THE SHAPE OF AN INSTANT: MEASURING AND MODELING PERCEPTUAL ATTACK TIME WITH PROBABILITY DENSITY FUNCTIONS

(IF A TREE FALLS IN THE FOREST, WHEN DID 57 PEOPLE HEAR IT MAKE A SOUND?)

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF MUSIC
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Matthew James Wright

March 2008

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Chris Chafe) Principal Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Jonathan Berger)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Julius Smith)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(David Wessel)

Approved for the Stanford University Committee on Graduate Studies.

Abstract

A musical event's *Perceptual Attack Time* ("PAT") is its perceived moment of rhythmic placement; in general it is after physical or perceptual onset. If two or more events sound like they occur rhythmically together it is because their PATs occur at the same time, and the perceived rhythm of a sequence of events is the timing pattern of the PATs of those events. A quantitative model of PAT is useful for the synthesis of rhythmic sequences with a desired perceived timing as well as for computer-assisted rhythmic analysis of recorded music. Musicians do not learn to make their notes' *physical onsets* have a certain rhythm; rather, they learn to make their notes' *perceptual attack times* have a certain rhythm.

PAT is notoriously difficult to measure, because all known methods can measure a test sound's PAT only in relationship to a physical action or to a second sound, both of which add their own uncertainty to the measurements. A novel aspect of this work is the use of the ideal impulse (the shortest possible digital audio signal) as a reference sound. Although the ideal impulse is the best possible reference in the sense of being perfectly isolated in time and having a very clear and percussive attack, it is quite difficult to use as a reference for most sounds because it has a perfectly broad frequency spectrum, and it is more difficult to perceive the relative timing of sounds when their spectra differ greatly. This motivates another novel contribution of this work, *Spectrally Matched Click Synthesis*, the creation of arbitrarily short duration clicks whose magnitude frequency spectra approximate those of arbitrary input sounds.

All existing models represent the PAT of each event as a single instant. However, there is often a range of values that sound equally correct when aligning sounds rhythmically, and this range depends on perceptual characteristics of the specific sounds such as the sharpness of their attacks. Therefore this work represents each event's PAT as a continuous probability density function indicating how likely a typical listener would be to hear the sound's PAT at each possible time. The methodological problem of deriving each sound's own PAT from measurements comparing pairs of sounds therefore becomes the problem of estimating the distributions of the random variables for each sound's intrinsic PAT given only observations of a random variable corresponding to difference between the intrinsic PAT distributions for the two sounds plus noise. Methods presented to address this draw from maximum likelihood estimation and the graph-theoretical shortest path problem.

This work describes an online listening test, in which subjects download software that presents a series of PAT measurement trials and allows them to adjust their relative timing until they sound synchronous. This establishes perceptual “ground truth” for the PAT of a collection of 20 sounds compared against each other in various combinations. As hoped, subjects were indeed able to align a sound more reliably to one of that sound’s spectrally matched clicks than to other sounds of the same duration.

The representation of PAT with probability density functions provides a new perspective on the long-standing problem of predicting PAT directly from acoustical signals. Rather than choosing a single moment for PAT given a segment of sound known a priori to contain a single musical event, these regression methods estimate continuous shapes of PAT distributions from continuous (not necessarily presegmented) audio signals, formulated as a supervised machine learning regression problem whose inputs are DSP functions computed from the sound, the detection functions used in the automatic onset detection literature. This work concludes with some preliminary musical applications of the resulting models.

Acknowledgements

First I'd like to thank all of the music teachers and friends who have taught me anything about rhythm over the years, including Tameem Afzali, Mark Applebaum, Kim Atkinson, Dan Auvil, John Baily, Souren Baronian, Jorge "Alabê" Bezerra, Boca Rum Bezerra, David Brown, Edmund Campion, Mitch Chakour, George Chittenden, Faith Conant, Sinan Ali Erdemsel, Suleyman Feldthouse, Polly Tapia Ferber, Brian Ferneyhough, Ernie Fischbach, Evan Gavriel Fiske, Tim Fuson, Richard Harrington, Bizmillah Iqbal, Scott Kettner, Aashish Khan, Swara Samrat Ustad Ali Akbar Khansahib, Zia Khawaja, Mohammad Rahim Khushnawaz, Ben Kunin, Mark Lamson, Juliet Lee, John Maggs, Haig Manoukian, Jorge Martins, Karim Nagi Mohammed, Naser Musa, Nininho (Maracatu Badia), Chris Overseas, Marcio Peeters, Curtis Pierre, Justino Rogers, Homayoun Sakhi, W. Andrew Schloss, Robert Schoville, Selim Sesler, Bruce Silverman, Michael Spiro, Klaus Urban, and Derek Wright.

I also acknowledge indispensable technical assistance large and small from kindly researchers including Jonathan Abel, Ed Berdahl, Nick Collins, Roger Dannenberg, Kelly Fitz, Henkjan Honing, Patty Huang, Lauri Kanerva, Edward Large, Sasha Leitman, Fernando Lopez-Lezcano, Michelle Logan, Andrew Schmeder, Malcolm Slaney, Julius O. Smith III, Edward Tufte, George Tzanetakis, Carr Wilkerson, and David Zicarelli.

I would particularly like to thank Prof. Julius O. Smith III for his online books on musical signal processing (Smith 2007a, 2007b, 2007c, 2007d), which have been invaluable references for me and many other people. I have made an effort to link into this rich web of content in many footnotes not because I think he personally made up everything he taught me, but because each URL is an excellent starting point for learning more about any given topic.

Thanks also to the 16 students of Professor Jonathan Berger's Music 151 course during Spring 2007 who served as subjects for my pilot study.

Finally, I acknowledge the various forms of support I received in the course of this dissertation from Juliet Lee, Michelle Logan, Sonia Rosenbaum, W. Andrew Schloss, George Tzanetakis, Kris Willits, Derek Wright, James Wright, U.C. Berkeley's Center for New Music and Audio Technology (CNMAT), and Stanford's Center for Creative Research in Music and Acoustics (CCRMA).

In spite of all this help I have made many errors, and take full credit for any that remain.

Table of Contents

List of Tables

List of Illustrations

Chapter 1 Background and Motivation

Rhythm in some form underlies almost all human activity, and it is certainly one of the most important elements of music worldwide. A comprehensive theory of rhythm should include a generalized view of sound existing on a continuum of time and explain how listeners make sense of sound by noticing various forms of quasi-repetition and by inferring discrete events at specific attack times. It must also explain how the mind actively groups these musical events temporally into both phrase structures and metric structures, that is, sets of quasi-even pulses organized hierarchically with additive and divisive operations, operating within certain temporal limits of human perception and movement.

Finally, a theory of rhythm should address questions of metric phase relations in general (and especially the notion of "downbeat" and its epistemology), syncopation, microtiming, ornamentation, rhythmic relationships in polyphonic, polyrhythmic, polymetric, and polytempic contexts, the close relationship between perception and production (embodiment), prediction, and the role of learning and acculturation (for both performers and listeners) in all of the above.

1.1 Some Questions

The following questions inspired the studies into musical rhythm that eventually became the present dissertation.

- *Rhythm*. What is rhythm? How do the human auditory and cognitive systems group and make sense of the timing of what we hear? What is special about the perception of time in music versus the perception of time in general? Why and how do certain sequences of sound generate expectation of the timing of future sound? Why and how are human listeners able to "find the beat," for example, tap their feet to the pulse of music? Why, how, and when do we perceive discrete "notes" and other sonic events amidst the smooth continuum of time?

- *Meter*. Why do almost all cultures have some form of metric structuring that places musical events at discrete time points within an essentially repeating cycle of pulses, accents, and durations? Why and how do some specific sequences of onset times and durations give rise to a sensation of meter while others do not? Music can easily be notated in the "wrong" meter; what makes one plausible metric interpretation of a given figure "correct" and others "incorrect"? Given a repeating pattern, on what basis does one choose the metric beginning (downbeat) of the pattern?

- *Syncopation*. What is syncopation? Why and how do certain syncopated rhythms retain a clear sense of pulse and meter, even if the beats themselves are not articulated? At what point does syncopation break the senses of pulse and/or meter? How can one syncopate against an irregular meter?
- *Influence of natural rhthyms*. What is the connection between human biological rhythms (such as heartbeat, respiration, gait, sexual intercourse) and musical rhythm? What is the interaction between evolutionary universals of human rhythmic experience and culturally specific musical rhythm?
- *Microtiming*. Computer-generated performances with exact, strictly metronomic timing have an unmistakably inhuman sound, even if all other nuances of expression (loudness, timbre) are taken from an expressive human performance. Adding randomness to the timing to "synthesize" human expression generally sounds both inhuman and imprecise. Therefore there is structure to microtiming[1]. What is that structure in general and in specific musical styles?

Why ask these questions? I believe (and many agree) that rhythm is an essential part of what it means to be human, and that pursuing these questions will teach us more about what it means to be human. This might be labeled my "humanistic" motivation. Another reason is that concrete, quantitative answers or partial answers to some of these questions could take the form of models that can be implemented by computer. Working software models of rhythm will enable exploratory compositional play. This might be labeled the "creative" or "expressive" motivation. Finally, a large part of my goal in pursuing these questions could be labeled my "aesthetic" motivation: music produced by computers falls overwhelmingly into either "too much rhythmic regularity": strictly metronomic, highly repetitive electronic dance music; or "not enough rhythmic regularity": essentially all "Computer Music" coming out of the academic Western art music tradition, where either large amorphous masses of sound lack all sense of articulation, or event onsets are so irregular as to eschew any sense of pulse. I believe that both extremes give up one of the important avenues of musical expression, hence my desire to develop better computer tools.

[1] "Microtiming" is jargon referring to the small time differences between a strictly metronomic performance and the times that notes are actually performed.

1.2 Structure of This Dissertation

The research presented here focused on just one part of one of the above questions: “when do we perceive discrete sonic events amidst the smooth continuum of time?” This is the issue known as *Perceptual Attack Time* (“PAT”) or, in the speech community, *Perceptual Center* (“P-Center”). My novel theoretical contribution is to treat each sound’s PAT not as a single instant but instead as a continuous probability density function.

Chapter 2, “Towards a Comprehensive and Culturally Universal Theory of Musical Rhythm,” considers the big picture, attempting to address some of the above questions with a literature review as well as introducing some important terminology. It also demonstrates the centrality of the PAT question to musical rhythm in general so as to motivate subsequent material.

Chapter 3, “On Perceptual Attack Time,” takes up the question of PAT in detail, first defining the term and discussing the various methods for measuring PAT and their specific difficulties. The observation that there can be a range of perceptually “correct” judgments of a sound’s PAT motivates the treatment of PAT with probability density functions. The theoretical heart of this work suggests a probabilistic interpretation of results from PAT measurements and offers a variety of methods for estimating each individual sound’s PAT (as a probability density function) given the measurable data of the relative difference in PAT between pairs of sounds. Finally it takes up the question of what reference sounds should be most effective for measuring a given sound’s PAT, which motivates *Spectrally Matched Click Synthesis*, the creation of arbitrarily short-duration clicks with spectra matching that of any given input sound; I show that this problem is reducible to the well-studied problem of finite impulse response (FIR) filter design.

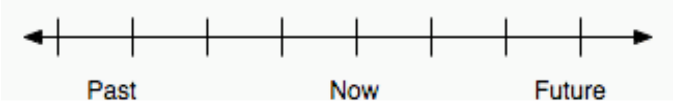
Chapter 4, “Listening Experiment” describes an experiment I performed to test PAT. Volunteer subjects downloaded custom software (described in detail in Appendix A “Software for Administering the Listening Experiment”) that presented them with a series of trials aligning the PAT of pairs of sounds. The chapter describes the experiment in detail, presents the results, and interprets them according to the theory developed in Chapter 3.

Chapter 5, “Motivations, Implications, and Future Work ,” takes up the question of estimating any sound’s PAT directly from acoustic properties of the signal, again from a perspective representing PAT as probability density functions rather than discrete instants. I present two approaches, both framed as regression problems in the context of supervised machine learning, with the training data coming from the results of Chapter 4. The first approach estimates PAT for each individual sound, as do all existing predictive models of PAT and P-Center. The second approach bypasses Chapter 3’s thorny theoretical issue of recovering each sound’s own PAT from measurements of relative PAT for pairs of sounds, and instead directly estimates the relative PAT for two sounds. The chapter concludes by suggesting some possible uses for these predictive models.

Chapter 2 Towards a Comprehensive and Culturally Universal Theory of Musical Rhythm

The goals of this chapter are to introduce the terminology and concepts used in the rest of the dissertation and to build towards a theory of musical rhythm based on a solid scientific epistemology while attempting to keep all of the world's musical cultures in mind.

2.1 Scientific View of The Time Axis



Musical rhythm almost[2] always manifests as sound, which consists of patterns of variation of pressure as a continuous[3] function of time. Time is one of the few fundamental quantities of physics and other sciences. To the extent than an audio recording captures a musical performance, that music is nothing more than a function of time.[4] Indeed the time axis is fundamental to most of science and engineering as well as to performing arts such as dance, theater, and film. Thus, the time axis is the foundation of this theory of rhythm.[5] In this view, time has an objective reality independent of any observer, and clocks can measure and divide time accurately.

The Einstein-Minkowski model of space-time tells us that time itself actually passes differently at every location according to relative velocity, gravity, etc.[6]; luckily these effects are negligible for all human music making.[7] An effect that we must take into account, however, is the slow speed of sound (approximately one foot per millisecond), which guarantees that if there are two or more

[2] For a counterexample, consider Mark Applebaum's 1995 composition *Tlön* for three conductors and no players (markapplebaum.com/tlon01.html). Also, musical rhythm can occur entirely in the imagination, as when a musician silently rehearses by mentally running through a piece. Certainly any non-sounding example of musical rhythm has meaning only in reference to the normal case of music that is heard; furthermore even these silent examples take place in time more or less in the same way as sounded music.

[3] For practical purposes we approximate continuous time with the successive discrete samples of digital audio.

[4] A stereo recording is *two* functions of time: left and right.

[5] Most theories of musical rhythm instead begin with Western musical notation and therefore wind up ignoring many important issues. My goal is to base this theory on a more general epistemology (Cook 2002) based on physics, results from psychology, and my personal understanding of musics of many cultures.

[6] (Hawking 1988) is a good reference on these ideas for the lay reader.

[7] (Shlain 1991) argues that great visual art depicts nonintuitive aspects of the physical world such as the relativity of time.

sound sources (be they loudspeakers, musical instruments, human voices, etc.), then the relative timing of each source will be different at every listening location.[8]

2.2 Repetition / Periodicity

The next idea is a function of time that repeats exactly at regular time intervals. We call this *repetition* when we perceive that something discrete keeps happening, for example, a dripping faucet, whereas we call it *periodicity* when we perceive that the overall shape keeps repeating without there necessarily being a clear beginning and ending point, like the up and down pitch of a siren. The only difference is in how we think of the unit that repeats.

