sound. Two movement features of the hand are studied: Hits are the moment where the hand changes direction at the end of the striking gesture, and acceleration peaks are sharp peaks in magnitude acceleration as the hand decelerates. Hits and acceleration peaks are also detected for the movement of the wrist. It is found that the acceleration peaks are more useful than the hits because they occur earlier and with less variability, and their timing changes less with note speed. It is also shown that timing differences between hand and wrist features can be used to group performers into different movement styles.

Introduction

Most musical instruments require the musician's touch. When we strike a drum, bow or pluck a string, or blow into a flute, our body, in direct contact with the instrument, provides the energy that produces sound. Even in instruments where the acoustic energy is not provided by the player—such as most organs or the majority of electronic and digital instruments—control of the instrument relies on direct manipulation of a key, slider, rotary knob, etc.

With the advent of electrical sensing it became possible to control an instrument with gestures "in the air," without manipulating or physically contacting the instrument. Early examples include the Theremin in 1919 (Glinsky 1992), which is controlled by empty-hand movements in space, and later the Mathews (1990) Radio Baton and Buchla Lightning (Rich 1991), which sense the movement of handheld batons.

The recent proliferation of affordable motion-sensing technologies, such as the Microsoft Kinect or hand-held devices containing inertial sensors, has led to a surge in new "air-controlled" musical interfaces, where performers control sound by moving their bodies freely in space. These interfaces seem to work well when the movement and control of sound are smooth and continuous. From my own experience and observations of others' work, however, it has proven difficult to heuristically design a system that will trigger discrete sounds with a precision allowing for a rhythmically complex performance. To the performer of such systems the timing of the resulting sound often feels wrong.

This article describes research into what people do when they want to trigger a sound with a sharp onset, such as a percussive sound, using gestures of their arms in free space. The goal is to understand the nature and timing of these movements, and to use this knowledge to design more rhythmically precise gesture-controlled musical instruments.

Air Gestures

I define *instrumental air gestures* as purposeful movements performers make with their bodies in free space to control an immediately responsive sound-generating instrument. *Discrete air gestures* are meant to trigger a sonic event at a precise time, and are contrasted with *continuous air gestures* in which some movement quality (e.g., the height of the hand) is continuously mapped to some sonic variable (e.g., a filter cutoff frequency).

Computer Music Journal, 39:2, pp. 47–66, Summer 2015
doi:10.1162/COMJ_a_00298

In popular usage, *air drumming* refers to miming the gestures of a percussionist in time to another musician's (usually prerecorded) performance. I call this typical case *mimetic air drumming*. For the sake of this research, air drumming can be expanded to include *productive air drumming*, i.e., discrete instrumental air gestures in which a performer mimics the striking of an imaginary surface to trigger a sound with a sudden attack. Air drumming is not the only type of discrete air gesture. For example, jerky movements such as in the dance style known as "popping and locking" might also be used to trigger percussive sounds.

Motivation and Overview

The ultimate aim of this research is to improve the design of air-controlled instruments so that discrete air gestures can be used to reliably trigger sounds with timing that feels natural to the performer. Most systems that enable discrete air gestures use some heuristic or arbitrary measure to decide when to trigger the sound. I contend, however, that when performers make such a gesture, they have a sense of when the sound should occur. If we knew when this moment occurs, and what in the performer's gesture it corresponds to, we would design our air instruments accordingly. In other words, I assume that when people make discrete air gestures, they do something with their body to create an internal sense of a discrete event, and that they intend this event to correspond to the sonic event of a drum sound. I want to know what this something is, and to characterize its timing with respect to the sonic event.

To this end, I conducted a study of mimetic air drumming, in which participants gesture in time to a simple prerecorded rhythm. Participants' movements were recorded in a motion capture system, and these data were analyzed to address the following question: What aspect of the air drummer's movement corresponds to the sound?

I examined two candidate movement features and analyzed their timing with respect to the onset time of the corresponding drum sound. The *hit* is the moment where the hand suddenly changes direction at the end of a strike gesture. When striking a real drum the hit would occur when the hand (or drumstick) contacts and rebounds off the drumhead, and the hit and the resulting sound would occur at the same moment. For air drumming, however, there is no physical object to limit the hand's travel, and it is not clear that the moment of the hand's direction change, or hit, necessarily corresponds to the time at which the air drummer intends the sound to occur.

The second movement feature is an *acceleration peak*, which occurs before the hit as the hand decelerates. Accelerations can be due to muscular effort, and thus these peaks may better correspond to the air drummer's internal sense of when the sound should occur.

I also detect the hits and acceleration peaks of the movement of the wrist in space. The timing of these features with respect to the hand features is analyzed, and I show how the timing differences between the hand and wrist features can be used to group people according to movement style.

The participants in my study mimed in time to a prerecorded musical performance. They were asked to move as if they were performing the sounds they heard. Thus I assume that the correspondence between gesture and sound would be the same if they were triggering the sounds themselves, and I expect the results of my analysis to also reliably describe productive air drumming and be useful in improving the timing of air instruments.

Related Work

A number of systems for playing musical sounds from discrete air gestures have been described in the literature, as has research on gestures for drumming (for both real and air instruments) and conducting. Findings from the field of sensorimotor synchronization are also relevant.

Discrete Air-Gesture Performance Systems

Real-time air instruments have used a variety of techniques for triggering sounds from movement. The Mathews (1990) Radio Baton senses the spatial location of two handheld wands. To trigger discrete sounds, it uses a spatial height threshold that corresponds to the height at which the baton contacts the surface of the sensor, giving the user tactile feedback. So, although the Radio Baton senses continuous air gestures, the discrete events that it affords are not air gestures.

Havel and Desainte-Catherine (2004) track sensors on the ends of drumsticks and use a velocity threshold to trigger sounds. Kanke and colleagues (2012) use data from acceleration and gyroscopic sensors in drumsticks to differentiate between striking a real percussion instrument and an air instrument. Strikes are registered when the acceleration exceeds a threshold.

In an earlier paper (Dahl and Wang 2010) I describe *Soundbounce*, a piece for mobile phone orchestra in which sound parameters are controlled by virtual balls that performers bounce upwards with flicking gestures of their arms and wrists. These "hits" were detected when low-pass-filtered acceleration data, which we can think of as approximating velocity, exceeded a threshold. That this instrument, although fun to perform and observe, was difficult to control precisely, was one inspiration for the research described here.

Studies of Discrete Air Gestures

A few studies of discrete air gestures have been conducted. Mäki-Patola (2005) studied participants striking a virtual drum surface in time to a metronome click. They compared the use of a physical stick held in the hand with a virtual stick. Clap sounds were triggered when the tip of the stick first intersected a horizontal virtual drum surface at a specific height. Among the findings was that hits lagged behind metronome clicks by 20 msec, which Mäki-Patola attributes to the "perceptual attack time" of the clap sound that was used.

Collicutt, Casciato, and Wanderley (2009) compared four cases of drumming: on a real drum, on an electronic drum pad, with the Radio Baton, and with the Buchla Lightning II. In all cases they tracked the height of the hand (even though their participants held sticks), and used vertical minima to determine when strikes occurred. They note that this did not work for one participant whose strikes corresponded to smaller minima before the actual minimum, however. They found that the Lightning, the only true air instrument, had the second best timing variability. They attribute this to the different way in which users control their movements when there is no tactile feedback.

Studies of Real Drum Gestures

Dahl (2004) made motion-capture recordings of drummers playing a simple rhythm on a real snare drum. She found that subjects raised the stick higher in preparation for accented strikes, and that preparatory height correlated with higher peak velocity. Drum strikes were detected as points in time that satisfy two criteria: The local minima of stick tip height must pass below a threshold, and the difference between two subsequent changes in vertical velocity (proportional to the third derivative of position, also known as jerk) must surpass a threshold.

Conducting Gestures

Conducting gestures for keeping time can be considered noninstrumental, discrete musical air gestures. Luck and Toiviainen (2006) studied how the beats of a musical ensemble synchronize to the conductor's gestures. They found that musical beats correlate closely with periods of maximal deceleration along the path made by the tip of the conductor's baton, with the musical beats lagging behind the gesture by about 30 msec. They also found that the vertical velocity of the baton also correlated highly with the musical beat, but at a larger lag of around 77 msec.

Sarasúa and Guaus (2014) wanted to create interactive musical systems controlled by conducting gestures. Their work, which took place concurrently with mine, uses a methodology similar to the one I use. They record mimetic conducting gestures (i.e., conducting to prerecorded music) to understand productive conducting gestures. They find that peaks in acceleration, which for their gestures also correspond to vertical minima, tend to occur before musical beats, and that participants differ with respect to the average lag.

Figure 1. Mimetic air drumming recorded in the lab.

Figure 2. Productive air drumming in performance.