Much of the theory of mathematics and signal processing is based on the ideal of exact repetition. For example, the sine function has the property that $sin(x)=sin(x+n2\pi)$ for any x and any integer n, so by modeling any function of time as a sine wave, we are assuming exact repetition. Fourier's technique of decomposing any signal into a sum of sine waves applies only to the case of infinitely long signals that repeat exactly.[9] Taking the Discrete Fourier Transform (DFT)[10] to find the spectrum of a finite-length signal s implies the concept of *periodic extension*[11], meaning that what we are really finding is the spectrum of an infinitely long, exactly repeating signal of which s is a single period (Smith 2007b).

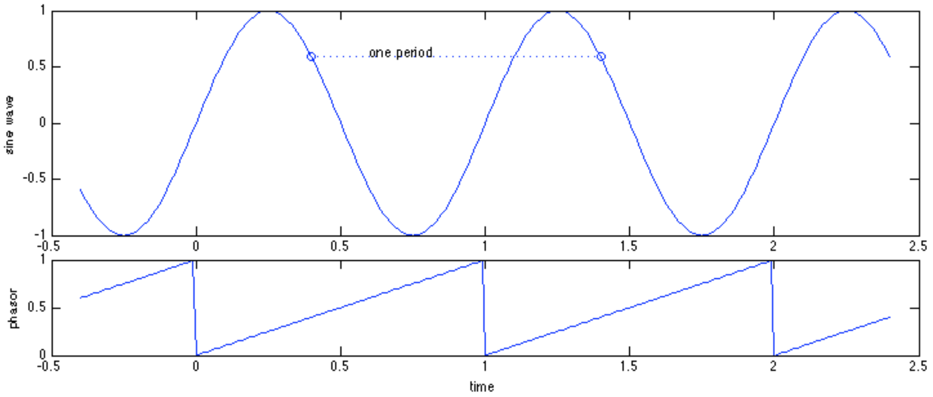


Figure 1: Example of an ideal sine wave oscillator

[8] Except for unlikely special cases such as two sound sources and two listeners all positioned exactly along a single straight line.

[9] More formally, given any finite-duration input signal, we can perform Fourier analysis to derive a Fourier series; the expansion of that Fourier series will be a "periodic function," meaning an infinitely long signal consisting of infinitely many exact repetitions of the original input signal. This is periodic extension in the continuous time case.

[10] In practice people use the more efficient FFT (Fast Fourier Transform) to compute the DFT.

[11] http://ccrma.stanford.edu/~jos/mdft/Modulo_Indexing_Periodic_Extension.html

Because this theoretical ideal is so useful and widespread, it is worth introducing some of the terminology. Figure 1 shows almost three cycles of a sine wave and its phase[12] as functions of time. An *oscillator* is an abstract mathematical tool that cycles through a repeating pattern; in this case the graph can be interpreted as the output of a sine wave oscillator. A *period* is the amount of time that an oscillator takes to go through its repeating pattern one time, in this case, one second. *Frequency* is the rate of repetition, that is, the reciprocal of period, in this case, one Hertz. The *phase* of a signal at any given time is its proportion of the way through the period at that time; the bottom graph in Figure 1 shows the phase as a function of time. Frequency is the rate of phase change.[13] A *phasor* is an oscillator whose output is just its current phase (so it makes a "sawtooth" shape). The choice of which particular point in the cycle to label as "phase zero" is a matter of convention; the important thing about phase is that it increases steadily from zero to one, then "wraps around" back to zero and starts over. In this example, phase zero is the point where the oscillator's output is zero and heading from positive to negative.[14]

We can bring together musical and mathematical terminology by using an oscillator as a model of an ideal metronome. It "ticks" when the phase is zero, and the frequency is proportional to the tempo.[15] The period is the duration of one beat. The current phase is the proportion of time that has elapsed between the previous tick and the next tick.

As an example of *phase relationships*, imagine two ideal metronomes set to the exact same tempo. If they start at exactly the same time, they will always tick at exactly the same time infinitely into the future. In this case we say the oscillators are *in phase* with each other; that is, their phase difference is zero. If the metronome on the left starts first, and then the metronome on the right starts exactly halfway between the ticks of the metronome on the left so that they alternate evenly "left, right, left, right…," then we say they are exactly *out of phase* with each other; that is, their phrase difference is 0.5.[16] A phase difference of 0.1 would mean that the second metronome always ticks 10% of the way between the first metronome's ticks, etc.

[12] Here I'm using the convention that phase goes between 0 and 1, as is typically seen when applying these concepts to musical rhythm; in most other situations phase is defined between 0 and 2π, or between –π and π.

[13] Frequency in Hertz is the rate of phase change when phase is defined to go from 0 to 1. If phase goes from 0 to 2π than radian frequency (which is 2π times frequency in Hertz) is the rate of phase change.

[14] In other words, this is the *sine* function. For a *cosine*, phase zero is the point at which output is the highest.

[15] Musical tempo is usually measured in beats per minute ("BPM"); divide this by 60 to get beats per second, which is the frequency of the metronome in Hertz.

[16] Again, I'm using the convention of phase from 0 to 1. "Exactly out of phase" would correspond to a phase difference of "180 degrees" or "pi radians."

2.2.1 Quasi-Repetition

There is no exact repetition in the real world, only quasi-repetition.[17] Sound never repeats exactly, because there is always a noise floor[18] in any acoustic environment. Figure 2 illustrates the same sine wave as Figure 1, but with noise added.[19]

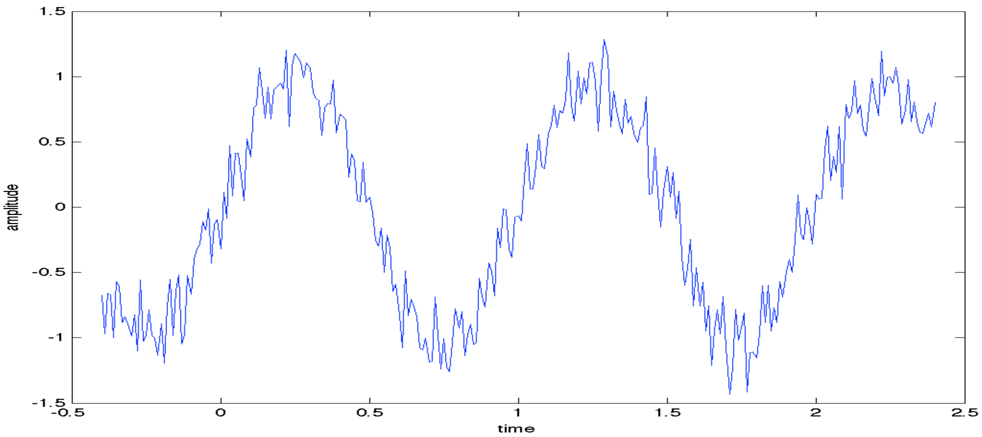


Figure 2: Example of an ideal sine wave oscillator with added noise.

Any quasi-repeating sound produced by human activity will have small "errors" in timing caused by what is known as "motor noise"; these are never much less than about 1 millisecond and often much more.[20] Only mechanical means such as tape loops and their digital equivalents can produce repetition that is exact to within the limits of human perception, and indeed a lot of popular music and the vast majority of electronic dance music is constructed digitally with exact repetition of rhythmic sequences or of a segment of prerecorded audio. Exact repetition in

[17] For example, the astronomical cycles that most affect life on earth, namely the day (rotation of the earth), the lunar month (phase of the moon), and the year (orbit of the earth around the sun), are essentially the same from cycle to cycle, but slightly different as the length of the day changes throughout the year, as the earth very gradually slows down in its orbit around the sun, etc.

[18] Even in an ideal recording studio with perfect sound isolation from the outside world (e.g., in outer space), the Brownian motion of the heat of the vibrating medium adds a nondeterministic and therefore non-repeating component to any sound. Also all microphones, mixers, recording devices, loudspeakers, etc., have a level of noise that they add to any signal.

[19] In this example the noise is Gaussian with mean zero and standard deviation 0.2.

[20] This motor noise turns out to be difficult to measure. One method is to ask subjects to tap at a variety of steady frequencies and measure the variance, then decompose this variance into a "central clock variance," i.e., the mind's inability to maintain a perfectly steady tempo, plus the motor noise as an added source of variance (Wing and Kristofferson 1973a, 1973b). With these methods "typically the motor variance is in the range of about 25 ms-squared (i.e., standard deviation of about 5 ms) and changes little with tempo" (Bruno Repp, personal communication, January 31, 2008). Rubine and McAvinney put this figure around 1.5 to 4ms (Rubine and McAvinney 1990), while Desain and Honing say 10-100 ms (Desain, Honing, and Rijk 1989), both citing (Vorberg and Hambuch 1978). See also (Lago and Kon 2004). Another method is to ask a performer to play a piece "the same way" two or more times and then measure the correlation in note durations among the results (Repp 1995). This approach also suffers from the methodological problem of separating the mind's variance from the body's.

rhythm is analogous to sustaining the tonic chord in harmony: it is the most basic arrangement, and in its pure form, artless and boring.

2.2.2 Perceiving Quasi-Repetition at Different Time Scales

Perception of quasi-repetition is qualitatively different depending on the period of repetition, as shown in Table 1 and Figure 3. One very important length of time for human perception is the *perceptual present*, a short-term auditory memory storing approximately the last 2-6 seconds of sound input (Clarke 1999). Another important length of time is around 50ms, the boundary between our perception of rhythm and pitch. Although many have noted the mathematical equivalence of periodicity in these two time scales, for example, (Scheirer 1997; Stockhausen 1957), we perceive them quite differently.

Period	***Regime of perception***
More than a human lifetime	Repetition cannot be perceived.
More than a few minutes	Repetition can be noticed by comparing new input to long-term memory.
More than about 2 seconds	Repetition can be noticed by comparing new input to short-term memory.
About 100 ms to 2 seconds	Repetition takes place within the "psychological present" and is perceived more or less automatically as a rhythm or meter.
About 50-100 ms	Grey area between pitch and rhythm: generally sensory roughness.
0.05 to 50 ms (i.e., 20-20000 Hertz)	Pitch
Less than 0.05 ms	Repetition cannot be perceived.

Table 1: Regimes of perception of quasi-repetition

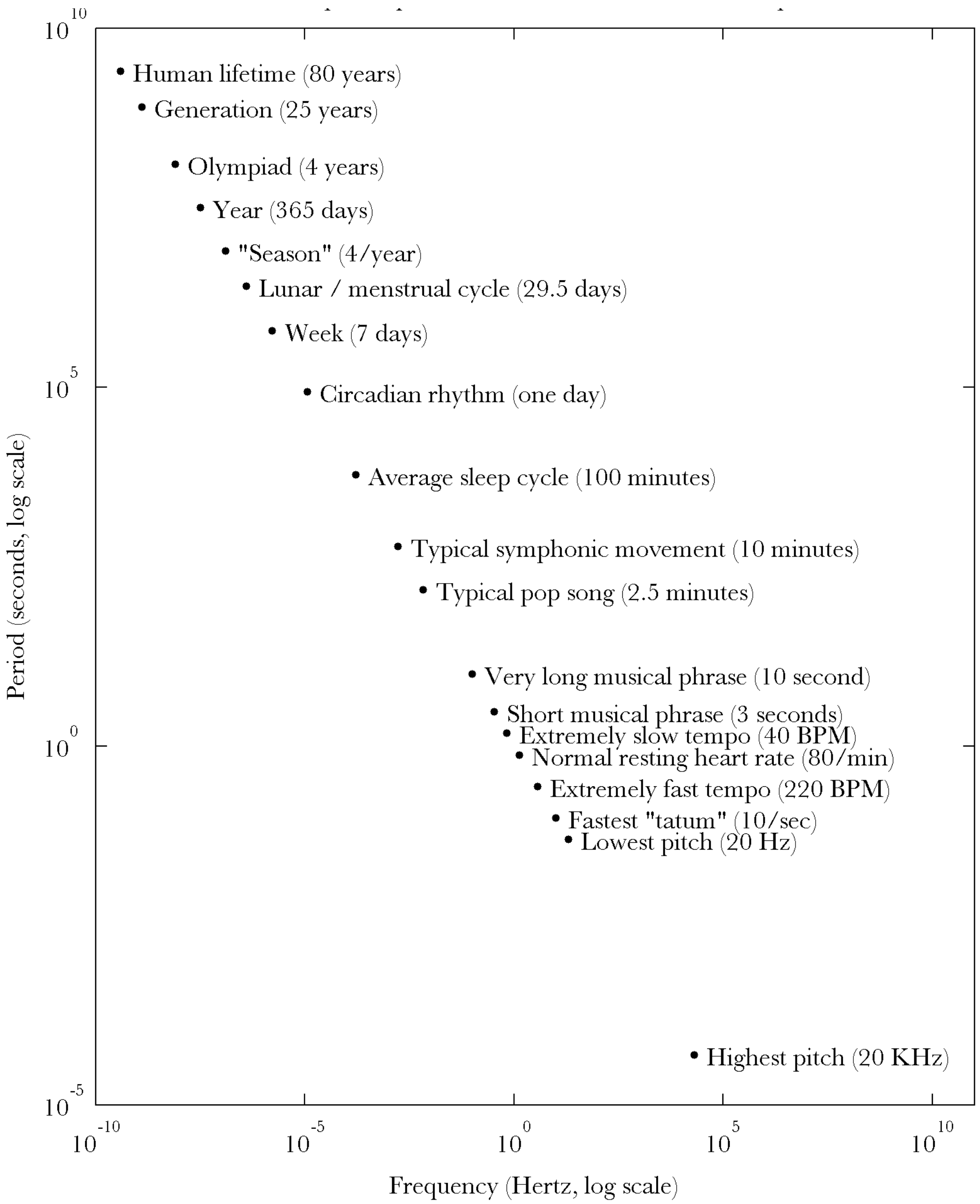


Figure 3: Some example frequencies and time-scales of human experience, charted simultaneously on log-frequency and log-period scales.

All points lie along a straight line because frequency and period are reciprocals. Many of these numbers are obviously example values drawn from a range of possibilities. These are the frequencies and time-scales that we perceive as quasi-repeating cycles. We can perceive even shorter durations in the inter-ear timing differences that give us cues about the direction of a sound source, but we cannot perceive cyclic repetition at these fast rates. On the other extreme, we can perceive, for example, that we are living in the 21st century, but we cannot perceive the centuries themselves as repeating cycles. This figure is inspired by the brilliant figure "The Time Domain" from (Roads 2001, 5).

2.2.3 Signal Processing Techniques for Detecting Quasi-Repetition

In signal processing, the term *periodicity estimation* is used to mean automatic techniques for detecting quasi-repetition. Two are well known in the computer music literature, originally applied in the regime of pitch perception to do fundamental frequency estimation and also applied more recently for analysis of rhythmic and metric structures: autocorrelation and harmonic spectral product.

2.2.3.1 Autocorrelation

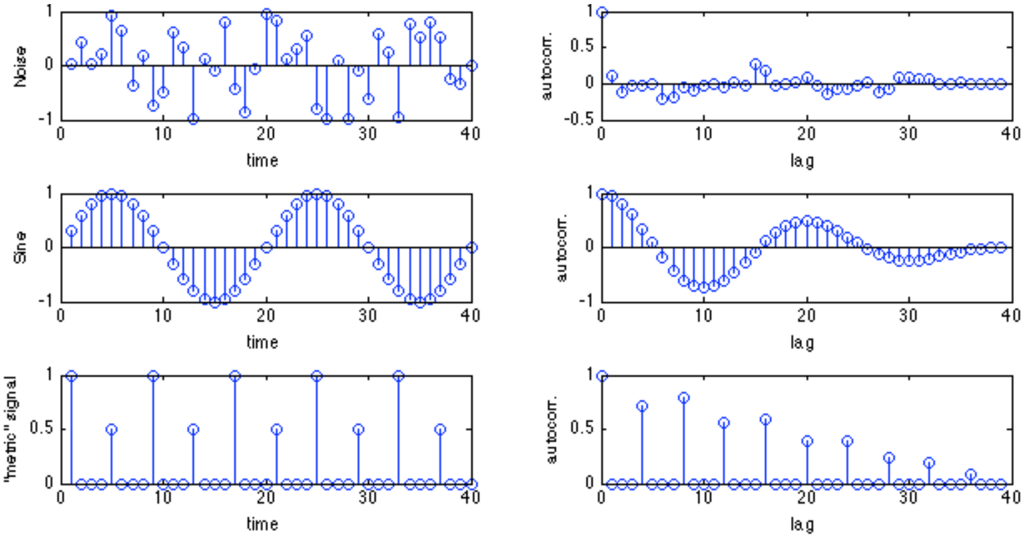


Figure 4: Autocorrelation of three short signals

Autocorrelation of three short signals: noise (top), sine wave (middle), and a primitive idealized "metric" signal (bottom)

Autocorrelation finds the correlation[21] of a signal against different versions of itself time-shifted by various amounts. Each time-shift amount is called a *lag time*. The output of an autocorrelation is the correlation amount[22] as a function of lag time. The maximum value will always be at a lag of zero, since a signal is always perfectly correlated with an exact copy of itself.[23] Other peaks in the

[21] In this sense *correlation* between two signals x and y simply means the scaled sum of the pointwise product: $k\Sigma x_i y_i$, what a statistician would call the "sample cross correlation"; see http://ccrma.stanford.edu/~jos/mdft/Cross_Correlation.html

[22] I've scaled the values so that the maximum correlation, i.e., the correlation between two copies of the same signal, is one (by using the 'coeff' argument to Matlab's *xcorr* function). A correlation of zero means that the two signals have nothing in common.

[23] For some signals the (unbiased) correlation amount at other lags might be equal to that at lag zero. For example, a completely constant signal will have equal correlation amounts at all lag times. A perfectly repeating signal's autocorrelation at a lag equal to the period will be the same as that at lag zero. An *unbiased* correlation

autocorrelation indicate lag times at which the signal is relatively highly correlated with itself; these can be interpreted as periods at which the signal quasi-repeats. In other words, autocorrelation is based on the idea that a quasi-periodic signal will resemble itself in the time domain when time-shifted by a duration (nearly) equal to the period.

Figure 4 shows three short signals alongside their autocorrelation functions. For the noise signal, other than the peak at lag zero, there does not seem to be any structure to the autocorrelation. The sine signal changes gradually, so the autocorrelation also changes gradually; this helps explain why autocorrelation is good at finding not-quite-exact repetition when the signal is somewhat smooth. The third example is a “metric” signal of a loud impulse alternating with a quieter impulse, separated by three units of silence. In this case, the lag of 4 yields a peak, since it makes the impulses all line up with other, while any lag amount that is not a multiple of 4 yields a zero because it makes the impulses line up with silences. The lag of 8 has an even higher correlation than the lag of 4, because in addition to making all the impulses line up with each other it also makes the loud impulses line up with each other.

Many researchers have used autocorrelation for rhythmic analysis: (Alonso, David, and Richard 2004; Brossier 2006, 105-110; Brown 1993; Davies and Plumbley 2004; Davies and Plumbley. 2005; Frieler 2004; Paulus and Klapuri 2002; Peeters 2005; Scheirer 1997; Toiviainen and Eerola 2005; Tzanetakis, Essl, and Cook 2001).

A recent development in the use of autocorrelation is the *autocorrelation phase matrix* (Eck 2007; Eck and Casagrande 2005), which outputs a two-dimensional (2D) matrix showing the correlation amount as a function of both lag time and phase. The distribution of autocorrelation energy in this space can reveal rhythmic structure even in cases where the autocorrelation alone provides no insight.

Related to autocorrelation is a method based on *comb filtering* (Scheirer 1998), in which an input signal passes through a collection of recirculating feedback delay lines. For example, the output of the one-second delay line is equal to the input plus a quieter version of the input from exactly one second ago, plus an even quieter version of the input from exactly two seconds ago, etc. So if the input contains periodicity at or near the one Hertz frequency, the amount of energy in the one-second delay line will tend to increase.

(http://ccrma.stanford.edu/~jos/mdft/Unbiased_Cross_Correlation.html) corrects for the fact that higher lag times correspond to shorter durations of the overlap between the original and time-shifted versions of the signal; the graphs in Figure 2 show regular biased autocorrelation.

2.2.3.2 (Harmonic) Spectral Product

The magnitude spectrum of a quasi-repeating signal should have a peak corresponding to the frequency of repetition. The *harmonic spectral product* method (sometimes called just "spectral product") is based on the assumption that the spectrum of a quasi-repeating signal will also have relatively strong peaks at frequencies corresponding to the first few harmonics of the frequency of repetition. This method works by first finding the magnitude spectrum (for example, with an FFT), then successively compressing that spectrum by factors of 2, 3, etc., up to M, then multiplying together all M spectra.[24]

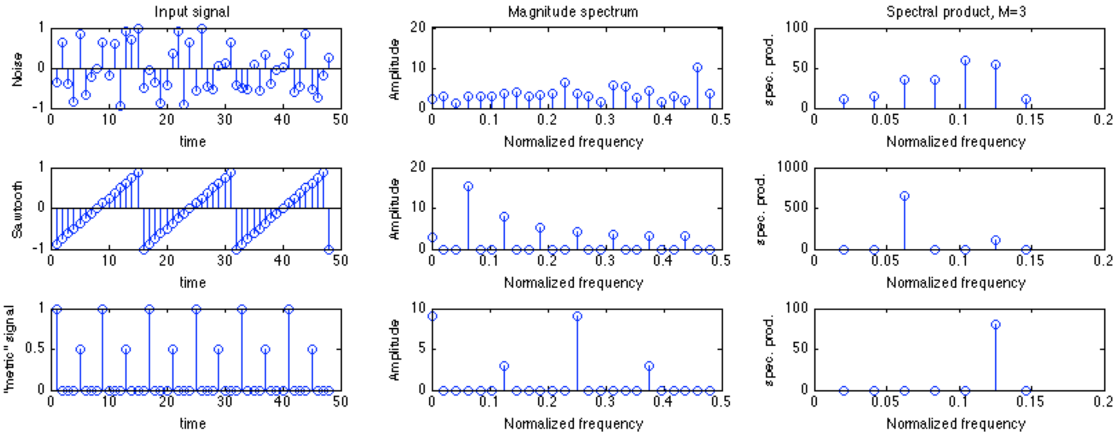


Figure 5: Harmonic spectral product of three short signals: noise (top), sawtooth wave (middle), and a primitive idealized "metric" signal (bottom)

Figure 5 shows three short signals, their magnitude spectra, and their harmonic spectral products with M=3. For the noise signal the magnitude spectrum is basically flat and any structure to the spectral product is random.[25] The sawtooth wave has a harmonic spectrum exactly like what this method expects to see, and indeed the spectral product has a huge peak at the sawtooth's fundamental frequency. The "metric" signal is perfectly periodic with a harmonic spectrum and so again the spectral product technique easily finds the fundamental frequency.

[24] Alonso (Alonso, David, and Richard 2004) writes this formula for spectral product, where f is normalized frequency and $P(e^{j2\pi f})$ is one bin of the FFT of the input signal:

$$S(e^{j2\pi f}) = \prod_{m=1}^{M} |P(e^{j2\pi mf})| \quad \text{for } f < \frac{1}{2M}$$

[25] The apparent structure in this example is due to the short duration (only 48 samples) of noise. As the number of noise samples increases the magnitude spectrum and therefore spectral product become flat.

Alonso has used harmonic spectral product to detect periodicity in the domain of rhythm (Alonso, David, and Richard 2004). The use of this technique to detect pitch has a rich history going back at least to 1969 (Noll 1969).

2.3 Discrete Events and Discrete Instants

The idea that sound consists of independent, distinct *events* (for example, musical *notes*) is both useful and dangerous. It is useful because many sounds are indeed produced by distinct, all-or-nothing physical actions, for example, striking, plucking, dropping, plosive consonants, etc. Many other sounds are produced by continuous physical actions, for example, singing or speaking vowels or voiced consonants, wind, crumpling, the sound of an engine, a vibrating reed, bowing, scraping, etc., and even in these cases the notion of a discrete event (for example, the beginning, a change of state, arrival at a quasi-steady value…) is often a good match to human perception and therefore very useful.

The danger arises from adopting a worldview in which all music a priori consists of discrete events. Western music notation, the MIDI protocol (Moore 1988), and most music software support this worldview by providing notes and other events as primitives. However, musical meaning and even rhythm can also be conveyed by continuous shapes of time with no clear division into distinct events. Martin Clayton makes a distinction in the context of North Indian râg singing between *syllabic* style, in which each vocal utterance is a distinct rhythmic event at a specific time point in the rhythmic structure, versus *melismatic* style, in which the singing is mainly about melodic connections between pitches and much less about marking time points (Clayton 2000, 48-52). Eric Scheirer's influential paper critiques models of music perception that proceed bottom-up via a stage that represents music entirely in terms of notes, which he terms the "transcriptive metaphor" (Scheirer 1996).[26]

I will follow common usage by using terms like "a sound" or "the sound" to refer to these discrete sound events. The fascinating question of how our minds organize perceived sound into these discrete events is part of the question of *auditory scene analysis* (Bregman 1990).

The subjectivity in the perception of discrete events manifests as difficulties in trying to establish perceptual "ground truth" from human listeners. For example, Leveau, Daudet, and Richard each hand-labeled the beginning times of all of the discrete musical events in 17 short musical

[26] Scheirer's view was supported by Gouyon's comparison of the performance of 12 beat-tracking algorithms on a common database of music (Gouyon 2005): he found that in general the algorithms that began by looking for discrete notes performed worse than those that worked directly from frame-by-frame features.

excerpts (6 to 30 seconds) in a variety of musical styles. Even though they all used the same software tool to perform the labeling, which they had themselves written, they often disagreed not only on the exact timing of the events, but also on the number of events in the excerpt. (Leveau, Daudet, and Richard 2004). Tanghe et al. came across the same problem trying to get expert percussionists to mark all discrete drum events in various recordings: "Brushes for example have a typical "dragged" sound which is hard to annotate as a single percussive event. In this case most annotators chose to register the accents of the brush sounds. Snare rolls do consist of a series of discernable percussive onsets, but it's very hard to annotate the many fast strokes accurately. The same is true for "flammed" drums (typically the sbare drum) where to hits of the same drum type are deliberately played almost (but not quite) at the same time, leading to the sensation of a ghost note occurring slightly before a main note" (Tanghe et al. 2005, 54).

2.3.1 Example: The Piano

The vast majority of quantitative studies of musical timing focus specifically on music played on the piano (Palmer 2005) (Repp, Windsor, and Desain 2002), usually on the European art music from the 1800's (Clarke 1995; Dixon and Goebl 2002; Dixon, Goebl, and Cambouropoulos 2006; Honing 2006; Scheirer 1995) (Repp 1998; Sundberg, Askenfelt, and Frydén 1983) (Zanon and Poli 2003) (Dixon and Goebl 2002) (Dixon, Goebl, and Cambouropoulos 2006) (Repp 1995) (and countless more). The piano's 88 keys each map to a distinct pitch; all that can be done to a note is start it (by choosing a key and depressing it at a certain speed) or stop it (by releasing[27] the key). In other words, a piano performance consists entirely of discrete notes. The piano's ubiquity in western music education (combined with the emphasis on written scores) has helped spread the dangerous misconception that music consists only of a sum of notes.

2.3.2 (Counter?)Example: Indo-Pakistani Vocal Alap

Sound example *shafqat-alap-derbari* is a single phrase excerpted from the *alap* section of a performance of *Rag Derbari Kanra* by the singer Shafqat Ali Khan.[28] Figure 6 displays the fundamental frequency and amplitude of this phrase as functions of time. Where are the perceptually salient discrete events in this example? Perhaps the most obvious events are the

[27] When releasing a piano key, the key is coupled directly and continuously to the damper, whereas when pressing a piano key there's a certain point at which the hammer is "thrown" at a certain velocity, after which the pianist relinquishes control. Therefore a pianist has more control over the exact shape of the end of a piano note than over the beginning.

[28] I had the pleasure on a few occasions of working and performing with this extraordinary musician; see (Wessel, Wright, and Khan 1998).

beginning and end of the entire phrase, at the transition points between zero and nonzero amplitude.

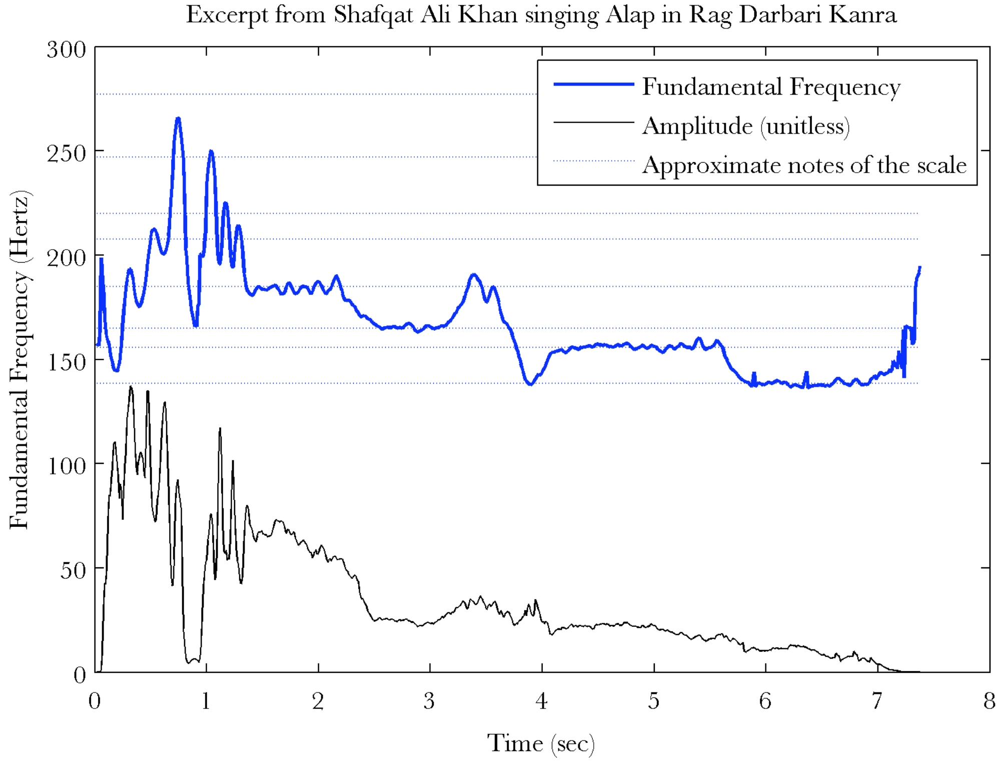


Figure 6: Pitch and amplitude of a phrase from an alap in Rag Derbari Kanra sung by Shafqat Ali Khan. The blue upper line is fundamental frequency (in Hertz) and the lower black line is amplitude (unitless). The dashed lines show the frequencies of the notes of the scale as approximated by equal temperament. Where are the discrete notes?