Sensorimotor Synchronization

Air drumming in time to music is a form of synchronizing movements to sound. Research into sensorimotor synchronization goes back decades (see Repp 2005 for a review). One of the primary findings is that when tapping in time to an audible beat (usually a metronome click), most people tap before the beat. This "negative mean asynchrony" is often a few tens of milliseconds, but may be as great as 100 msec.

Research in sensorimotor synchronization is relevant to the current study because I assume that, like the subjects in tapping experiments, the air drummers are synchronizing some physically embodied sensation to the beat. The finding of negative asynchrony suggests that the physical event (the tap or the internally sensed movement event), although seeming to the performer to be identical to the sonic event (the metronome click or drum sound), may not necessarily occur at the same time as the sonic event.

Studying Air Drumming

The goal of this research is to better understand productive air-drumming gestures, i.e., drum-like gestures in free space that are used to trigger sounds in some real-time system. It is difficult, however, to study this behavior directly in a way that would give us access to the data needed. We want to know what aspects of the gesture correspond to the intended timing of the sound, but we don't have a way of knowing the time at which the air drummer intends the sound to occur.

The Study

To get around this difficulty, I use mimetic air drumming as a proxy for productive air drumming. Instead of recording people generating sounds from gestures, I record people gesturing in time to predetermined, known drum sounds, and they are asked to gesture as if they are generating the sounds. With this methodology the recorded drum sounds, whose timing I know precisely, become the "ground truth" against which the gestures can be analyzed.

To illustrate my methodology, Figure 1 shows mimetic air drumming in which a computer plays a sound, the participant gestures in time to that sound, and the gestures are recorded and stored for later analysis. In my study I record mimetic air drumming in the lab, but my goal is to improve real-time productive air-drumming systems, shown in Figure 2, in which a performer makes a gesture,

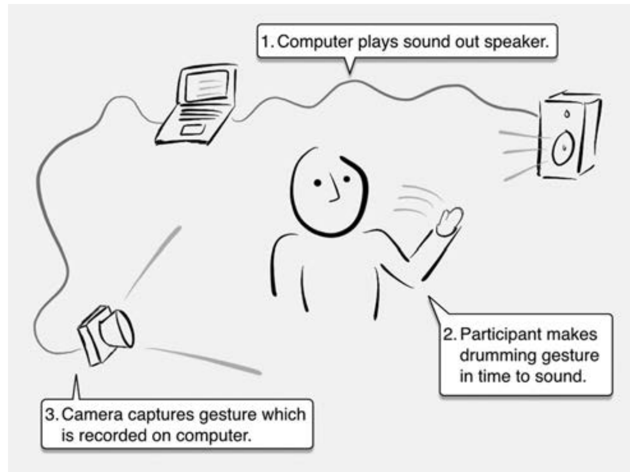


Figure 1

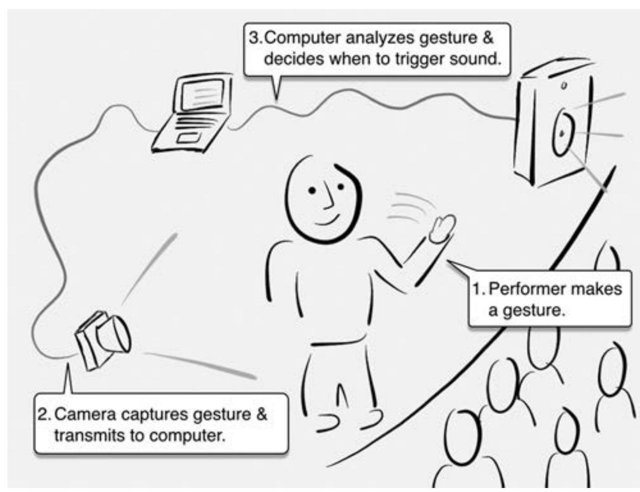
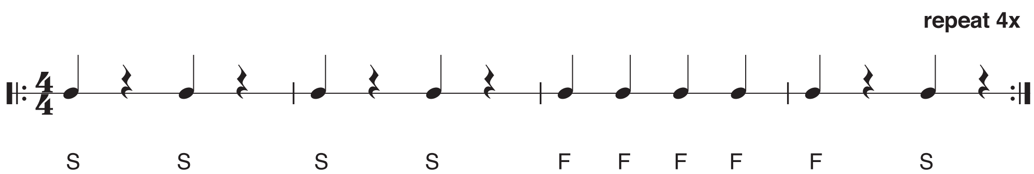


Figure 2

Figure 3. The stimulus rhythm. Slow notes are labeled “S” and fast notes “F.”



which is captured and analyzed in a computer which detects the correct moment to play a sound.

The Task

I recorded the movements of people performing air-drumming gestures in time with the simple rhythm described in Figure 3. They were asked to gesture as if striking a drum somewhere in front of them with a closed, empty right hand, and to act as if they are performing the sounds they hear. Because I am interested in gestures someone might make while performing an air instrument in free space, I did not provide further specification as to the location or style of the strike.

Air-drumming gestures may differ when performed at different repetition speeds, and so the rhythm is designed to have an equal number of “slow” notes (quarter notes with rests in between), and “fast” notes (quarter notes with no rests). The rhythm was played by the sound of a synthesized tom-tom at a tempo of 100 beats per minute (where a beat is one quarter note.) At this tempo both note speeds have inter-onset intervals less than 1.8 sec. According to Paul Fraisse, for intervals longer than 1.8 sec the “perception of time” gives way to the “estimation of time,” and the negative asynchrony disappears as people are no longer able to reliably synchronize their movements to the sound (see the review by Clarke 1999).

For each trial, participants performed the rhythm four times in succession without stopping. Two trials were recorded for each participant, resulting in a total of eight repetitions of the rhythm. A four-beat click was used to start each trial.

A video of air-drum stimulus, available on the *CMJ* Web site (www.mitpressjournals.org/doi/suppl/10.1162/COMJ_a_00298/AirDrum-StimulusAnimation.mp4, shows an animation of Subject 1’s movements, for the first two repetitions of the rhythm along with the stimulus that she heard.

In addition to this task (i.e., gesture in time to a prerecorded rhythm), participants were given a second task, where they were asked to vocalize the rhythm while gesturing. This task did not provide meaningfully different results (see Dahl 2014), and is not discussed further here. Participants also performed a third task, in which they performed air drumming to a rhythm with different dynamic levels. These data will be analyzed in future research.

Participants

Ten participants were recruited with the requirement that they have some experience playing a musical instrument and that they be able to read notated music. They were five women and five men, ranging in age from 22 to 57 years, with a median age of 23.5 years. All were right-handed. They reported between 13 and 48 years of musical experience, with a median of 16 years. Four had some formal dance training (between three and seven years). One (Participant 10) was a percussionist. Before recording, it was verified that each participant could read the simple rhythm and perform the desired task.

Equipment

Participants were outfitted with 14 reflective markers on their right arm and upper torso (see Figure 4), and their movements were recorded at 200 frames per second by a Motion Analysis motion capture system with twelve cameras mounted around the participant.

Participants could read the rhythm on a music stand one meter to their front right. The stimulus rhythm was played by a MIDI clip in Ableton Live

Figure 4. A participant with markers and motion capture cameras. (The participant's identity has been obscured.) The markers used for tracking the location of the hand (RMCP3 on the right third metacarpal) and wrist (RFAradius and RFAulna on either side of the distal end of the right forearm) are labeled. Markers on the first right metacarpal, the medial side of the elbow, and the back of the neck are not visible.

and presented over a Behringer MS40 studio monitor placed a meter to the front left of the participant. The stimulus was initiated at the beginning of each trial by the experimenter. Stimulus sounds were recorded into the motion capture system at a sample rate of 20 kHz via an analog input and synchronously with the motion data, thus simplifying the alignment of audio and motion capture data for later analysis.

Movement Features

If we want to trigger sounds by tracking a single body location, it makes sense to start with the hand. In normal drumming, the hand strikes the drum or holds the stick that strikes the drum. Because the hand lies at the end of the kinematic chain, its movements are the result of movements of the torso and the shoulder, elbow, and wrist joints.

In this section I describe the calculation of two movement features: hits and acceleration peaks. The timing of these movement features will be compared with the timing of the drum sounds to which they are intended to correspond. Thus, the first stage of analysis is to determine the onset times of the drum sounds for each trial.

Detecting Audio Onsets

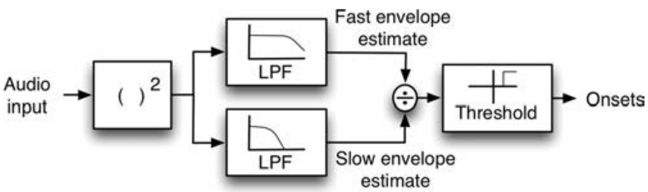
To detect audio onsets, the squared audio signal is passed in parallel through two DC-normalized one-pole low-pass filters. These are used to estimate two energy envelopes of the audio. One is "fast," with a time constant of 0.5 msec, and the other is "slow," with a time constant of 10 msec. When the ratio of the fast estimate over the slow estimate exceeds a threshold, a potential onset is registered (see Figure 5). Similar techniques have been used to detect the first arrival time of echoes in geophysical prospecting (Coppens 1985), and have been adapted for detecting audio onsets by Herrera and Kim (2014). Potential onsets for which the slow estimate is very low are removed (these are false events in the background noise), as are those that occur within 200 msec of an earlier onset (in order to keep only the first moment of attack). The times of the detected onsets are stored for later processing.