Other salient events begin around times 1.5, 2.6, 4.2, and 5.8 seconds, the relatively long and flat segments of the frequency envelope where the singer holds a steady note. However, upon close inspection (as shown in Figure 7) there is no obvious exact instant when these notes begin. In each case Shafqat reaches the new pitch via a microtonal glissando from the previous pitch. Around time 2.5 (left plot in Figure 7) we see that Shafqat slightly overshoots[29] the target pitch, so that the local minimum around time 2.58 is perhaps the first half-cycle of vibrato. Around time 4.1 (right plot in Figure 7), we see instead the opposite situation, where the slope of the upward glissando changes at around time 4.06, after which the average pitch continues to climb until about time

[29] By my use of the term "overshoot" I don't at all mean to imply an error or that Shafqat failed to meet some ideal of precision. (Furthermore, his ideal of intonation is certainly not equal temperament.) He is, in fact, one of the most accurate singers I've ever heard. My point is only about the difficulty of choosing a single instant as "the" point of arrival at each note.

4.5, but now modulated by a vibrato process. My point is that these long-held steady notes are clearly salient musical events, but that it is very difficult, and ultimately a matter of interpretation, to pinpoint a single instant at which one of these events begins.

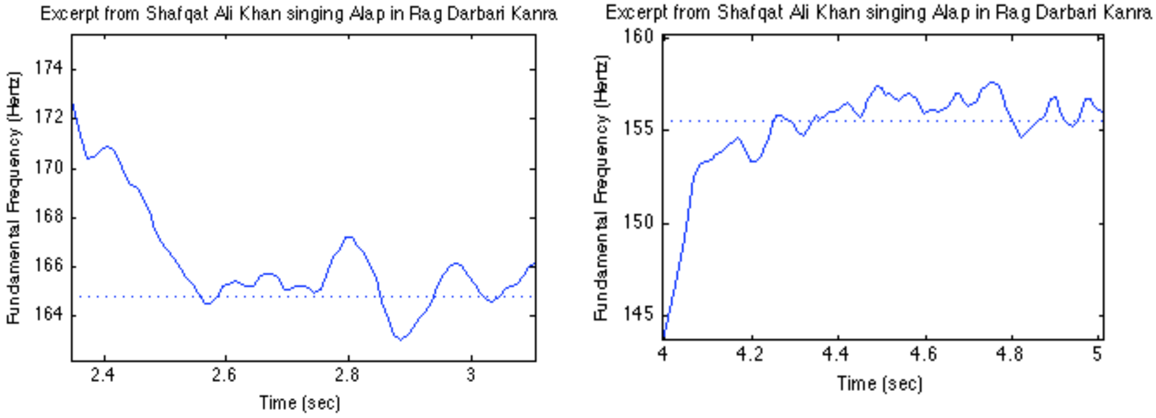


Figure 7: Detail of two "moments" of arrival at steady pitches from Figure 6

What about the virtuosic flourish at the beginning of this phrase? Within the first second we see five relatively large local maxima of amplitude that are approximately equally spaced in time, and it is possible to hear these as five fast notes. On the other hand, it is also possible to hear this entire first second as a single opening gesture, continuous functions of pitch, amplitude, and timbre that have musical meaning only in their entirety.

The next section, from about time 1 to time 1.5, is an ornament leading to the first steady note of the excerpt. Again, this can be heard as discrete notes, but to my ear the beauty of this example lies in the exact shape of pitch as a function of time, especially its connectedness.

2.3.3 Onset Detection

The beginning of a musical event is called its onset, and onset detection is the task of automatically finding onsets in an audio signal. Applications of onset detection include tempo and meter tracking (Beek, Peper, and Daffertshofer 2000; Cemgil and Kappen 2002; Collins 2004a; Desain 1989, 1992; Desain and Honing 1992b, 1999; Desain, Honing, and Rijk 1989; Dixon 2001; Jensen and Andersen 2003; Large and Jones 1999; Large and Kolen 1994; Large and Palmer 2002; Nagai 1996; Nava 2004; Seifert, Rasch, and Rentzsch 2006; Seppänen 2001a, 2001b), analysis of expressive timing in recordings (Clayton 2000; Clayton, Sager, and Will 2004; Scheirer 1995), compositional techniques for real-time sampling and rhythmic playback (Collins 2004a, 2004b), temporal rearrangement of recorded music (Jehan 2004), transformation of rhythmic features of recorded music (Gouyon, Fabig, and Bonada 2003), adaptive sound effects processing, computer accompaniment systems (Grubb and Dannenberg 1997, 1998), changing the relative volume of various drum sounds in a commercial stereo recording (Yoshii, Goto, and Okuno 2005), sound segmentation (Collins 2005c; Smith 1994), time scaling[30] (Duxbury, Davies, and Sandler 2003), spatial feature extraction (Supper, Brookes, and Rumsey 2006), "intelligent" user interfaces for working with audio (Chafe, Mont-Reynaud, and Rush 1982), audio compression (Samsudin et al. 2006), Music Information Retrieval (generally in the service of beat tracking), and automatic transcription (FitzGerald 2004; Hainsworth 2004; Klapuri 1997, 2004; Marolt, Kavcic, and Privosnik 2002; Moorer 1975; Schloss 1985).

This subfield of signal processing has taken off greatly in recent years (Abdallah and Plumbley 2003; Bello et al. 2005; Collins 2005b; Dixon 2006; Duxbury et al. 2004; Hainsworth and Macleod 2003). Most onset detection systems have the same basic architecture: optional *preprocessing* (often tuned to match human perception, for example, frequency-dependent amplitude scaling according to equal loudness contours), followed by a series of DSP operations that compute one or more *detection functions* such as energy or spectral centroid (see Section 5.1.1 "Detection functions" on page 125) whose amplitudes are supposed to increase sharply near the time of event onsets, and finally a *peak picking* step that selects (and perhaps combines) instants from one or more detection functions, usually based on some kind of fixed or adaptive threshold (Bello et al. 2005, 1036). Most methods look for an increase of energy in one or more frequency bands or a redistribution of energy (for example, towards higher frequencies) (Alonso, David, and Richard 2004; Collins 2005a; Grubb and Dannenberg 1998; Jehan 2005; Klapuri 1999; Smith and Fraser 2004), though there are also methods based making sense of fundamental frequency envelopes (Collins 2005d) and on phase continuity (Bello et al. 2005; Bello et al. 2004; Bello and Sandler 2003; Dixon 2006; Duxbury et al. 2003a, 2003b; Duxbury, Sandler, and Davies 2002).

Collins distinguishes two possible aims of automatic onset detection: either to match human perception of event onsets or to "reverse engineer" the input audio to determine "all distinct sound producing events" (Collins 2005b). Either way, the correct answers (i.e., the "true" onset times, usually called the *ground truth*[31]) inherently come with some uncertainty. Researchers

[30] A commercially important special case of this segments looped drum patterns into individual drum notes, as performed by the programs *Recycle!* (from Propellerhead software in Stockholm: http://www.propellerheads.se/products/recycle) and *Acid Pro* (developed by Sonic Foundry and later sold to Sony: http://www.sonycreativesoftware.com/Products/product.asp?pid=383). This enables transformations such as tempo change simply by changing the times at which each individual note is played, avoiding the difficulty and possible artifacts of other methods of time-scale modification.

[31] The phrase "ground truth" comes originally from analysis of aerial photographs and satellite imagery, in which conclusions drawn from such images are double-checked by information collected on site, in other words, on the

generally evaluate their onset detection algorithms with respect to ground truth created by painstaking manual annotation by expert listeners: a detected onset is usually considered correct when it is within 50 ms of an onset found by the human expert. (Leveau, Daudet, and Richard 2004) discusses some issues of subjectivity and inter-subject variability. Even if we ignore the inherent subjectivity of human perception and consider only the timing of sound production, there is some arbitrariness in the selection of a single instant as the time that a note begins. Of course with a purely synthetic tone we can define the onset as the instant of the first nonzero sample generated by the synthesizer, but with natural sound the situation is less clear. One factor is the presence of the noise floor: distinguishing random variations in background noise from true onsets requires some kind of method with inherent uncertainty. Even if we could observe the exact mechanics of sound production there would still be uncertainty; consider the case of plucking a stringed instrument. Does each note begin when the plectrum first touches the string, or when the plectrum releases the string, or when the body of the instrument begins to radiate the energy caused by the pluck, or some other time?

The way the onset detection problem is generally posed, the input is a continuous audio signal and the output is a sequence of discrete instants at which onsets were detected; this formulation fails to acknowledge the uncertainties inherent in the task of labeling event onsets. Many systems avoid the peak-picking step and instead directly use a continuous signal whose magnitude indicates the degree of "onsetness" at each moment of the signal, such as some combination of the detection functions mentioned above. For example, many researchers have modeled metric structure directly from the raw audio, without first segmenting the input signal by onsets (Atlas 2003; Atlas and Shamma 2003; Davies and Plumbley 2004; Eck 2001, 2002a, 2002b; Eck and Casagrande 2005; Large 2000; Scheirer 1998; Todd 1994). These are able to take advantage of a larger amount of data in forming a model of the sound's temporal structure, but are not useful for segmentation or many of the other applications of onset detection.

2.3.4 Perceptual Attack Time (PAT)

A musical event's *Perceptual Attack Time* ("PAT") is its perceived moment of rhythmic placement. Note that this is a subjective, perceptual parameter. For highly percussive sounds, the perceptual attack time might be the same as, or just a few milliseconds after, the onset time, but for sounds

ground. The term has made its way into the jargon of machine learning to refer more broadly to any data for which the correct "answers" are known.

with a slow attack, for example, bowed violin, the PAT might be dozens of milliseconds after the onset. Chapter 3 discusses PAT in much greater detail.

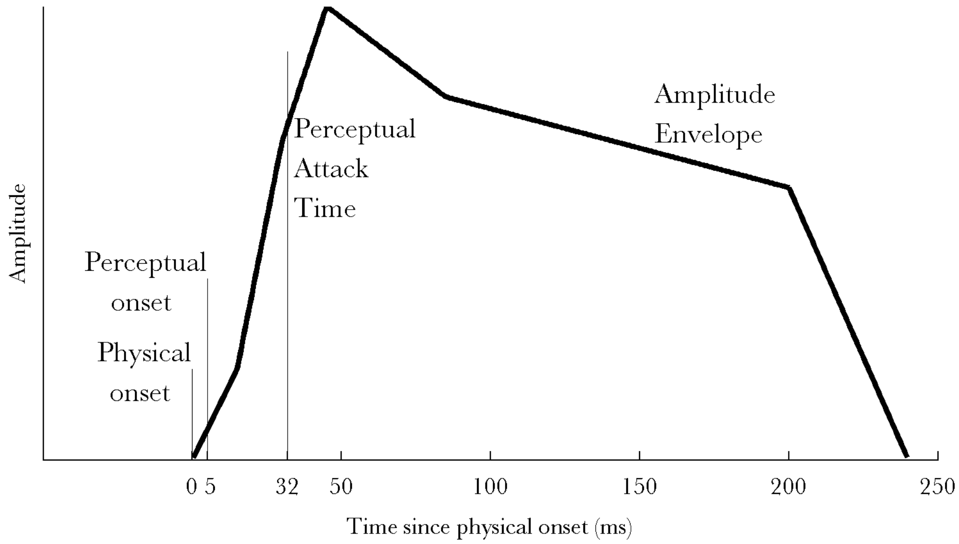


Figure 8: Amplitude envelope of a hypothetical sound, displaying hypothetical physical and perceptual onsets and perceptual attack time.

In contrast to perceptual attack time, an event's *physical onset* is the actual acoustic beginning of the event, that is, the moment that the event's amplitude first becomes greater than zero.[32] The *perceptual onset* is the moment at which a listener can first hear that an event has begun (Vos and Rasch 1981). PAT will in general be after both of these forms of onset. Figure 8 illustrates these three moments for a hypothetical sound event: time zero is defined as the time of the physical onset, the perceptual onset comes shortly later, the PAT is yet later, and the amplitude/energy maximum is even later. The *transient* at the beginning a sound event is the segment of time starting at the physical onset and lasting until the sound achieves some kind of steady state.[33]

[32] For synthesized sound, it is easy to say when the amplitude first becomes nonzero. For recorded acoustic sound there is always a noise floor, and so there is always some ambiguity and/or subjectivity in choosing the instant that the event's amplitude first rises above the noise floor.

[33] There may be other transients later in a sound event and especially at the end. Here is another definition: "As a preliminary informal definition, transients are short intervals during which the signal evolves quickly in some nontrivial or relatively unpredictable way" (Bello et al. 2005, 1036).

Note that all three of the terms "physical onset," "perceptual onset," and "perceptual attack time" have meaning only with respect to a discrete sound event. [34] Onset and attack times are not meaningful for musical sound that is not perceived in terms of discrete events. Furthermore, there are some discrete sound events with clear physical and perceptual onsets that nevertheless do not have perceptual attack times, for example, a sound that fades in very slowly and gradually (say, over two seconds) might be perceived as an event with a relatively clear perceptual onset, but there would likely not be any sense of rhythmic emphasis and hence no PAT.

Perceptual Center or *P-center*[35] is the corresponding concept to PAT in the relatively vast literature on speech, where the discrete event is a *syllable* (Harsin 1997; Howell 1988a, 1988b; Janker 1995, 1996a; Marcus 1981; Morton, Marcus, and Frankish 1976; Patel, Lofqvist, and Naito 1999; Rapp-Holmgren 1971; Scott 1998; Soraghan et al. 2005; Villing, Ward, and Timoney 2007; Villing, Ward, and Timoney 2003; Vos, Mates, and Kruysbergen 1995).

The issue of PAT or P-center first came up in the context of trying to synthesize a musical or spoken phrase by splicing together individually recorded notes or words. Without taking PAT into account, one would naïvely assume that to produce a rhythmically even sequence one should evenly space the beginnings of each individual sound segment. Doing this will not produce a perceptually even sequence unless every segment's PAT is at the same relative delay from its physical onset. As Chapter 5 will suggest, PAT has practical importance well beyond this particular issue.

2.4 Pulsation and Microtiming

Pulsation is a quasi-isochronous[36] series of instants that are possible event attack times. Each pulse can be thought of as a container that holds either the PAT of an event or not, that is, only some of the pulses have notes. In the ideal (a.k.a. "metronomic") case, the pulses are exactly isochronous, and each sound's attack time coincides precisely with one of the pulses. In the real-world case of music performed expressively by human beings there are two complicating factors (Honing 2001; Iyer et al. 1997; Palmer 1997), known collectively as *microtiming*:

[34] For ease of exposition I will use the term "sound" in this chapter to mean "discrete sound event," as in Schloss' and Collins' definitions above.