Detecting Hits

The first movement feature I examine is the end of the striking gesture, which I refer to as the hit. In a real drum strike, the hit would correspond to the moment when the drumstick hits the head of the drum, imparting energy to the instrument (thus initiating the sound), and rebounds in the direction from which it came.

For a striking gesture in free space, where no physical contact is made, where is the end of the strike? As Collicutt, Casciato, and Wanderley (2009) discovered, and as I found with my own data, the hit does not necessarily correspond to the moment when the minimum height is reached. Furthermore, participants' movements are not restricted to any particular plane or direction (they are instructed to act as if they are striking an invisible drum

Figure 5. An algorithm for detecting audio onsets. The squared audio signal is passed, in parallel, through a low-pass filter (LPF) with a high cutoff and another low-pass filter with a low cutoff. An onset is detected when the ratio of the resulting fast and slow estimates of the audio envelope exceeds a threshold.



"somewhere in front of them"). Thus I define a hit as the moment at the end of a striking gesture, where the hand suddenly changes direction.

To that end, I designed a sudden-direction-change detector. The design takes inspiration from the onset detector described earlier, which compares slow and fast estimates of audio energy. The sudden-direction-change detector uses a slow and fast estimate of the hand's 3-D velocity vector. The intuition is that during a sudden change of direction, the slow estimate will lag behind the quickly reacting fast estimate, and the angle between these two estimate vectors will be large. Upon inspecting the data I found that the moment I believed was the hit most reliably corresponded to a positive peak in the rate of change of this angle (see Figure 6).

The following is a detailed description of the sudden-direction-change detector:

1. Extract the position of the hand from the motion-capture data, using the marker RMCP3 on the back of the hand at the base of the middle finger (see Figure 4). The position is represented as (x, y, z) coordinates over time, where x is the direction in which the participant is facing and z is upward.
2. Smooth the position data in each dimension by approximating each point as a weighted least-squares quadratic fit of the point and its seven neighbors on either side.
3. Calculate the 3-D velocity vector of the hand, $\mathbf{v}_{hand}$, as the first difference of the smoothed hand position.
4. Create two smoothed versions of the velocity vector by passing it through two "leaky integrators" (i.e., DC-normalized one-pole low-pass filters). One, $\mathbf{v}_{slow}$, has a time constant of 100 msec, and the other, $\mathbf{v}_{fast}$ has a time constant of 5 msec. These are implemented as recursive filters on the 3-D velocity vector according to the following difference equations:

$$\mathbf{v}_{fast}[n] = (1 - a_{fast})\mathbf{v}_{hand}[n] + a_{fast}\mathbf{v}_{fast}[n-1]$$

$$\mathbf{v}_{slow}[n] = (1 - a_{slow})\mathbf{v}_{hand}[n] + a_{slow}\mathbf{v}_{slow}[n-1],$$

   where a_{slow} and a_{fast} are the pole locations corresponding to the slow and fast time constants.
5. At each time point n, calculate the angle θ between $\mathbf{v}_{slow}$ and $\mathbf{v}_{fast}$:

$$\theta[n] = \cos^{-1}\left(\frac{\langle \mathbf{v}_{slow}[n], \mathbf{v}_{fast}[n]\rangle}{\left\|\mathbf{v}_{slow}[n]\right\| \cdot \left\|\mathbf{v}_{fast}[n]\right\|}\right)$$

6. Calculate θ_{slope} as the first difference of θ.
7. Find all peaks of θ_{slope} that exceed a threshold. The times of these peaks are considered as moments when the hand changed direction, and they are stored as candidate hit times.

The next step is to find the change of direction associated with each strike gesture. The following algorithm is used to find the hit for each audio onset:

1. Because a hit occurs after a fast movement of the hand, find all peaks of the magnitude hand velocity, $\|\mathbf{v}_{hand}\|$, that exceed a threshold.
2. For each of these peaks, find the next candidate hit time (i.e., a large peak in θ_{slope} as described earlier).
3. To prevent choosing changes of directions that occur after a preparatory upwards movement, remove hits for which the distance between the hand and the shoulder is less than a threshold.
4. For each audio onset, find the hit candidate that is closest in time and store this as the hit time for that onset.

Does this method find the correct moment where a hit occurs? There is no way to know for sure, because the hit does not exist in any objective sense. That is, we have no ground truth. Figure 7 shows the detected hit time for a slow note by one

Figure 6. An algorithm for detecting sudden changes of hand direction. The velocity vector of the hand is calculated as the first time difference of the hand-position vector. Parallel low-pass filters (LPFs) produce fast and slow estimates of the velocity vector. The angle between these estimates is calculated, and peaks in the first time difference of this angle are detected as changes of direction.

Figure 7. Detecting the hit for a strike gesture.

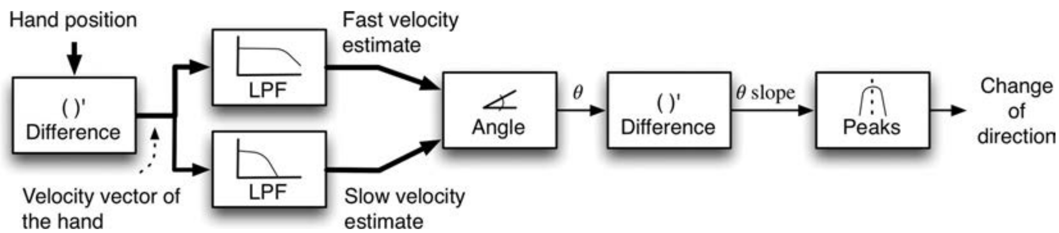


Figure 6

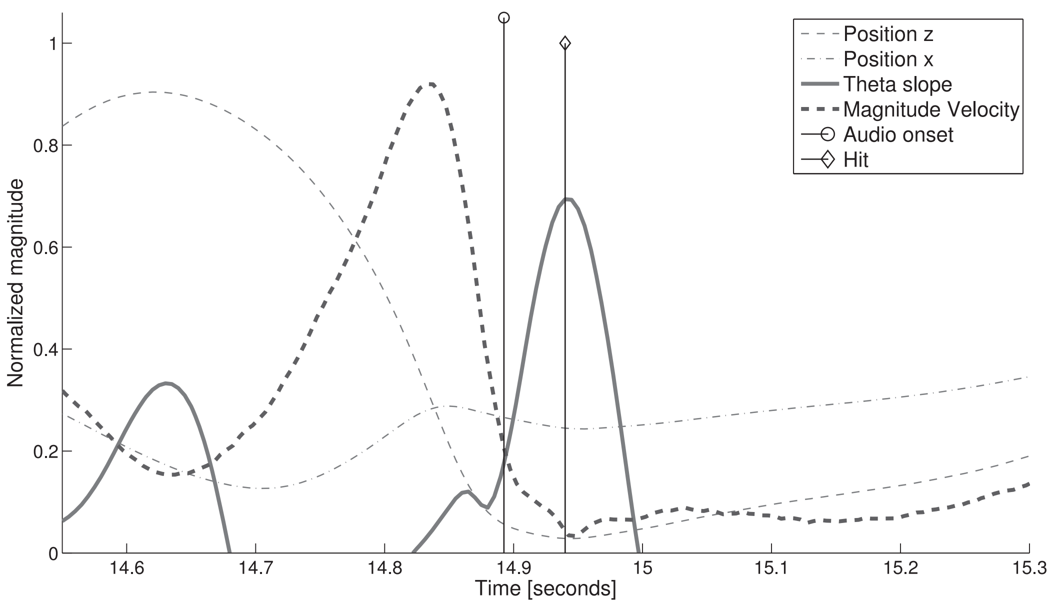


Figure 7

participant. The striking gesture happens primarily in the x and z directions, and we can see that the hit happens at extrema in both these dimensions. That the hit coincides with a distinct minimum in the magnitude velocity of the hand validates this choice.