[35] The spelling "centre" is also often used.

[36] "Isochronous," literally, "equal timed," means "exactly repeating at a fixed frequency." Beware that in computer networking and hardware architecture "isochronous" means "a signal that encodes its own clock."

1. The series of pulses is not exactly isochronous, but instead the frequency changes slowly as a function of time. These "tempo curves" (Desain and Honing 1992a; Honing 2001) have been studied and modeled extensively for modern day performance practice of European music of the Classical and Romantic periods (Todd 1995; Widmer and Goebl 2004); they're generally correlated to the phrase structure that the performer wants to bring out. Many researchers have also charted the variation of tempo in recordings of non-Western music (Bilmes 1993; Clayton 2000; Clayton, Sager, and Will 2004, 29; Schloss 1985; Tzanetakis et al. 2007, 14).

2. Each event's attack time does not exactly coincide with the time of one of the pulses, but instead may be early or late by tens of milliseconds. This "asynchrony" or "deviation" might form a regular repeating pattern as in Jazz "swing" (Collier and 2002; Dixon, Gouyon, and Widmer 2004; Friberg and Sundström 1997; Friberg and Sundström 1999, 2002; Gouyon, Fabig, and Bonada 2003; Lindsay and Nordquist 2006, 2007; Lindsay 2006; Waadeland 2001, 2003) or Brazilian "swingee" (Lindsay and Nordquist 2006, 2007; Lindsay 2006; Wright and Berdahl 2006), "systematic variation" of note durations in, e.g., Viennese Waltz or Swedish folk tunes (Gabrielsson 1982), or Clynes' controversial "composer's pulse" (Clynes 1983) for historical Western composers. In addition, each individual note might have its own deviation from the time of the pulse for reasons including accentuation, differentiation from other instruments, etc. (Iyer 1998). Finally, human motor control is not completely precise, so in addition to all the above intentional factors there is also "motor noise" adding at least about 1 ms of random jitter to each note's attack time even for the most skilled performers.

Note that these two factors are not independent. For any given sequence of event attack times, one could construct a (possibly wildly varying) tempo curve that predicted all of the event times without any per-note asynchrony, or one could assume a perfectly flat tempo curve and explain all of the actual event times with per-note asynchrony, or any of infinitely many compromises between these two extremes.[37] Bilmes (Bilmes 1993) handled this problem in the case of Afro-Cuban Rumba music by looking at the spectra of the note timings and seeing an obvious low-

[37] Another factor that could explain an irregular sequence of event attack times is different ratio-of-integer-related ideal "note values" for each inter-attack-interval, e.g., the first note was supposed to be two beats, the second note 1.5 beats, the third note 1/3 of a beat, etc. In this case the problem of converting the observed non-ideal times to supposed ideal times is *quantization* (Desain and Honing 1992b; Desain, Honing, and Rijk 1989; Takeda, Nishimoto, and Sagayama 2004).

frequency cluster (which he modeled as a tempo curve) and an obviously high-frequency cluster (which he modeled as per-note asynchrony).

Nonlinear oscillators (Eck 2001, 2002a; Large 1996, 2000; Large and Jones 1999; Large and Palmer 2002; McAuley 1995) are an excellent model of production and perception of pulsation. A nonlinear oscillator has a time-varying phase and frequency (because it is an oscillator), as well as an internal source of energy that causes it to keep oscillating (thereby making it nonlinear). Nonlinear oscillators can *synchronize* their phases and/or frequencies to match an incoming signal, thereby *entraining* to an external pulsation (Pikovsky, Rosenblum, and Kurths 2001; Strogatz 2003); this allows them to "hear the tempo" in a robust way, even in the face of microtiming.

The competing explanation for humans' production and perception of pulsation is the "Timekeeper" model (Beek, Peper, and Daffertshofer 2000; Palmer 1997; Wing and Kristofferson 1973a, 1973b), which posits that some part of the brain functions like a ticking clock, outputting events separated by a fixed interval (plus timing noise) that then drive motor behavior.

2.5 Learning

Figure 9 depicts the main feedback loop in the process of learning to play music. This is what might be called the "forward model," in which we assume that the musician starts with an idea of what sound to make ("Intention").[38] The brain translates these into nervous system messages that propagate to various muscles ("motor control"), causing them to contract in varying degrees as functions of time ("body") so as to control a musical instrument ("instrument").[39] This then produces a sound (or not), to which the musician listens[40] ("perception"). Some part of the brain then compares what the musician perceives to what was intended. The gap between intent and perception then drives the learning process, as the musician consciously and unconsciously adapts[41] the mapping from intention to motor control.

[38] A more complete model would take into account the very important exploratory aspects of discovering sounds that can be produced. There can be a sense of discovering the possibilities of an instrument in an acoustic space, rather than a predefined sonic goal to reproduce. Wherever the sources and dynamics of the musician's intention, the rest of the feedback loop works as described here.

[39] Here the term "instrument" includes the human voice, clapping, etc.

[40] Depending on the instrument, the musician might also feel the effects of touching the instrument, see the instrument, or, in principle, smell or taste it. Visual and haptic feedback are optional, yet sometimes very important, aspects of musical instruments, and when present they are part of this feedback loop. I believe that predominantly a musician's intention is an idea of what *sound* he or she wants to produce, though in many interesting cases music is organized (at least in performers' minds) according to the body's movement patterns on a particular instrument; see (Baily 1991).

[41] The musician's body itself also adapts as a result of practice, e.g., guitarist's calluses, wind players' strong embouchures, etc.

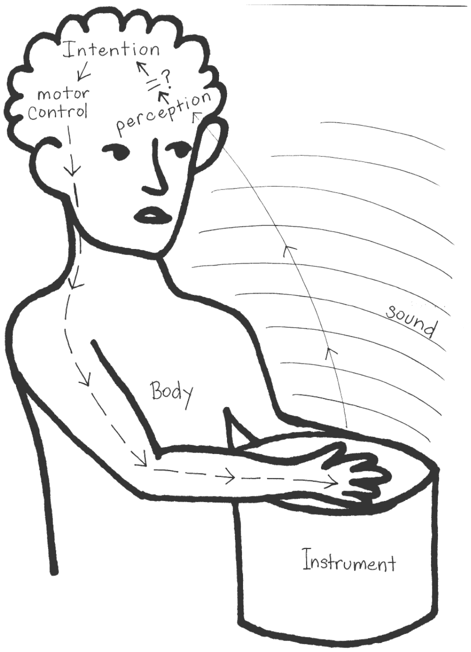


Figure 9: The primary feedback loop in learning to play music.[42]

I want to emphasize the role of Perceptual Attack Time in this process. It is an integral part of the "perception" link, and therefore part of the entire feedback loop. There are already tens or hundreds of milliseconds of lag time between when the brain issues motor control commands and when the body actually produces sound, due to the slowness of chemical message propagation in the nervous system, the body's and instrument's inertia, and other physical considerations. Our brains are very good at generating motor commands the appropriate amount of time before we want an action to take effect, and at fine-tuning these kinds of time relationships when learning various physical activities.

Therefore a violinist, for example, has learned to time her motor behavior so that notes' perceptual attack times follow the desired rhythm, not the notes' physical onsets. The delay between a note's onset and its PAT is just another lag that the brain learns to compensate for; that is why, for example, pipe organists can learn to play accurate rhythms even when there are hundreds or thousands of milliseconds of delay between pressing a key and hearing the corresponding note.

Therefore any fine-grained empirical study of musicians' timing must take PAT into account. Onset detection has great promise for computer-assisted analysis of musical recordings, but we also need computational models of PAT.

[42] Thanks to Michelle Logan for drawing this figure.

2.6 Musical Meter

Justin London's masterful book (London 2004) is the best explanation of musical[43] meter, and I can only crudely summarize it here. Meter involves pulsation no faster than about 100 ms (the psychological limit on perceiving very fast distinct events) and no slower than about 6 seconds (the duration of the "perceptual present" (Clarke 1999), a form of short-term "echoic" memory during which our minds are able to recall fairly exactly what we have just heard). Meter requires at least two levels of pulsation (also called *metric levels*) at different rates, with well-defined phase and frequency relationships among all levels; in other words a single stream of pulses is not in itself meter. The "main" or most salient metric level is the *beat* or *tactus*, and is usually defined as the metric level to which most listeners will tap their feet.

London's "many meters hypothesis" is that every possible hierarchical arrangement of metric levels is a distinct meter, so instead of the relatively small number of time signatures used in western notation (2/4, 3/4, 4/4, 6/8, etc), he considers there to be dozens of distinct meters: for example 4/4 with only eighth notes is one meter, but if one eighth note were replaced by a pair of sixteenth notes it would become a different meter.[44]

Curt Sachs introduced the terms "divisive" and "additive" to describe meter (Sachs 1953). [45] *Divisive* meter is what we usually see in Western music, in which longer metric units consist of an integer number of quasiequal shorter metric units.[46] For example, in 4/4 meter a bar contains 4 beats, each beat takes the time of two 8th notes, each 8th note takes the time of two 16th notes, etc. In *additive*[47] meter, longer metric units consist of sequences of nonidentical shorter metric units, for

[43] The concept of meter also applies to poetry and speech; the parallels between spoken meter and musical meter are numerous, interesting, and beyond the scope of this work.

[44] For readers not familiar with the logic of time signatures in Western notation, "4/4" simply means two metric levels, with the slower one at four times the period of the main one. (In other words, every fourth beat is the beginning of a metric cycle.) Terminology such as "quarter note" and "eighth note" come from a bias towards treating 4/4 as the standard meter, so that a "whole" note has a duration of four beats.

[45] London instead uses "isochronous" and "nonisochronous" respectively for these ideas; I prefer Sachs' terms because of their familiarity to many musicians and because "isochronous" already has enough other meanings. Also, even in additive meters, there is usually a fast quasi-isochronous underlying pulse grouped into sequential units of twos and threes. The distinction is that in divisive meters *all* metric levels are quasi-isochronous, while in additive meters only some metric levels are quasi-isochronous.

[46] The "metrical well-formedness rules" of Lehrdahl's and Jackendoff's influential "Generative Theory of Tonal Music" (Lerdahl and Jackendoff 1983, 69-74) consider only divisive meters to be "well-formed."

[47] People also refer to additive meter as "Balkan rhythms," because these meters are used extensively in the dance music of Albania, Bulgaria, Greece, etc. Since this kind of meter is also used extensively outside of the Balkans (e.g., in North Africa, the Middle East, Iran, Afghanistan, India, etc.), I prefer to avoid that term. Another synonym I prefer to avoid is "odd time": although many additive meters would be written in Western notation with a time signature with an odd numerator (e.g., 5/8, 7/8, 13/8…), other additive rhythms would be written as 10/8 (3+2+2+3 or 2+3+2+3 or…) or 12/8 (2+2+3+2+3, etc.). Likewise, 9/8 (subdivided as 3+3+3) and 3/4 are time signatures with a odd numerators but are in fact divisive meters.

example, a seven-beat metric cycle might consist of a group of three beats followed by two groups of two beats each.

A *subdivision* is a quasi-integer frequency relationship between metric levels. In the divisive case this is known as *binary* or *duple* when the integer is 2 or 4, and *ternary* or *triple* when the integer is 3. In additive meter the units being added together are usually groups of either two or three of the next faster level of pulses. Part of the reason for this may be that a group of four naturally divides into 2+2, a group of five naturally divides into 2+3 or 3+2, etc.

2.6.1 Downbeat

The beginning of a metric cycle is the *downbeat*.[48] Many theories of rhythm say that the downbeat is a "strong" beat and that it (that is, any event placed on it) should always be emphasized. Although this is true for some (particularly Western European) styles of music, it is certainly not true in general. Although the misconception that a downbeat must necessarily be stressed was debunked over 30 years ago (Kolinski 1973) it still manages to persist; in particular, almost every computational model of meter estimation is based on this assumption (or at least on the assumption that they will tend to be stressed), for example, (Alonso, David, and Richard 2004, Dixon, 2001 #134, Eck, 2002 #2; Eck 2001, Goto, 1999 #143, Jensen, 2003 #130; Jensen and Andersen 2003, Scheirer, 1998 #96; Lee 1985). Chernoff points out that for many if not most African musical cultures, musicians will tend *not* to articulate the downbeat (Chernoff 1979, 47-9). He also notes that musical phrases in Africa will tend not to *start* on the downbeat, but to *resolve* to the downbeat on the final note; this relates to the Indian (Clayton 2000) and Afghan (Baily 1988) *tihai/seh* practice of resolving rhythmic complexity by completing a pattern with the last note on a downbeat.)

Arom reports that "Central African music… uses neither the notion of 'measure' nor the strong beat involved in this notion" (Arom 1989, 92). Although this music is certainly based on repeating rhythmic figures and at least one level of pulsation, the claim is that there is no "zero phase" point means that this music is actually not metric according to London's definition.

[48] The etymology of the "down" part of "downbeat" comes from the motion of the conductor's baton in the Western orchestral setting (Rushton); likewise an *upbeat* is the beat just before the downbeat. This correspondence between direction of hand motion and metric position comes from the 15th and 16th century notion of a *tactus* (what I call a "beat") which "comprised two hand motions, a downbeat and an upbeat (*positio* and *elevatio*, or thesis and arsis)" (Brown and Bockmaier). Beware that some people use "downbeat" to mean what I call "beat," in other words, any primary pulse, not necessarily the beginning of a metric cycle, for example, (Eck 2002a).

2.6.2 Syncopation

Syncopation (also known as “offbeat”) is a phenomenal accent or the existence of a sound event at all in a metrically weak position as compared to the nearby metrically strong position (in other words, lying between the pulses of a higher metric level). Syncopation operates by defying the expectation that every beat will be articulated by a sound event. Without a sense of beat there can be no syncopation. Syncopation against additive meters brings its own challenge, because the pattern of metrically accented time points is itself irregular. Syncopation is considered a special challenge for beat tracking (Desain and Honing 1994), and many beat trackers and quantizers work by searching for the least syncopated metric interpretation of the input.

This raises a question: if a given rhythm can be interpreted as being less syncopated by shifting the sense of the downbeat to a different phase point in the main metric cycle, why not shift phase in this way? What determines the downbeat in these cases? Sometimes the music starts with less syncopation, establishes the metric framework, and then listeners retain that sense of meter as the syncopation increases. Sometimes there might be conflicting metric clues such as harmonic changes or simpler parts played by other instruments that continue to establish the “correct” metric interpretation. In other examples the answer is determined purely by standards specific to each musical culture. A famous example is “the” African 12/8 bell pattern, known by many other names including “the standard pattern” (Agawu 2006; Anku 2000; Chernoff 1979; Iyer 1998; Pressing 1983; Temperley 2000; Toussaint 2003).[49] This pattern is very syncopated and open to multiple plausible rhythmic interpretations.[50] To those not familiar with these musics it may still be easy to hear that the pattern repeats every 12 pulses, but there is no way to choose one of those pulses as the downbeat or to decide whether the beat comes every 2, 3, 4, or 6 pulses. Yet not only does this bell part fit unambiguously within a metric framework, it actually *defines* the metric framework, providing a reference point against which other instruments might syncopate even further. Cuban music uses syncopated *clave* parts in the same way (Toussaint 2002; Tzanetakis et al. 2007). In Western culture the closest equivalent is the *backbeat* of rock and funk music, played (usually by a snare drum) on beats 2 and 4 of a 4-beat metric cycle. Sometimes there may be a great deal of syncopation, yet the sense of meter is retained not by events on the downbeat but by these backbeats.