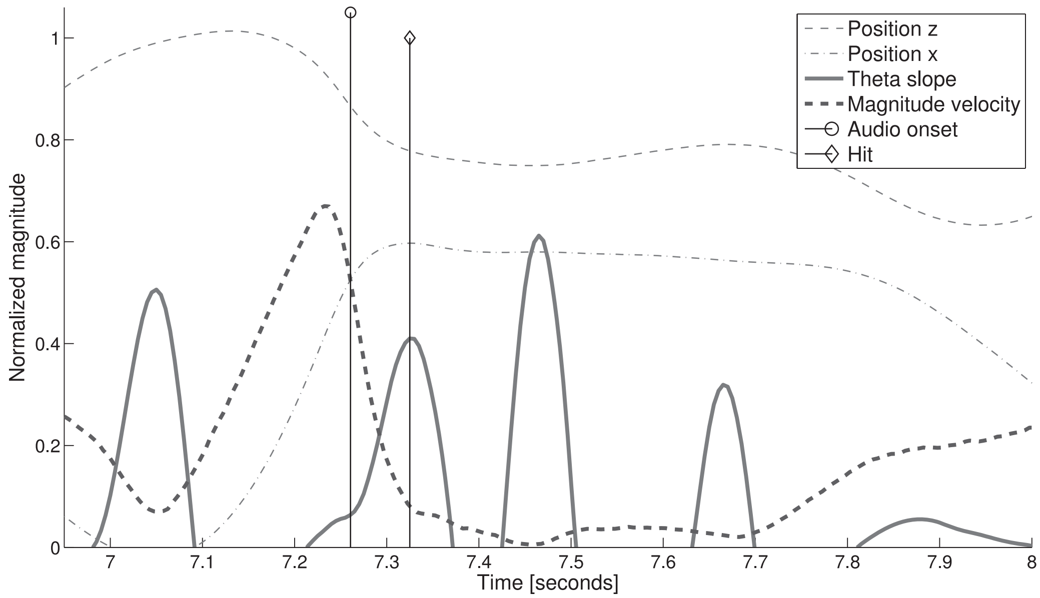
The striking gesture of another participant, shown in Figure 8, is more complex, however. This participant tended to add a short hook to the end of her strike. This is detected as multiple direction changes (large peaks in θ_{slope}). The algorithm chooses the first such peak after a large magnitude velocity, which corresponds to an extrema in the x direction and a sudden change of slope in magnitude velocity.

Detecting Acceleration Peaks

While examining the data, I noticed that large peaks in the magnitude acceleration often occur close to the audio onset. For an unimpeded movement, acceleration of the hand is the result of a muscular force, and so an acceleration peak may correspond to the internal movement event that air drummers create to correspond with the sound. (In fact, these peaks are decelerations as the participants sharply brake their strikes.) The following algorithm is used to pick the highest peak corresponding to each strike:

1. Calculate the acceleration vector, **a**, as the first difference of the hand velocity vector.

Figure 8. Detecting the hit for a more complex strike gesture.



2. Calculate magnitude of the acceleration vector, $\|\mathbf{a}\|$.
3. Look for times where $\|\mathbf{a}\|$ first exceeds a threshold and call these T_{up}.
4. For each T_{up}, find the next point where $\|\mathbf{a}\|$ passes below a second threshold and call these T_{down}.
5. For each interval $[T_{up}, T_{down}]$, find the time of the highest peak in $\|\mathbf{a}\|$ and save this as a prospective acceleration peak.
6. For each audio onset, find the prospective acceleration peak that is nearest in time, and store this as the acceleration peak time for that onset.

Figure 9 shows the acceleration peak of the strike gesture for a slow note. We can see that it occurs much more closely to the audio onset than does the corresponding hit.

Timing Analysis of Hand Features

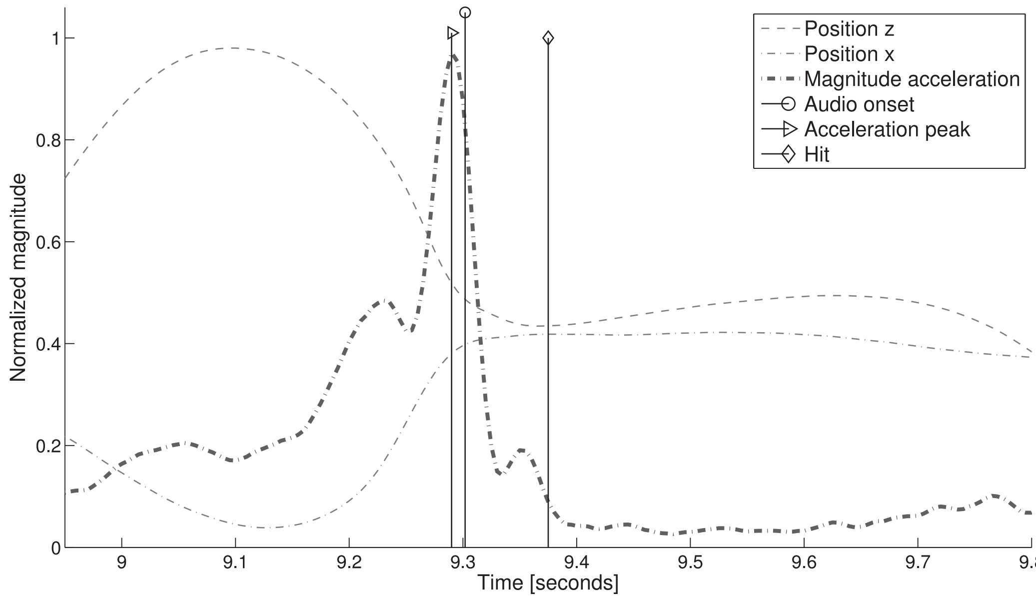
The audio onset times, hit times, and acceleration peak times were calculated for each note, as described earlier. From each hit time and acceleration peak time, the associated audio onset time is subtracted to get the time offset (or asynchrony) between the audio event (the onset of the sound) and the detected movement event (the hit or acceleration peak). A negative offset means the movement event preceded the audio event, and a positive offset means it came after. All subsequent analysis is performed on these offsets.

Computing Timing Statistics

Because there were two trials for each task, the data from each trial for each participant is aggregated and then split into the slow note and fast note conditions. For each participant this leads to a total of 40 events for each condition (five events per four-bar rhythm for each condition × four repetitions of the rhythm × two trials per task).

In order to reject bad data due to detector errors or participant mistakes, events whose offset is greater than half the time between notes (600 msec for slow notes, 300 msec for fast notes) are removed. Events that lie more than two standard deviations from the mean for each condition for each participant were rejected as outliers. This led to the removal of 21 slow hits, 21 fast hits, 18 slow acceleration peaks, and 23 acceleration peaks (out of 400 total for each case).

Figure 9. Detecting the acceleration peak for a strike gesture.



For the following results we want to know whether various conditions differ in the greater population. Specifically, we want to know wether the mean and standard deviations differ between conditions. For a given condition and movement feature, the mean tells us how much the movement feature tends to precede or lag behind the audio onset for that condition. The standard deviation tells us how much the timing of the performer's gestures varies randomly, and we can think of this as a measure of the "noise" in the mental and physiological processes that generate his or her movement.

To infer whether two conditions differ in the population, the mean (or standard deviation) of each participant's offset times for the conditions we wish to compare are computed. Then a two-sided paired-sample t-test of the ten participants' means (or standard deviations) for the two conditions is conducted.

For example, to compare whether the standard deviation of hit times is different between slow notes and fast notes, I first calculate the standard deviation of each participant's slow hits. I then calculate the standard deviations of each participant's fast hits. With these ten sample standard deviations for each condition I conduct a t-test with nine degrees of freedom.

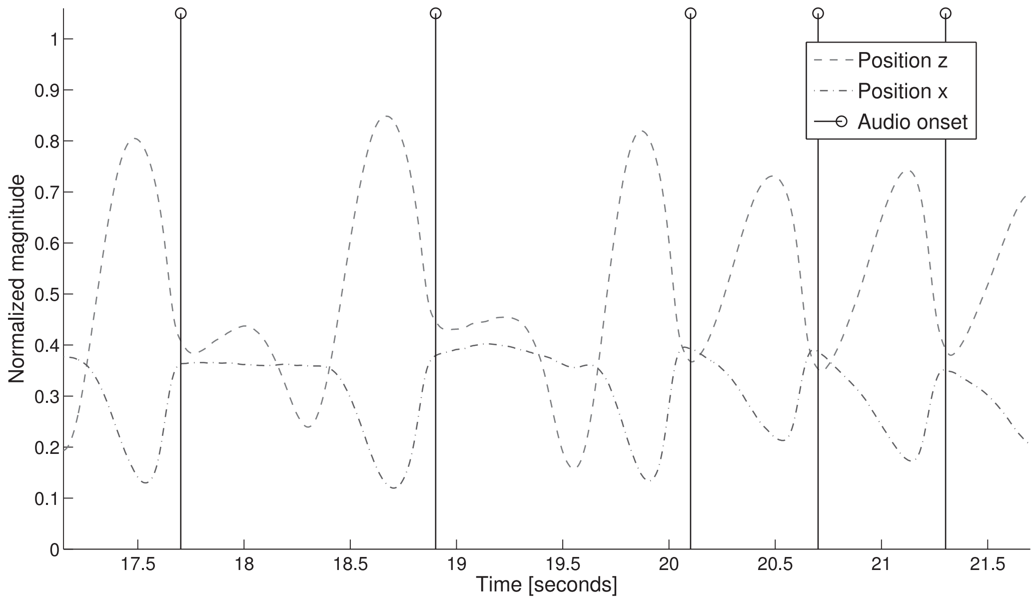
Effects of Note Speed

In Figure 10 we see that gestures for slow and fast notes are qualitatively different. For most participants, the gestures for slow notes have short pauses or bounces between them, whereas those for fast notes were simpler and more sinusoidal. It may be the case that the tempo of repeated discrete air gestures also affects the timing of their performance. If this is true, an air instrument needs to somehow take into account the tempo and rhythmic level of the intended notes.