[49] This pattern could be notated as x-x-xx-x-x-x, which reveals the downbeat, or as x-x -xx -x- x-x, which reveals both the downbeat and the “correct” way of grouping into four groups of three. This pattern is also used extensively in Afro-Cuban and Afro-Brazilian musics.

[50] In fact, some have suggested that this ambiguity is a source of musical opportunity, and might explain the popularity of this particular bell pattern (Toussaint 2005).

2.7 Accents

An *accent* is "A stimulus (in a series of stimuli) which is marked for[/by] consciousness in some way" (Cooper and Meyer 1960; London 2004).[51] Lehrdahl & Jackendoff's (Lerdahl and Jackendoff 1983, 17) taxonomy consists of *phenomenal* accents ("points of local intensification caused by physical properties of the stimulus such as changes in intensity, simultaneous note density, register, timbre, or duration"), *structural* accents ("points of arrival or departure in the music that are the consequence of structural properties such as tonality-the cadence being the most obvious example"), and *metrical* accents ("time points in music that are perceived as accented by virtue of their position within a metrical scheme" (Clarke 1999, 482)).

As mentioned in the previous section, an event on a downbeat (which by definition is metrically accented) is not necessarily phenomenally accented.[52] What then is the epistemology of the downbeat? I'd say that although there are plenty of cases where the downbeat is in fact marked by some kind of phenomenal accent[53], or the downbeat is established by phenomenal accents and then continues to be felt as a downbeat even when it is not phenomenally accented, the perception of downbeat and of metric accent in general can be completely subjective and is culturally learned. For example, in the highly syncopated *Maracatu de Baque Virado* ("Maracatu of the Turned-around Beat") drumming from the Northeast of Brazil the downbeat is often indicated very clearly to properly enculturated listeners by a bell part that plays not on the downbeat but one sixteenth note after the downbeat (Crook 2005, 164).[54]

This circularity in explanation of the downbeat ("it's (metrically) accented because I hear it as the downbeat / I hear it as the downbeat even though it's not otherwise accented") has a parallel in the phenomenon of "subjective rhythmicization,"[55] in which listeners presented with a perfectly isochronous sequence of perfectly identical sounds (for example, the ticking of a clock or the dripping of a faucet…) imagine an alternating sequence of accented and unaccented sounds. I believe this shows an innate human propensity to organize perceived sound metrically, even when there is no objective basis to do so.

[51] Composers and performers mark events "for" consciousness, as Cooper and Meyer say, but for listeners the events are marked "by" consciousness, as London amends their definition.

[52] In fact there may be no note at all on the downbeat, the extreme case of lack of phenomenal accent.

[53] Indeed, the success of computational downbeat-finding models built on this assumption indicates that case is common in the musical examples used to test these systems.

[54] This bell part also has notes on the second beat and on the 8th note after the second beat; it might be notated as ".x.. x.x." using "x" for a note and "." for a rest.

[55] Better terms for this include "subjective metricization" and "subjective accentuation."

2.8 Phrasing

Phrasing is the perceived connection among discrete events that occur nearby in time, usually together within the perceptual present.[56] One question is how our minds group perceived events into phrases. Lerdahl's and Jackendoff's *Generative Theory of Tonal Music* includes many rules for grouping musical notes together into phrases (Lerdahl and Jackendoff 1983). Bregman has another set of rules for grouping events into auditory streams (Bregman 1990). Todd's "rhythmogram" (Todd 1994) is a graphical representation of musical material that brings out grouping structure. Rothstein's "tonal phrases" require harmonic motion by definition (Rothstein 1989).

Saying which groups of events "group" together is only part of the story about phrasing; there are also musical parameters that vary over certain shapes across entire phrases. For example, most attempts at computer-generated stylistically correct renditions of piano music vary the tempo systematically over the course of each musical phrase (Friberg 1995; Sundberg, Askenfelt, and Frydén 1983; Sundberg, Friberg, and Bresin 2003; Widmer 2002; Widmer and Goebl 2004). Bilmes' system for machine learning of microtiming in the Afro-Cuban rumba genre used a sophisticated form of lookup based on the assumption that similar phrases will have similar microtiming (Bilmes 1993).

2.9 Prediction

Finally, I want to emphasize the role of prediction by listeners (and performers) in musical rhythm. Our brains do not just wait passively for information to trickle up through the auditory nerve and form events, phrases, meter, and other structure; they actively predict what we will hear through top-down models as well as these bottom-up processes (Slaney 1997). Hawkins goes so far as to define intelligence completely in terms of prediction (Hawkins and Blakeslee 2004). There are just as many nerves running from our brains back to our ears as in the other direction, and our cochleas are not just passive encoders of incoming sound, but nonlinear active systems with feedback, giving us increased frequency and amplitude resolution perhaps by fine-tuning the mechanics to focus on what we expect to hear (Zwicker and Fastl 1999).

David Huron has proposed a wide-reaching model of expectation that illuminates many aspects of music (Huron 2006). Mari Reiss Jones suggested that the ability to pay attention to incoming

[56] A "phrase" is also a grammatical unit, and musical phrases are related to spoken phrases, for example, the prosody that organizes the words of a spoken sentence to clarify the meaning uses the musical parameters of pitch and timing.

sound is a limited quantity, and proposed a nonlinear oscillator model that uses meter to predict when new events will occur so as to best utilize these attentional resources (Large and Jones 1999). Even the simple delay lines in Scheirer's beat tracker can be interpreted as a form of prediction of what sound will come in the future (Scheirer 1998, 593).

Finally, Snyder and Large were able to measure metric prediction in the brain's gamma-band activity via an EEG. Subjects heard an isochronous sequence of synthetic stimuli with some of the events randomly omitted. While *evoked* gamma band activity depended on whether or not the subject had heard an event, *induced* gamma band activity began *before* the expected time of the event and increased up to that time, even in cases where no event actually sounded.

Chapter 3 On Perceptual Attack Time

3.1 Definition of Perceptual Attack Time

A musical event's *Perceptual Attack Time* ("PAT") is its perceived moment of rhythmic placement. Here are some other definitions:

- "The time [a] sound is perceived as a *rhythmic event*" (Schloss 1985, 23)
- "The perceptual attack time (PAT) is the compensation for differing attack components of sounds, in the case of seeking a perceptually isochronous presentation of sounds" (Collins 2006, 923).
- "The time a tone's moment of attack or most salient metrical feature… or rhythmic emphasis is perceived" (Gordon 1987, 88).

Figure 8 (page 20) illustrates a hypothetical sound event's physical onset, perceptual onset, PAT, and overall amplitude envelope; in general perceptual onset will be at or after the physical onset, and PAT will be at or after the perceptual onset.

The corresponding term in the speech literature is "Perceptual Center" or "P-Center." Even when Morton first introduced this term (Morton, Marcus, and Frankish 1976) he suggested that it also applied specifically to music and to dance, but in the musical literature we have instead been using the synonym "perceptual attack time" (Collins 2006; Gordon 1987).

3.1.1 Example: Short bowed violin

Figure 10 shows the waveform of a short musical note played by bowed violin.[57] We can see that the amplitude increases gradually over about the first 100 ms of the sound. Time zero is the physical onset.

[57] Specifically, this is the "Violin" sound described on page 73.

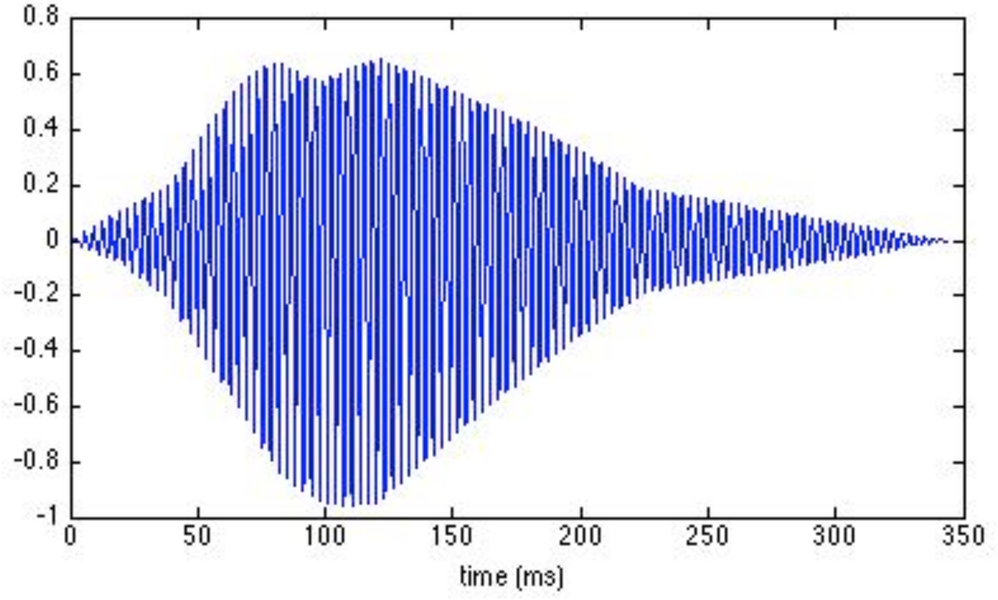


Figure 10: Time-domain waveform of a short synthesized bowed violin note.

According to my personal subjective experience, the PAT of this note is about 34 milliseconds after its physical onset. I chose the number 34 by aligning this sound in time with a percussive sound with a very sharp attack.[58] The following sound examples illustrate this alignment process:

- Sound Example *Violin*: This violin note alone.
- Sound Example *Violin+stick_together*: This violin note mixed with the percussive sound, with the two starting at exactly the same time. In this example, it sounds to me like the percussive sound comes first and the violin comes after. If the percussive sound were a metronome tick that the violinist was trying to play along with, I'd say that the violinist was late.
- Sound Example *Violin_first_then_stick*: The same mix of the violin with the percussive sound, but with the violin starting first and the percussive sound starting 34 milliseconds later. This example sounds rhythmically together to me.

The point is that making two sounds be physically synchronous does not necessarily make them sound perceptually synchronous.

[58] For purposes of illustration I'm assuming that the short percussive sound's PAT is equal to its physical onset. In fact all I have determined by aligning the two sounds is that the PAT of the violin sound occurs 34+*x* milliseconds after its onset, where *x* is the time that the PAT of the percussive sound occurs after its onset. I will take up this issue in detail below.

3.2 Relativity: What is the "zero" point of PAT?

As the X axis in Figure 8 is "time since physical onset," we see that the PAT occurs 32 ms after the note's physical onset. "In practice, a sound's PAT is useful only in how it relates to that sound's time of physical onset or some other sound's PAT. We can thus define *relative* perceptual attack time (RPAT) as the temporal interval between physical onset and PAT" (Gordon 1987, 88).

PAT is defined only for sound events. Each sound event has a physical onset time t_{onset} such that its amplitude $a(t) = 0$ for all $t < t_{onset}$. Note that for any $t_{earlier} < t_{onset}$ it will also be the case that $a(t) = 0$ for all $t < t_{earlier}$. In practice, it is often convenient to define "time zero" as some $t_{earlier}$ that precedes the true physical onset of the event in question, e.g., when we extract an event from a sound recording, there will be some splice point at which we choose to begin the excerpt.[59]

It also makes sense to talk about the PAT of individual events in the context of a longer recording, in which case "time zero" might be the beginning of the entire recording, or to talk about the PAT of events happening live in real-time, in which case actual time-of-day (e.g., "8:33:07.234 pm PST, Sunday August 12, 2007") makes sense.

3.3 Methods of Measuring Perceptual Attack Time

Perceptual attack time is in many ways inherently relative. In terms of PAT's meaning to meter and rhythm, what matters is the spacing among the PATs of the multiple events that make up the rhythm; an isolated single event could reasonably be said to have "no rhythm."

Soraghan et al. measured PAT directly from subject's brains via auditory evoked potentials (Soraghan et al. 2005). After averaging measurements from 1500 trials per sound[60] they fit a 15th degree polynomial to the result and then took the global minimum. The time difference between these minima had a strong and significant correlation ($r=0.9$, $p<0.001$) to relative PAT as measured with the isochrony method (described below), with an average difference of about 35 ms for spoken digits and of only about 7.2 ms for 1 kHZ sinusoids under various amplitude envelopes.

All other known methods for measuring PAT work by measuring the subjective temporal relationship between the sound and something else, either a second sound or a physical action

[59] Even if this is the moment the recorder was turned on.

[60] Subjects had no task but to listen to each sound over and over again. With about 400 ms per sound plus a wait/rest period, each trial took 1.5 seconds, so it took about 38 minutes of listening to find the PAT of each sound.

(usually tapping) performed by the subject. There is always some kind of repeating loop so that the subject can predict when the sound will next occur. When there are two sounds, they are the *test* sound, whose PAT is being measured, and the *reference* sound; in these kinds of experiments the subject adjusts the relative timing of the two sounds until the desired perceived temporal relationship is achieved.

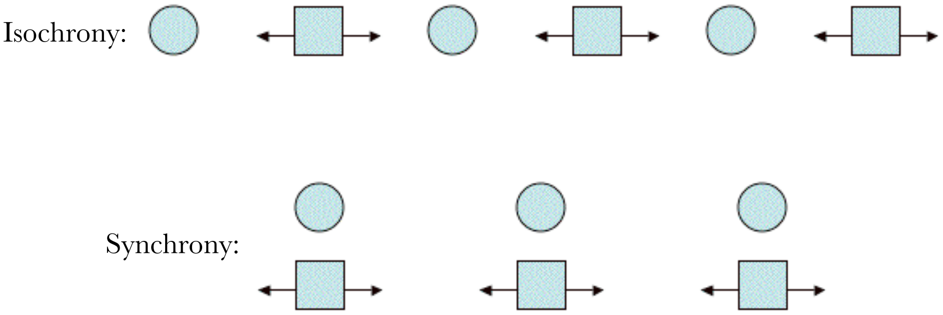


Figure 11: Schematic illustration of the "isochrony" and "synchrony" methods for measuring PAT.

In both cases the subject hears a fixed repeating reference sound (shown as a circle) and a moveable repeating test sound (shown as a square), and adjusts the physical onset time of the test sound relative to the reference sound. In the **isochrony** ***method (above) the subject's goal is a perceptually even alternation of the two sounds; the subject adjusts the timing until the two sound like they are rhythmically alternating. In the*** **synchrony** ***method (below) the subject's goal is for the two sounds to be perceptually synchronous; the subject adjusts the timing until the two sound like they are playing "together."***

There are two such temporal relationships known in the PAT and P-Centre literature, illustrated in Figure 11. *Isochrony* is a perceptually evenly spaced alternation between the two sounds or between the sound and the action. *Synchrony* is a perceptual rhythmic togetherness between the two sounds or between the sound and the action. (This distinction is also known as *alternating* versus *simultaneous* presentation of the two sounds (Collins 2006).)