To investigate whether the timing of gestures changes with note speed, I compared the offset times between fast notes and slow notes for both hits and acceleration peaks, to answer the following questions:

1. For hits, do slow and fast notes have different mean offsets? Yes, the mean offset times are significantly different ($t(9) = 4.5366$, $p = 0.0014$), with slow hits occurring 23.6 to 70.5 msec later than fast hits (this is the 95% confidence interval).

Figure 10. Position data for two slow notes and three fast notes.



2. For acceleration peaks, do slow and fast notes have different mean offsets? No, the difference between fast and slow notes that is seen for hits is not seen for acceleration peaks.

The mean offset time of hits changes with note speed, but the mean offset time of acceleration peaks does not. The implications of this finding will be discussed later.

These results suggest that, for repeated series of notes, the hit time may depend on the speed of repetition. Does this difference remain if we normalize offset times by the period of repetition (600 msec for fast notes, 1.2 sec for slow)? For hits the difference in mean time still exists, but it is only barely significant ($t(9) = 2.5420$, $p = 0.0316$), with fast hits preceding slow by 0.05 to 8.3 percent of a period. This normalization also affects acceleration peaks, which are now also barely significantly different ($t(9) = 2.2883$, $p = 0.0479$), with fast peaks preceding slow by 0.5 to 8.2 percent of a period.

Timing of Hits and Acceleration Peaks

Acceleration peaks mark sharp decelerations during the descent of the hand, and hits mark the time at which the hand changes direction. In order to find out whether these two movement features differ with respect to their offset times I use the data to answer the following questions:

1. Do hits and acceleration peaks have different mean offsets for slow notes? Yes ($t(9) = 4.8440$, $p = 9.1589 \times 10^{-4}$), with peaks preceding hits by 30.9 to 85 msec (with 95% confidence).
2. Do hits and acceleration peaks have different mean offsets for fast notes? Yes ($t(9) = 4.5294$, $p = 0.0014$), with peaks preceding hits by 15 to 44.8 msec (with 95% confidence).

Acceleration peaks do precede hits, and they do so by much more for slow notes than they do for fast notes.

Examining Wrist Movements

The instructions given to the participants—to gesture as if striking a drum somewhere in front of them—did not specify in detail how the gesture should be performed. Thus, it was observed that subjects differed with respect to the vertical range of the hand, the orientation of the hand, and the amount and timing of movement in the wrist joint.

The previous analyses were conducted using the location of the hand, whose motion is primarily the result of movements in the shoulder, elbow, and wrist joints, with possible small contributions from rotation and twisting of the trunk. To understand how wrist movements contribute to air-drumming gestures, hits and acceleration peaks were detected for the movement of each participant's wrist, where wrist location is tracked as the point half-way between the markers located on the medial and lateral sides of the wrist (markers RFAulna and RFAradius in Figure 4). The movement of the wrist joint can be studied by looking at the time difference between the hit time (or acceleration peak time) of the hand and that of the wrist.

Figure 11 shows histograms of all detected events for both the hand and wrist. A visual inspection reveals that, for most participants, the histograms for the hand and wrist tend to overlap for hits, whereas for acceleration peaks the hand and wrist histograms overlap less. Tests comparing the means of the hand and wrist event times across the test population produced the following results:

1. For slow note hits, the hand and wrist means have only a slightly significant difference $(t(9) = -3.1125, 0.0125)$, with hand hits preceding wrist hits by 1.5 to 8.9 msec (95% confidence interval).
2. For fast note hits, the hand and wrist have no significant difference in their means across the population.
3. For slow note acceleration peaks, the hand and wrist means are highly significantly different $(t(9) = -7.1416, p = 5.4159 \times 10^{-5})$ with hand peaks preceding wrist peaks by 9.8 to 18.9 msec (95% confidence interval).
4. For fast note acceleration peaks, the hand and wrist means are highly significantly different $(t(9) = -8.2435, p = 1.7406 \times 10^{-5})$ with hand peaks preceding wrist peaks by 10.3 to 18.1 msec (95% confidence interval).

These findings confirm what is observed in the histograms of Figure 11: In general, hand hits and wrist hits occur at about the same time, whereas acceleration peaks of the hand occur before the acceleration peaks of the wrist.

Grouping Participants by Movement Style

It may be that performers of air-drumming gestures tend to fall into one of a small number of movement styles, in which case case discrete air instruments might be improved by taking into account the particular movement style of the performer.

To explore this hypothesis, I used the k-means clustering algorithm to group participants. Each participant is represented as a four-dimensional vector consisting of the difference between the mean hand offset and the the mean wrist offset for the following four events: hits for slow notes, acceleration peaks for slow notes, hits for fast notes, and acceleration peaks for fast notes.

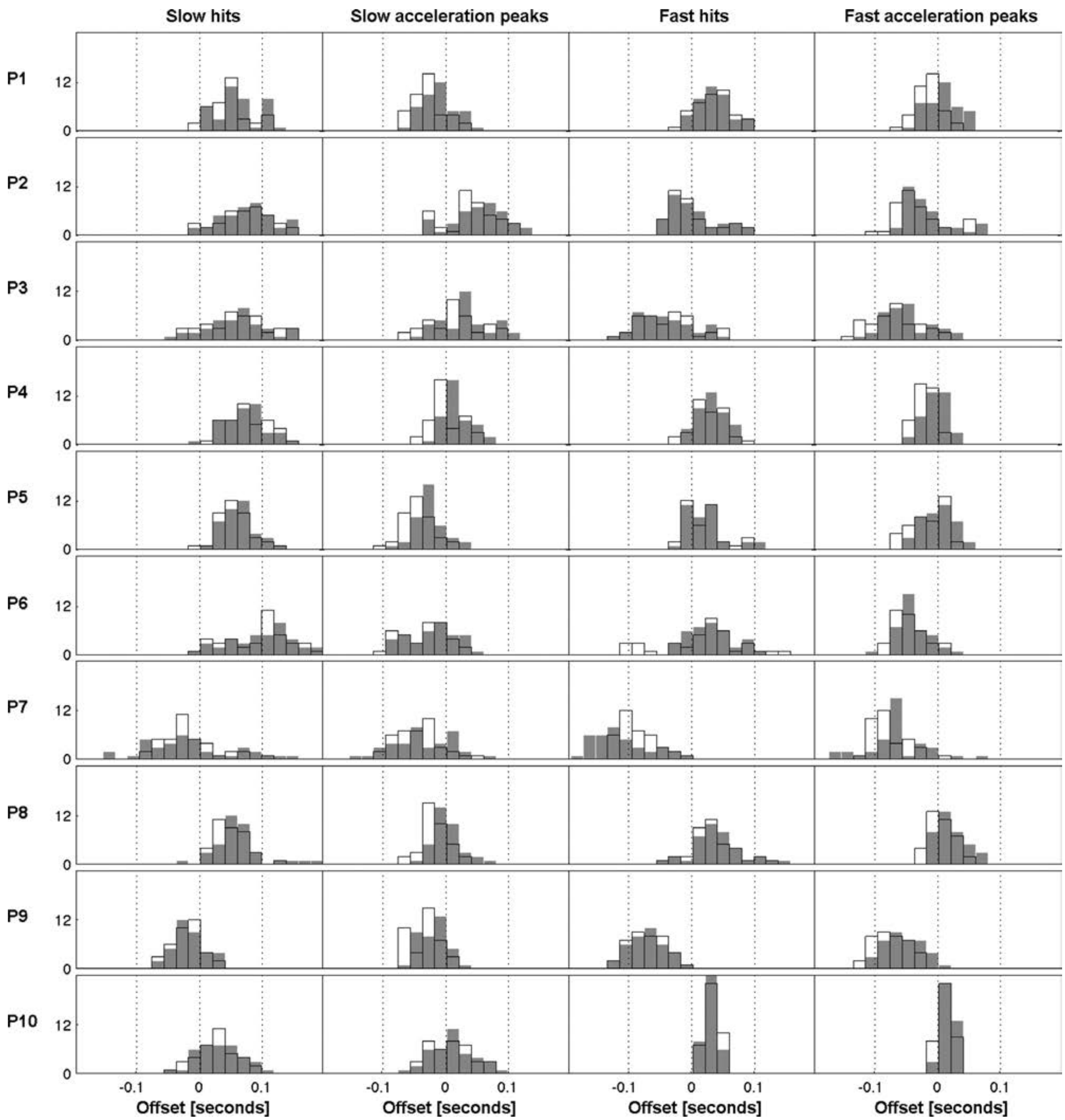
When the k-means algorithm was run with two clusters it returned one cluster with only Participant 7, and the second cluster with the other nine participants. Running the algorithm with three clusters returned a cluster with only Participant 7, another cluster with Participants 6 and 10, and a third cluster with the remaining seven participants. Table 1 shows the centroids for each of these clusters.

Initially, it was not clear from looking at the cluster centroids what a meaningful interpretation of these groupings might be. I then reviewed animations of the movement data for all ten participants, and noticed the following:

Participant 7, the sole member of Cluster 2, allowed movement in her wrist joint to occur only at the end of the strike, unlike all other participants whose wrists tended to move throughout the strike. In Table 1 we see that for slow notes, Participant 7's wrist hits came more closely after her hand hits, compared with the other groups. For fast notes Participant 7's hand hits came after the wrist hits, unlike any other group.

Participants 6 and 10, who make up Cluster 3, tend to gesture as if they are drumming holding a stick. They orient the palm of their hand almost vertically, whereas the members of Cluster 1 tend

Figure 11. Histograms of offset times of all detected events for the right hand and wrist for all ten participants. The hand histograms are transparent with black outlines, and wrist histograms are displayed behind them in gray. Time zero corresponds to the onset of the drum sound. For positive offset times the movement event came after the audio onset.



to orient their palms closer to horizontal, almost as if they were playing a hand drum (though their hands were closed, as instructed). The participants in Cluster 3 made wrist movements that had a much sharper "snap" than those of cluster one. This may be why Cluster 3 has much smaller delays between

Table 1. Clustering Participants

Cluster	*Participants*	*Slow Hits*	*Slow Acceleration Peaks*	*Fast Hits*	*Fast Acceleration Peaks*
1	1, 2, 3, 4, 5, 8, 9	−5.2	−17.7	−1.5	−17.2
2	7	−3.0	−4.7	+26.2	−9.7
3	6, 10	−6.1	−7.5	−6.7	−5.9

Participants are grouped into clusters according to the mean hand offset minus the mean wrist offset for each movement feature and note speed. The table shows the offset differences for each cluster's centroid. All times in milliseconds.

the acceleration peaks of the wrist and those of the hand than does Cluster 1.

Comparing Hands and Wrists

We now have four movement features: hand hits, hand acceleration peaks, wrist hits, and wrist acceleration peaks. Figure 12 shows box plots of the means for all four features.

Which Comes First?

Visual inspection of Figure 12 shows that for both the hand and wrist, and for both note speeds, the acceleration peaks occur, on average, before the hits. This appears to be true for both slow notes and fast notes, but the difference seems to be less for fast notes.

A repeated-measures two-way analysis of variance (ANOVA) was conducted on all offset means. The two within-subjects factors were movement feature, with three degrees of freedom, and note speed, with one degree of freedom. Significant effects are found for both movement feature ($F = 17.053$, $p = 2.057 \times 10^{-6}$) and note speed ($F = 10.911$, $p = 0.0092$). There was also a significant interaction between movement feature and note speed ($F = 15.726$, $p = 4.112 \times 10^{-6}$).

We already saw that for the hand, the difference between acceleration peaks and hits is significant at both note speeds. For the wrist two more post hoc t-tests reveal that:

1. For slow notes, wrist acceleration peaks precede wrist hits ($t(9) = 4.1889$, $p = 0.0023$) by 22.4 to 75 msec (this is the 95% confidence interval).
2. For fast notes, no significant difference was found between hits and wrist acceleration peaks of the wrist.

All three significant results (slow and fast notes for the hand, and slow notes for the wrist) pass a Bonferroni correction for four post hoc comparisons.

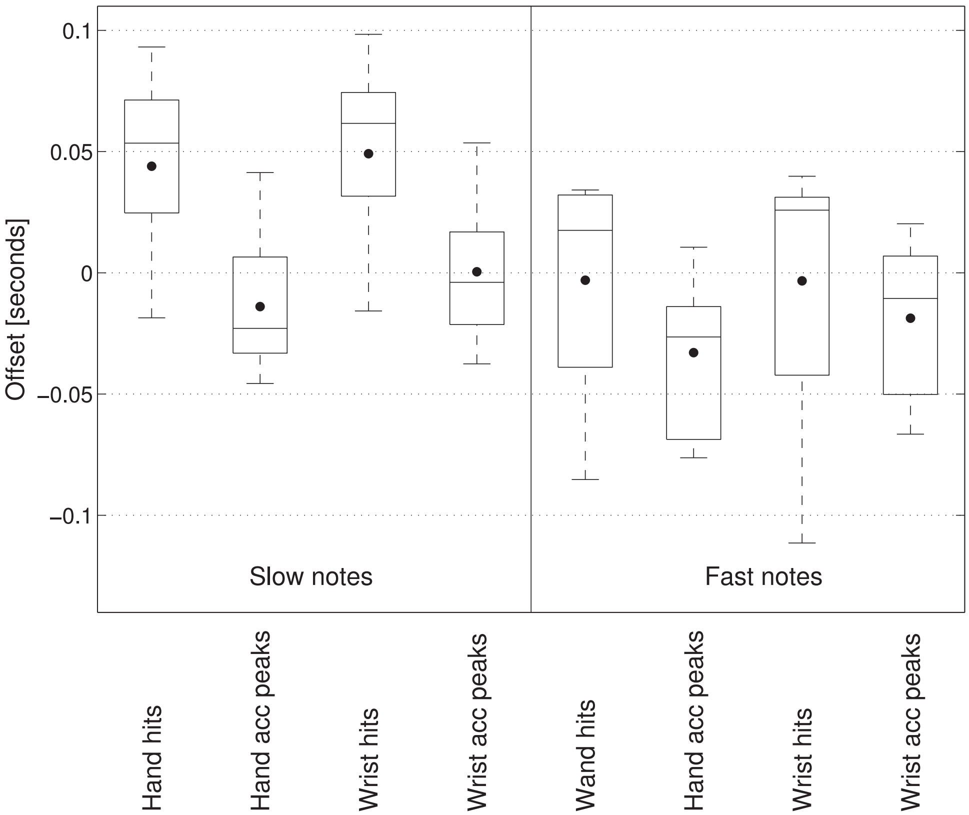
Effect of Note Speed

Earlier we found that for the hand, the timing of hits changes with note speed but the timing of acceleration peaks does not. From Figure 12 this also appears to be the case for the wrist. The wrist acceleration peak timing does not seem to change as much with note speed as does the wrist hit. These observations are confirmed by the following findings:

1. Wrist hits for fast notes occur earlier than those for slow notes ($t(9) = 5.2030$, $p = 5.6181 \times 10^{-4}$) by 29.6 to 75.2 msec (95% confidence interval).
2. For wrist acceleration peaks no significant difference is found in the timing of slow and fast notes.

These results are visualized in Figure 13, where we see that for hits there is a large and significant difference between the mean offset times for slow and fast notes for both the hand and the wrist. For acceleration peaks, there is also a difference between mean offset times for slow and fast notes, but it is much smaller and not significant. The implications of these results will be discussed later.

Figure 12. Box plot of mean offsets for all movement features (hits and acceleration peaks for hands and wrists). For each feature, the line is the median, the dot is the mean, the edges of the box are the 25th and 75th percentiles, and the whiskers extend to the most extreme data.



Analysis of Noise

For each participant, the mean and standard deviation of timing offsets were calculated for each movement feature. The earlier discussion focused on the mean offset, which tells us how much, on average, the movement features preceded or lagged behind the audio onset. I will now look at the standard deviations, which tell us how much variance there is in people's timing, or how difficult it is for people to be consistent in their timing.

One model from the sensorimotor synchronization literature suggests that variability in the timing of tapping is due to two sources, one from a clock mechanism in the central nervous system that indicates when to strike next, and the other from the motor system's ability to execute the intended gesture at a the desired time (Wing and Kristofferson 1973). There is another possible source of noise in my data, which is the amount of inaccuracy in my movement feature detection algorithms.