	Synchrony	***Isochrony***
Tapping	Subject taps at the time of the test sound's PAT.	Subject taps exactly between PATs of consecutive test sounds.
Reference sound	Subject adjusts the relative onset time of the test sound so that the two sounds give the impression of attacking together.	Subject adjusts the relative onset time of the test sound so that the two sounds give the impression of alternating in an even rhythm.

Table 2: The four known methods of measuring PAT

3.3.1 Choosing a Method for Measuring PAT

Every method for measuring PAT is problematic in some way.

The physical act of tapping seems to come with a large amount of temporal variance (Janker 1996b) including the motor noise discussed on page 8. Subjects also tend to tap early (Aschersleben 2002; Dunlap 1910; Vos, Mates, and Kruysbergen 1995) by varying amounts in the range of about 20-80 ms, which must be measured and accounted for.[61] Also, for technical reasons it requires a controlled hardware setup to control latency and jitter in the measurement of subjects' taps; for example, using a computer's QWERTY keyboard to register subjects' taps can mean tens of milliseconds of both latency and jitter depending on operating system and a large number of software configuration options (Wright, Cassidy, and Zbyszynski 2004).

With tapping there is also the question of what exactly it is that the subject times to line up with the sound (Vos, Mates, and Kruysbergen 1995). Some subjects may tap so that the kinesthetic feedback from their finger back to their brain arrives at the moment of the sound's PAT; this is one explantion for why subjects consistently tap early in these kinds of tasks. On the other hand, if tapping generates a sound, either acoustically from the physical act of tapping, or from a reference sound triggered in response to each tap, then the subject is likely to align the tapping sound with the test sound (in isochrony or synchrony depending on the task). Once the subject gets used to the PAT of the sound, he or she will learn to compensate for it as shown in Figure 9 (page 24), at which point this method has the same problems as other methods of specifying the relative timing of test and reference sounds, plus issues of motor noise.

Let's consider methods based on reference sounds. "There is no one signal for which the absolute P-centre is known. Thus, any direct measure of the P-centre location… is impossible, since no signals could be used as the baseline against which the other signals would be compared" (Scott 1998, 6). This requires making various assumptions and approximations to estimate the sounds' PATs, as described in Section 3.5 (page 49).

In choosing between the two methods based on reference sounds, "there are problems unique to each measurement method, and choosing one method over the other is somewhat arbitrary, but PAT is presumably the same regardless of the method used" (Gordon 1987, 90). Collins, however, found differences on the order of 20 ms in two small experiments comparing the two methods to measure the PAT of the same set of sounds: "a subject achieved a correlation score of 0.534

[61] By assuming that each subject's mean tap anticipation time remains constant throughout an experiment it's possible to calibrate for it and subtract it out, but there's no guarantee that it will actually be constant.

between alternating and simultaneous presentation modes for the [first] 25 [sounds], with absolute difference statistics showing an average discrepancy per sound on the order of 20msec, certainly noticeable as a timing change (mean 0.01908, standard deviation 0.01197, max 0.05625, min 0). In a between subjects test, two further subjects showed a correlation of 0.379 and stats of (mean 0.02742, standard deviation 0.0270, max 0.10425, min 0) between their responses on the second group of 25 recorded sounds" (Collins 2006, 925).

One difficulty with the isochrony method is the tradeoff in choosing the period of repetition of the reference sound: "There are interactions between the need to avoid fusion and masking phenomena through sound overlap, and the need to keep the separation between reference and test sound onset small to improve temporal acuity of subjects in judging isochrony (following Weber's law)" (Collins 2006, 924). In the case of measuring PAT for spoken syllables (i.e., P-center) with the isochrony method, Marcus notes the existence of "an order effect bias, resulting from a tendency to rotate the knob further clockwise or anticlockwise than the desired position" (Marcus 1981, 248). Friberg and Sundberg found a just-noticeable-difference of "about 10 ms for tones shorter than about 240 ms duration and about 5% of the duration for longer tones" in an experiment using the isochrony method (Friberg and Sundberg 1993).

The experiment described in Chapter 4 uses the synchrony method with a variety of reference sounds. One problem with this method is *masking*: when the two sounds are played together the louder one might cover the quieter one, making it impossible to hear whether the PAT of the quieter one is in fact at the same time as the PAT of the louder one. I addressed this by giving subjects complete control over the relative volume of the two sounds (as described on page 78), which in turn brings up the potential problem of inconsistency in the stimuli subjects heard: if PAT depends on volume then this is an uncontrolled variable.

A second issue with the synchrony method with pairs of sounds is the tendency for two sounds to fuse perceptually into a single event. At first glance this might not seem to be a problem: if the two sounds are perceptually synchronized so closely that they sound like a single event, then it would seem that the subject has aligned the PATs successfully. In general this is true, but Section 4.4.2 (page 118) discusses the fact that many subjects ended up placing a very percussive click *before* the physical onset of a sound with a slower attack, possibly because the resulting composite sound has a percussive attack followed by a sustaining portion, like many natural musical sounds.

Finally, using the synchrony method to align a sound against an exact copy of itself produces *comb filtering*[62] that often leads to audible spectral differences among the delay times close to zero, even when there is no perceivable rhythmic difference. If subjects attend to this spectral difference rather than to the rhythmic togetherness of the two copies of the sound, then this could bias the results.

The main reason I chose the synchrony method is that it can measure the PAT of sounds in a musical context (for example, a continuous loop of metric music), not just isolated individual notes cut out of sound recordings or synthesized. The isochrony method requires the sounds being measured to be discrete individual sonic events.

3.4 Representing Perceptual Attack Time with Probability Density Functions

A musical event's *Perceptual Attack Time probability density function* ("PAT-pdf") is a statistical probability distribution that gives the likelihood of a listener hearing the event's PAT as a function of time.

3.4.1 Prior Work in Probability Density Function Representations for Music

Desain proposed a causal model of rhythm perception in which the rhythmic structure of events that have already occurred generate an expectancy function giving the model's prediction of the probability of an event as a continuous function of time (Desain 1992).

Grubb and Dannenberg developed a computer accompaniment system that represented the position in the score of a vocalist with a probability density function they termed "score position density." The input to the score follower is current pitch estimate, spectral envelope, and estimate of whether a note onset just occurred, all of which are noisy and might have errors. They did not represent these factors with probabilities, but instead trained their system to produce prior distributions for, e.g., the histogram of likely pitch tracker outputs conditioned on each pitch class in the score. On each iteration the system updates its probabilistic estimate of the score position given the new observations plus its estimate of the current score position and tempo (Grubb and Dannenberg 1997, 1998).

[62] http://ccrma.stanford.edu/~jos/pasp/Feedforward_Comb_Filter_Amplitude.html

Researchers usually evaluate the output of onset detectors by comparing it to ground truth onset times labeled by hand by an expert listener (Leveau, Daudet, and Richard 2004). Of course these results are almost never exactly the same to the level of a digital audio sample, or even one millisecond; it's typical to consider a detected onset correct if it is within 50 ms of a "true" onset. This method is equivalent to treating the output of the onset detector as the mean of a uniform probability distribution (in other words, a rectangular pdf).

As described in Section 2.3.3 (page 17), many systems estimate metric structure from continuous detection functions rather than the discrete peak-picked instants of an onset detector; this could also be interpreted as a stochastic model, with the detection function representing the probability of an onset at each time.

Frieler (Frieler 2004) proposed a beat and meter estimator that starts with a list of discrete instants representing (unquantized) onset times for a musical excerpt. He then gets from this discrete representation to a continuous function with "Gaussification," placing a Gaussian centered on each input point and then adding the results together. He did not, however, give a probabilistic interpretation of the resulting functions.

3.4.2 Motivation for Representing PAT with PDFs

The "perceptual" in "perceptual attack time" means that we can only measure it by recording the subjective impressions of human listeners.[63] Regardless of the measurement method (described in Section 3.3), multiple listeners will not agree on the exact instant of a sound's PAT. In fact, even the same listener will choose a slightly different instant for the same sound's PAT on each of multiple trials.[64]

John Gordon's early investigations of perceptual attack time for musical tones (Gordon 1987) found that different subjects would disagree on the exact PAT of synthetic musical notes, and that each individual subject would respond somewhat inconsistently from trial to trial. He addressed this issue by smoothing (with "a low-pass, zero-phase filter") the cumulative response distributions across all subjects and trials, differentiating ("approximated by a simple first-order difference equation"), plotting, and then interpreting these plots as probability density distributions (Gordon 1987, 94). He found that the shapes of these distributions varied according to certain musical

[63] In other words, PAT is a *subjective* measure, so we need *subjects* to measure it.

[64] Since time is continuous, if subjects have arbitrary control over timing then the probability of getting to the exact same relative alignment between sounds is zero. So we have to consider time intervals. In practice, we know there are perceptual limits on temporal acuity; see (Eggermont 2001; Gordon 1984; Ivry 2004; Krumbholz et al. 2003). So if we look at the histogram of results on the same task; obviously they won't all be sample-accurately the same.)

characteristics of the tones he tested: some are bimodal, some have a “plateau”, many appear approximately Gaussian, some are tall and skinny, etc. For example, “the curve for the flute exhibits a ‘plateau,’ which suggests that the flute's [PAT] may be better represented by a range of values instead of just a single [moment]. There is good reason for this plateau, which is related to the spectrum of the particular flute tone used in the experiment. The fundamental appears several milliseconds after all the other harmonics, resulting in a strong chiff effect. Evidently, some subjects placed more emphasis on the rise of the fundamental in determining the perceptual attack of this tone, while others placed more emphasis on the onset of the second and higher harmonics (or with the rise of the overall amplitude envelope). In other words, we can infer that there is a lack of agreement as to when the perceptual moment of attack occurs for the flute tone” (Gordon 1987, 94). I have reproduced one of Gordon’s figures (with permission) as Figure 12.

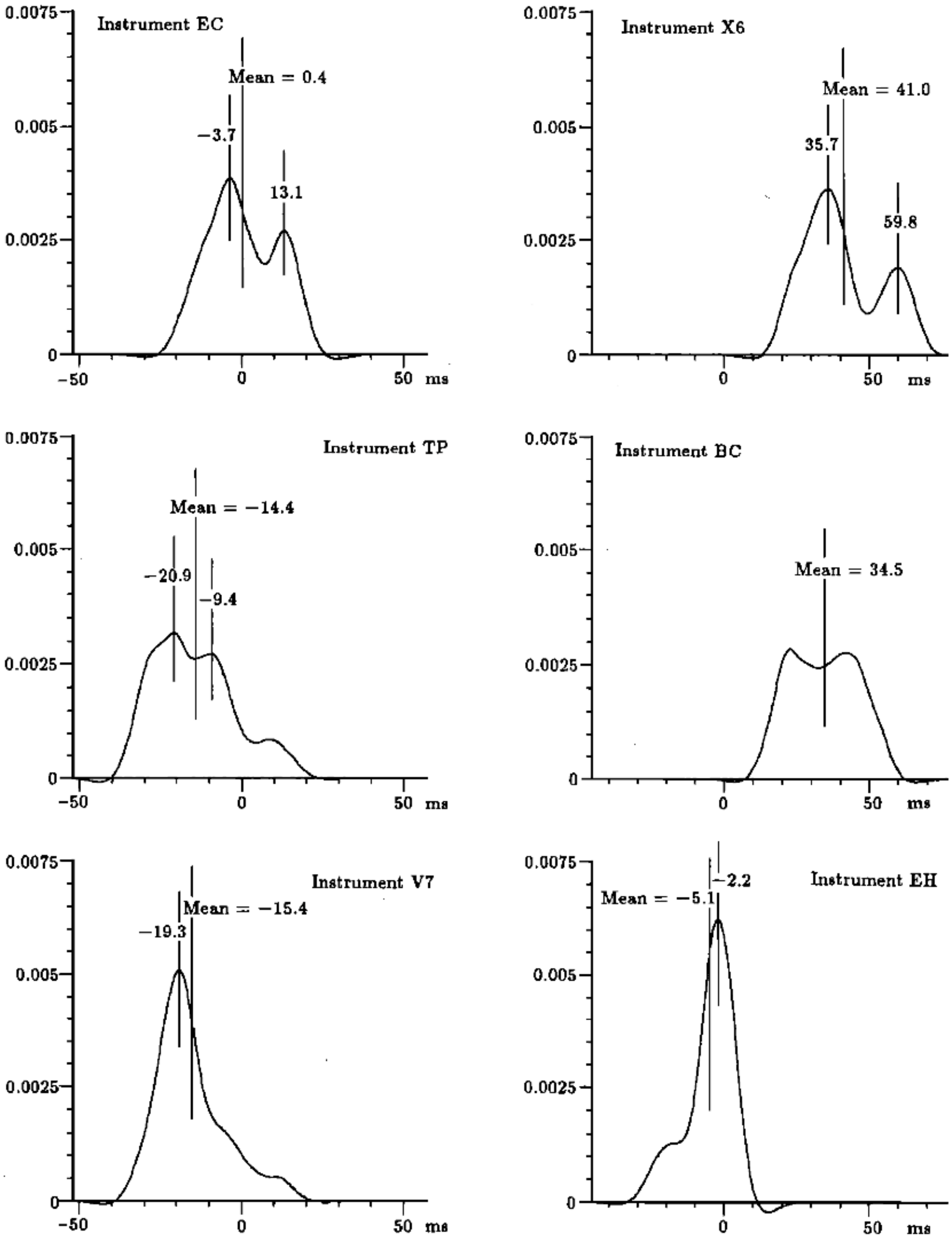


Figure 12: Probability distributions for the results of six measurements of PAT using the* synchrony *method, reproduced from Figure 7 of (Gordon 1987, 96) with the author's permission.

On the left, from top to bottom, are Clarinet, Trumpet, and Violin, against the Clarinet as a reference. On the right, from top to bottom, are Saxophone, Bass Clarinet, and English Horn, against the Bassoon as a reference. All of these sounds are Grey's synthetic orchestral tones as described on page 73. The zero point of the X axis represents the situation in which the two tones'

physical onsets are synchronous. "Vertical lines marking the mean response are shown for each instrument. In addition, other vertical lines mark certain modes."

However, after much discussion of the shapes of these distributions Gordon then turns to the task of modeling PAT from acoustic features of the signal, for which purpose he chooses a single instant as the most representative PAT for each tone.[65] Each of his PAT models therefore outputs a single time instant as the predicted PAT for any given input sound.