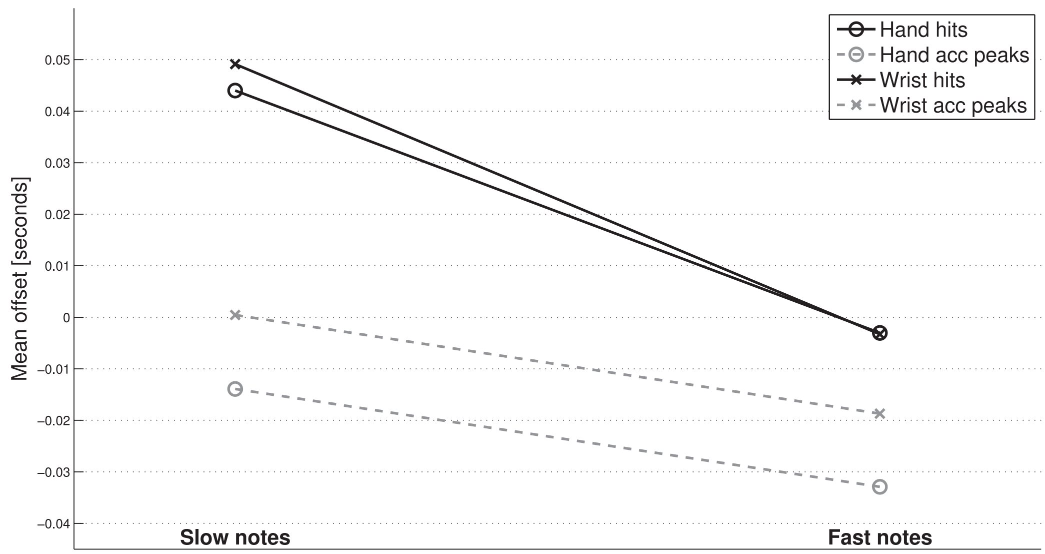
The question of interest here is, do some features have different amounts of noise? Features that can be executed and detected with less noise will be more successful at generating the sound at the time intended by the performer. Figure 14 shows box plots of the standard deviations for each feature for both slow and fast notes.

Noise in Acceleration Peaks and Hits

Upon inspecting Figure 14 we notice that for each body part, the mean standard deviation for the acceleration peak appears to be lower than the mean standard deviation for the hit.

A repeated-measures two-way ANOVA was conducted on the standard deviation for each participant. As with the calculations on the mean

Figure 13. Interactions between movement feature and note speed. The data points are the population mean for offset time means. Solid lines are significant differences, and dashed lines are not significant.



performed earlier, the two within-subjects factors were movement feature, with three degrees of freedom, and note speed, with one degree of freedom. A significant effect of movement feature was found ($F = 5.226$, $p = 0.0056$), but none was found for note speed.

Post hoc t-tests were conducted to compare standard deviations of hits and acceleration peaks, leading to the following four results:

For slow notes:

1. The noise for hand acceleration peaks is lower than the noise for hand hits ($t(9) = 4.5366$, $p = 0.0014$) by 1.4 to 7.8 msec (95% confidence interval).
2. The noise for wrist acceleration peaks is lower than the noise for wrist hits ($t(9) = 3.8552$, $p = 0.0039$) by 3.6 to 13.7 msec (95% confidence interval).

And for fast notes:

3. The noise for hand acceleration peaks is lower than the noise for hand hits ($t(9) = 2.4022$, $p = 0.0398$) by 0.5 to 16.1 msec (95% confidence interval).
4. The noise for wrist acceleration peaks is lower than the noise for wrist hits ($t(9) = 2.5487$, $p = 0.0313$) by 0.5 to 8.1 msec (95% confidence interval).

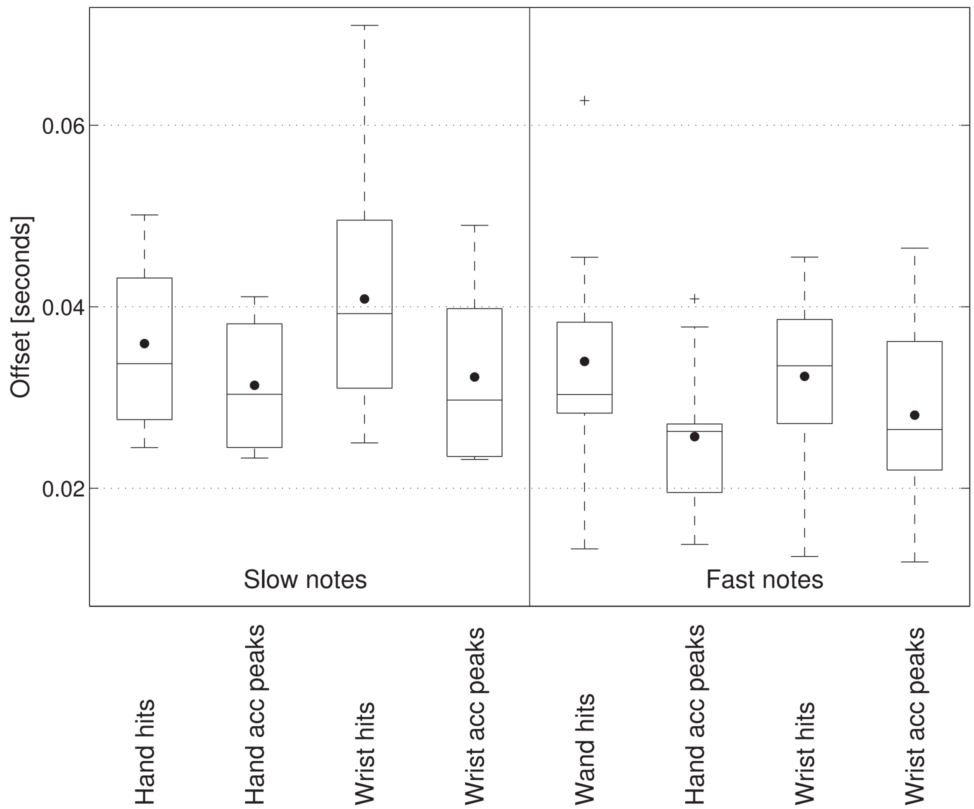
If we apply the conservative Bonferroni correction for four post hoc tests, requiring p values to be less than 0.0125, Results 3 and 4 would not be considered significant. Therefore, although it is not a very strong result, it does seem that the trend we observed in Figure 14 is confirmed: Movement features based on acceleration peaks have less noise than those based on detecting the sudden change of direction (i.e., hits), and the difference is more prominent for slow notes than for fast.

Discussion

How do these findings relate to other work on discrete musical air gestures? As described earlier, other real-time discrete air-gesture systems have used various movement features to trigger sounds, including thresholds in position (Mäki-Patola 2005; Collicutt, Casciato, and Wanderley 2009), velocity (Havel and Desainte-Catherine 2004; Dahl and Wang 2010), and acceleration (Kanke et al. 2012).

In my work I examined the moments where the hand and wrist change direction, which I call hits, and peaks in the magnitude acceleration of the hand

Figure 14. Box plot of offset standard deviations for all movement features. For each feature, the line is the median, the dot is the mean, and the edges of the box are the 25th and 75th percentiles, and the whiskers extend to the most extreme non-outliers. The crosses are outliers.



and wrist. I found, as did Collicutt, Casciato, and Wanderley (2009), that the moment of the hit does not necessarily correspond to a vertical minimum in the trajectory of the hand. This is why my hit detector is designed to find sudden changes of direction, irrespective of the orientation.

Mäki-Patola (2005) found that when striking a virtual drum surface in time to a metronome, the end of the strikes occurred after the metronome click. In my research I also find that hits occur after the note they are meant to correspond to. I believe this is because the hit is not the motion feature that performers are using to synchronize to the sound, and that acceleration peaks, which occur earlier, are more appropriate. This may help explain Mäki-Patola's finding.

Like Luck and Toiviainen (2006), Sarasúa and Guaus (2014) use maxima (i.e., peaks) in acceleration along the trajectory of the hand to detect the beats in conducting gestures. They find that these peaks correspond to vertical minima. For air-drumming gestures, however, I find that acceleration peaks and vertical minima do not necessarily coincide (Figure 9 shows an example). This difference may be what distinguishes conducting gestures from air-drumming gestures, and further research could prove illuminating.

Dahl (2004) used a threshold in jerk to detect drumsticks striking a real drum. These events (when the slope of acceleration exceeds a threshold) may occur very close in time to peaks in acceleration. The sudden acceleration of a drumstick however, is due, to the spring force of the drum head acting on the stick, whereas for air gestures, acceleration is the result of muscular torque exerted on the joints. The drummer may sense the rebound of the stick, but the air drummer must enact the muscular effort at the end of the strike.

Table 2. Offset Times: Means of Means and Standard Deviations

	Offset Mean		*Offset Standard Deviation*	
	Slow Notes	*Fast Notes*	*Slow Notes*	*Fast Notes*
Hand hits	43.00	−3.03	35.94	33.97
Hand acceleration peaks	−13.92	−32.90	31.34	25.69
Wrist hits	49.14	−3.29	40.86	32.33
Wrist acceleration peaks	0.45	−18.70	32.26	28.06

Grand means and mean standard deviations for offsets over all movement features. All times in milliseconds.

Which Feature Is Best?

If a digital instrument builder wants to design a system to trigger sounds with air-drumming gestures that has a timing that feels natural to the user, which movement feature should they use? There are a number of reasons why acceleration peaks are better.

As we saw here, acceleration peaks for the hand and wrist occur before the hit for the same body part. This makes acceleration peaks more useful in real-time systems, which need to predict when the audio event should occur early enough to take into account the various latencies in the system. Such latencies include the time it takes the motion-capture technology to deliver data to the detection algorithm, the time needed by the detection algorithm (e.g., to detect a peak in some value, some data that occur after the peak are needed), and the time it takes audio to pass through the sound hardware's buffering system.

We also saw that acceleration peaks had less variability in timing than their associated hits. Whether this variability is due to people's inability to produce gestures at the desired time or whether it is due to errors made by the movement feature detection algorithms, for a real-time air-gesture system it is better to use a feature with less noise.

Table 2 shows the mean, across all participants, of the offset mean and standard deviation for all four movement features. The hand acceleration peaks occur earliest and have the lowest standard deviation for both slow and fast notes, which suggests that acceleration peaks of the hand are a good feature to use for triggering real-time, discrete air gestures.

Effects of Note Repetition Speed

Acceleration peaks also perform better with respect to notes played at different tempi. The offset of hits was found to differ for slow and fast notes, whereas the offset for acceleration peaks did not change significantly with note speed.

When offset times for the hand are normalized for note speed, the differences between slow and fast notes are small (0.05 to 8 percent of a period) and barely significant. This suggests that the offset time for hits and acceleration peaks may scale with the speed of repetition. Gestures recorded at additional speeds are needed to investigate this hypothesis further, however.

Internal and External Perspectives on Movement

The hits of the hand and wrist locations are based on an external frame of reference. We track where these parts of the body are in space with respect to an origin and coordinate system that is external to the body. This makes them easy to observe with camera-based technologies such as marker-based motion capture.

I initially assumed that, when we make a discrete air gesture, we do something with our bodies to create an internal sense of a discrete moment in time. The forces we generate within our bodies, which we may experience as a sense of effort, and which result in accelerations on joints and body parts, are closer to our subjective experience of movement. This is why I find peaks in acceleration to be likely candidates for this internal sense of a moment in time. That acceleration peaks have less noise and don't vary as much with note repetition speed supports this view.

Furthermore, acceleration peaks tend to have a negative mean offset, like that found in sensorimotor synchronization experiments. When we tap we synchronize the tactile sensation of our finger hitting a surface to the sound, while when we make discrete air gestures, such as air drumming, it may be that we synchronize sudden muscular efforts to the sound.

Sonification of Movement Features

One way to experience these movement features is to listen to them. I created sonifications of the hits and acceleration peaks of the hand for part of the first trial of Participant 1, which can be heard in a video available on the *CMJ* Web site at www.mitpressjournals.org/doi/suppl/ (www.mitpressjournals.org/doi/suppl/10.1162 /COMJ_a_00298/AirDrum-HandFeatureSonification .mp4.

The video begins with an animation of air-drumming gestures in time to the audio stimulus, which is heard as a tom-tom sound. This is then accompanied by a higher-pitched tom-tom sound at the time of each detected hand hit. You can hear how the timing offset of hits varies during performance; that hits, in general, occur after the stimulus sound; and that they occur closer to the stimulus sound for fast notes. This is followed by a sonification of acceleration peak times as an even higher-pitched tom-tom sound. You may notice that, in general, acceleration peaks occur much more closely to the stimulus sounds than the hits did. These are sonifications of only a small segment of the data, but they do give a direct experience of the movement feature timing, and in a sense, hearing is believing.

Wrists and Movement Styles

Across the population of ten participants studied here, the timing of the hits was not significantly different between the hand and wrist, and the acceleration peaks for the wrist came after those of the hand. Because of real-time constraints, this suggests that if an air instrument can track only one point on the body, the hand is a better choice.

I showed how the difference in timing between movement features of the hand and wrist can be used to identify different styles of air-drumming gestures. This could be useful either to help calibrate a system to a user's preferred style, or to enable a system to detect different discrete air gestures from the same user.

Future Work

The hit and acceleration peak detection algorithms, as they are currently implemented, are not designed for real-time use. Both use thresholds that are calibrated to the range of the related variable over the length of a recorded trial, and the algorithm for choosing peaks relies on future values of acceleration. For real-time applications, these algorithms would need to be revised to work using only causal information.

Further analysis of these data may reveal other movement features, perhaps those based on joint angles or estimated muscular activation, that would more reliably indicate the correct time of the player's intended sounds. Other types of physiological data, such as electromyography, could also be studied.

One could also study nonstriking discrete air gestures, or gestures with other body parts. For example, bringing some part of one's body to a sudden halt is different from the drumming gestures studied here, which usually have a rebound.

Applications of Discrete Air-Gesture Research

The research described here, and future research into relationships between music and movement features, may have applications beyond triggering sounds from drum gestures. For example, holding a musical instrument in hand and gesturing in the air with it can be seen as a type of discrete air gesture. Thus controllers and "hyperinstruments" with motion sensors may be designed to more precisely trigger events or sudden changes in audio processing from air gestures made with the instrument. Similarly, systems that control musical processes from the movements of dancers may be made to have better timing with respect to the dancer's internally perceived sense of discrete movement events.

These results may also be helpful in designing nonmusical air-gesture-based interactions. The experience of playing gesture-controlled video games, as well as selecting items or pressing virtual buttons in generic gesture-based user interfaces, can all be improved by having timing that more closely aligns with the user's movement-generated sense of when the resulting event should occur.

Acknowledgments

This work was performed as part of my PhD research at the Center for Computer Research in Music and Acoustics (CCRMA) at Stanford University. The study was approved by the internal review board for human subjects research at Stanford University. Materials related to this work, including the videos mentioned herein, can be found online at purl.stanford.edu/kw169cw9976. Thanks to Chris Chafe, Takako Fujioka, and Jonathan Berger for their advice and support. Special thanks to Professor Fujioka for use of the motion capture system in the CCRMA NeuroMusic Lab. Thanks also to Jonathan Abel for suggesting the least-squares data-smoothing technique.

References

Clarke, E. F. 1999. "Rhythm and Timing in Music." In D. Deutsch, ed. *The Psychology of Music*. San Diego, California: Academic Press, pp. 473–500.

Collicutt, M., C. Casciato, and M. M. Wanderley. 2009. "From Real to Virtual: A Comparison of Input Devices for Percussion Tasks." In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pp. 4–6.

Coppens, F. 1985. "First Arrival Picking on Common-Offset Trace Collections for Automatic Estimation of Static Corrections." *Geophysical Prospecting* 33(8):1212–1231.

Dahl, L. 2014. "Triggering Sounds from Discrete Air Gestures: What Movement Feature Has the Best Timing?" In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pp. 201–206.

Dahl, L., and G. Wang. 2010. "SoundBounce: Physical Metaphors in Designing Mobile Music Performance." In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pp. 178–181.

Dahl, S. 2004. "Playing the Accent: Comparing Striking Velocity and Timing in an Ostinato Rhythm Performed by Four Drummers." *Acta Acustica united with Acustica* 90(4):762–776.

Glinsky, A. V. 1992. "The Theremin in the Emergence of Electronic Music." PhD thesis, New York University.

Havel, C., and M. Desainte-Catherine. 2004. "Modeling an Air Percussion for Composition and Performance." In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pp. 31–34.

Herrera, J., and H. S. Kim. 2014. "Ping-Pong: Using Smartphones to Measure Distances and Relative Positions." *Proceedings of Meetings on Acoustics* 20(1). Available online at dx.doi.org/10.1121/1.4860156. Accessed December 2014.

Kanke, H., et al. 2012. "Airstic Drum: A Drumstick for Integration of Real and Virtual Drums." In A. Nijholt, ed. *Advances in Computer Entertainment*. Berlin: Springer, pp. 57–69.

Luck, G., and P. Toiviainen. 2006. "Ensemble Musicians' Synchronization with Conductors' Gestures: An Automated Feature-Extraction Analysis." *Music Perception* 24(2):189–200.

Mäki-Patola, T. 2005. "User Interface Comparison for Virtual Drums." In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pp. 144–147.

Mathews, M. V. 1990. Three dimensional baton and gesture sensor. US Patent 4,980,519, filed 2 March 1990, and issued 25 December 1990.

Repp, B. H. 2005. "Sensorimotor Synchronization: A Review of the Tapping Literature." *Psychonomic Bulletin and Review* 12(6):969–992.

Rich, R. 1991. "Buchla Lightning MIDI Controller: A Powerful New Midi Controller Is Nothing to Shake a Stick at." *Electronic Musician* 7(10):102–108.

Sarasúa, Á., and E. Guaus. 2014. "Beat Tracking from Conducting Gestural Data: A Multi-Subject Study." In *Proceedings of the International Workshop on Movement and Computing*, pp. 118–123.

Wing, A. M., and A. B. Kristofferson. 1973. "Response Delays and the Timing of Discrete Motor Responses." *Perception and Psychophysics* 14(1):5–12.