Collins also conducted experiments to measure PAT on a collection of sounds. In his first experiment, 14 subjects each indicated the PAT of 25 synthetic tones twice each. "Ground truth was created for the 25 sine sounds by averaging relative PATs from those experimental subjects judged most consistent in their responses. There were six subjects where correlation scores between the first and second repetition were greater than 0.5 and mean absolute difference was less than 20 milliseconds with standard deviation also under 20 milliseconds" (Collins 2006, 926). Again, for modeling purposes the details of the shapes of the distributions of measured PATs were thrown away so as to represent each sound's ground truth PAT as a single instant. In Collins' second experiment, there were 100 sounds including a variety of recorded and synthetic examples. "Given the variability of subject data in the general experiment, and some subjectivity perhaps inherent in the task, it was found most consistent for modeling purposes to use ground truth provided by the author" (Collins 2006).

In the P-center literature researchers generally take the mean result from multiple trials (usually pooled across subjects) as the single instant of P-center (Harsin 1997; Janker 1995, 1996a; Patel, Lofqvist, and Naito 1999; Pompino-Marschall 1988; Scott 1998). Although they sometimes report the standard deviation of their observations, I know of no attempt to model the shapes of the resulting distributions. Marcus proposed a method for combining all results from a full-factorial experiment comparing all possible pairs of sounds to estimate the relative P-center of each individual sound as a single instant (Marcus 1981) (see Section 3.5.4.2 on page 53).

In short, all prior work on measurement and/or modeling considers PAT or p-centre to be a single moment. Gordon's research demonstrates that this model is inadequate, that even for very simple synthesized tones with definite attacks, with no polyphonic context, the histogram of subjects' PAT judgments will take on characteristic shapes for each tone. I believe that the

[65] Gordon made this selection somewhat by hand, taking into account each tone's mean PAT across trials and subjects and reference tones and normalizing to remove the per-reference-tone bias. "Since there were modal values in some of the response distributions that differed from the respective mean... values, a [PAT] value... for each instrument was chosen based on all of the mean and modal... values obtained empirically" (Gordon 1987, 100).

widespread urge to summarize these shapes with single time value is harmful, discarding information that could be useful in later processing steps. Therefore I consider a sound's PAT to be a probability density function rather than an exact single instant. Not only do I observe probability distributions in subjects' measurements of PAT, but I also consider the ground truth to be in the form of these distributions and hence pose the modeling problem (Chapter 5) in terms of estimating these kinds of distributions.

3.4.3 An Event's Intrinsic PAT-pdf

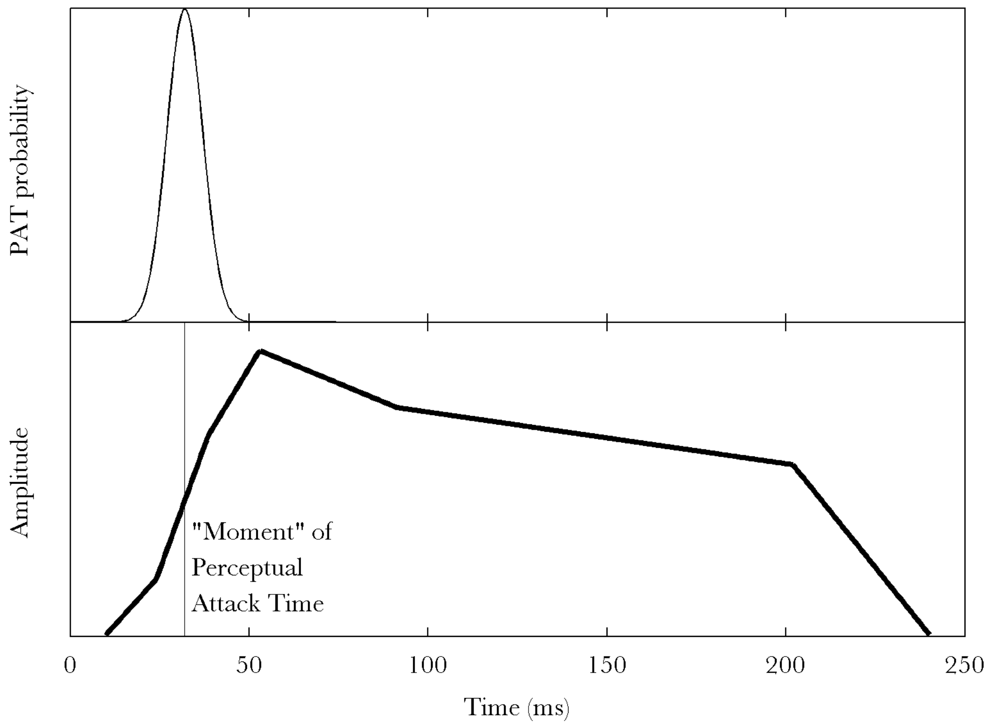


Figure 13: Probability density function of PAT for the same hypothetical sound as Figure 8.

The vertical line in the lower plot marks the moment at 32ms that Figure 8 called "the" PAT. The upper plot shows a hypothetical probability distribution function for the PAT for this example; it is a Gaussian with mean 32 ms and standard deviation 5 ms.

What we have been discussing so far might be called the *intrinsic* PAT of each event, in the sense that it's a property purely of the sound itself, divorced from all context with any other sounds. To my knowledge every existing model of PAT or P-center is based on the assumption that each sound actually has an intrinsic PAT. "The p-centre is context independent (e.g., doesn't depend on neighbouring sounds in a sequence)… All the models being reviewed make [this assumption]" (Villing, Ward, and Timoney 2007). Collins puts it like this: "A practical assumption of this work is that if any algorithm is established for PAT determination of isolated events, this PAT will remain valid even in playback situations with multiple streams" (Collins 2006, 924).

The notation I_A will mean "the intrinsic PAT-pdf of sound event A." I will also sometimes use *mean(A)* or μ_A as a shorthand for *mean(I_A)* and *var(A)* or σ_A^2 as a shorthand for *var(I_A)*.

3.4.4 Statistical Interpretation of PAT Measurement Results

If we[66] believe that each sound event has an intrinsic PAT-pdf, then whenever we measure a sound's PAT in terms of a reference sound, we must consider that the reference sound has its own intrinsic PAT-pdf. The simplest method for dealing with this is to assume that the reference sound has a definite and unambiguous PAT (i.e., an infinitely narrow pdf with zero variance). By arbitrarily defining the reference sound's PAT to be equal to the time of its physical onset, the zero point of all measurements then shifts by a constant equal to the true PAT of the reference sound. Although this is not absolutely correct, it is "relatively correct" in the sense that it measures each sound's PAT relative to a common reference; this suffices for practical scheduling applications[67] or for a comparative study of different sounds' PAT. However, in some situations it is necessary to know the absolute PAT of various sounds, for example, when trying to make use of PAT data measured against different reference sounds. Also, any method for predictive modeling of PAT (see Chapter 5) must take as its input properties of the event's time-domain audio signal, which are anchored to the absolute time axis.

If we treat the reference sound's PAT as a PAT-pdf, the situation becomes more complicated. Let R be a random variable representing the time a listener will perceive the PAT of the reference sound on any given trial, and let T be the corresponding random variable for the test sound. These are each relative to the physical onset of the respective sound. We assume that each sample of R and T comes from the intrinsic PAT-pdfs for the reference and test sounds (respectively), which are what we want to estimate. The perceived amount of time that the attack of the reference sound comes before that of the test sound when their physical onsets are

[66] I will switch to the first person plural throughout "our" statistical treatment of the theory both because it's more formal and as a sort of acknowledgement of my gratitude to the many people who have collectively helped me begin to understand statistics.

[67] As long as every sound is scheduled according to PATs measured against the same reference then the result will have the proper rhythm; the actual PAT of the reference sound corresponds to a uniform time-shift of the entire sequence.

synchronous is another random variable $D_{T,R} = T - R$.[68] (If $D_{T,R}$ is negative it means that when physical onsets are synchronous it sounds like the reference sound comes *after* the test sound.)[69] The only way to measure $D_{T,R}$ directly is to ask subjects to estimate the amount of time between when they hear the two PATs, which would certainly not be very accurate.

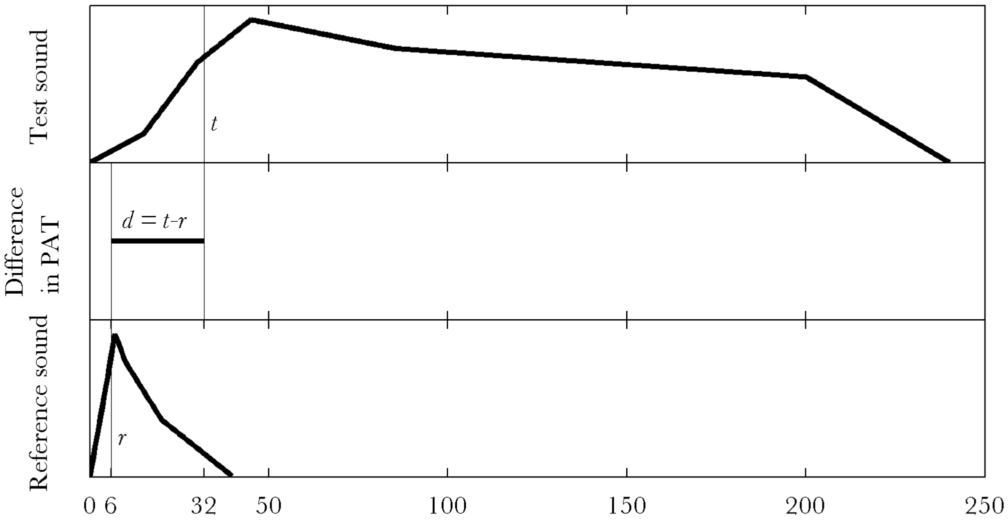


Figure 14: Illustration of what we can actually measure experimentally: the difference between the PATs of the test and reference sounds.

(This hypothetical example represents each sound with a made-up amplitude envelope.) The two sounds' physical onsets are synchronous. For this trial, the PAT of the test sound is* t *and the PAT of the reference sound is* r*. We can measure* d*, the difference between these two PATs.

We assume that on each trial a sample *r* is drawn from *R* and that a sample *t* is drawn from *T*, as shown in Figure 14.[70] The subject controls the time relationship between the physical beginnings of the test and reference sounds: specifically, the subject controls a (possibly negative) delay time *d*

[68] *D*'s subscripts indicate which pair of sounds were compared, e.g., $D_{A,B}$ is a random variable representing trials aligning test sound "A" with reference sound "B."

[69] Gordon use the notation ΔPAT to mean "the amount of time needed to shift a tone away from physical synchrony/isochronism (with some standard tone) in order to attain perceptual synchrony/isochronism (with that standard)" (Gordon 1987, 88).

[70] Note that we assume this sampling happens only once per trial, even though the subject might hear each sound repeated dozens or hundreds of times in the course of a single trial. The deeper assumption is that if the subject hears multiple exact repetitions of the same sound within a reasonably short period of time (perhaps within the "perceptual present" discussed on page 9) then he or she will "choose" the same moment of PAT from that sound's PAT-pdf on each repetition.

between the physical onset of the test sound and that of the reference sound.[71] For any value of *d*, the subject will hear the PAT of the test sound at time *t-d* after the physical onset of the reference sound.[72] In the synchrony method, the subject's task is to adjust the relative physical onsets of the sounds until the perceived time difference between PATs is zero,[73] in other words, the subject finds *d* such that *r* =*t-d*, as shown in Figure 15. Since the value of *d* produced in each trial is equal to *t-r* for that trial, *d* can be thought of as a sample of the random variable $D_{T,R}$.

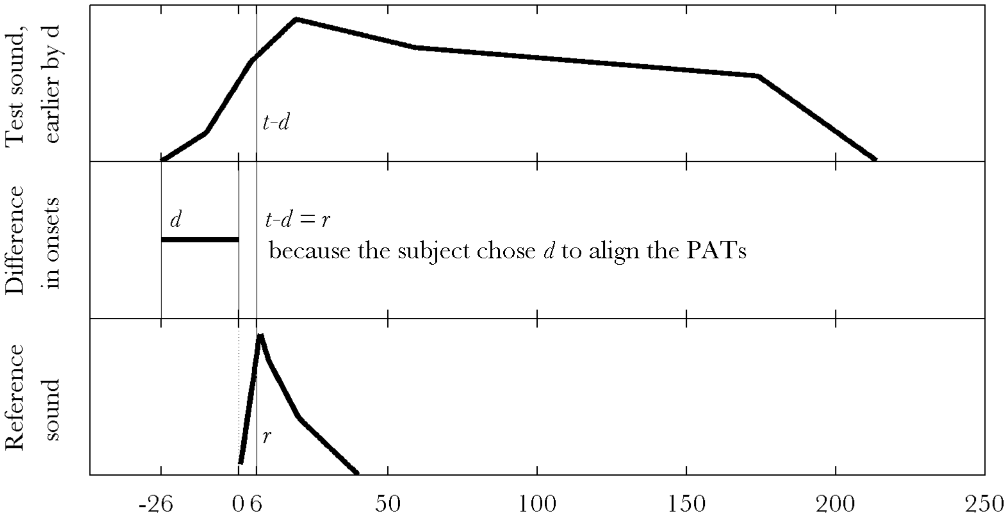


***Figure 15: The situation after a subject has aligned the PATs of the test and reference sounds (using the same hypothetical amplitude envelopes as Figure 14). Now the time difference between the physical onsets is our measurand* d.**

Suppose we perform multiple trials with the same test and reference sounds. This gives us many samples of the random variable $D_{T,R}$, which we can use to infer the underlying distribution for $D_{T,R}$, as shown in Figure 16. Unlike the previous made-up illustrations, Figure 16 displays real measured data from the listening experiment described in Chapter 4, namely the results of all 33 trials performed with "Clarinet" as the test sound and "Clarinet SMC12" as the reference sound. Each data point is the value of *d* that a subject found in one trial. The X axis is the delay (in ms)

[71] In fact the timing of the reference sound is fixed, so perhaps it would be more accurate to call *d* the amount of time by which the physical onset of the test sound precedes that of the reference sound. I chose this sign convention for *d* so that a plot of *D* will be exactly a plot of the test sound's PAT-pdf if we assume the reference sound's PAT has mean and variance of zero.

[72] This step relies on the common sense assumption that delaying the physical onset of a sound will delay its PAT by the same amount. This is just another way to state the assumption inherent in the "each event has an intrinsic PAT" model that an event's PAT has a fixed temporal relationship to its physical onset.

[73] In the isochrony method the situation is the same except for the addition of a time constant equal to half the period of repetition of the reference sound.

between the physical onset of the reference sound and the physical onset of the test sound; these are the units of each value of *d*. Negative values of X represent situations in which the reference sound's onset is *before* the test sound's onset; positive X means the reference sound's onset is *after* the test sound's onset. (In this case every trial resulted in a negative value, in other words, every listener felt the Clarinet's physical onset had to precede Clarinet SMC12's physical onset to achieve perceptual synchrony.) The estimated probability distribution gives the relative probability of perceiving the two sounds as synchronous as a function of the delay time between their physical onsets. So to interpret the estimated shape of the probability density function in Figure 16, we'd say that when the reference sound precedes the test sound by about 30-50 ms, most listeners will hear them as synchronous. Note that there are two data points on the right edge of the graph, near the zero point of physical synchrony; I believe these are due to a particular fusion effect discussed in Section 4.4.2 (page 118).

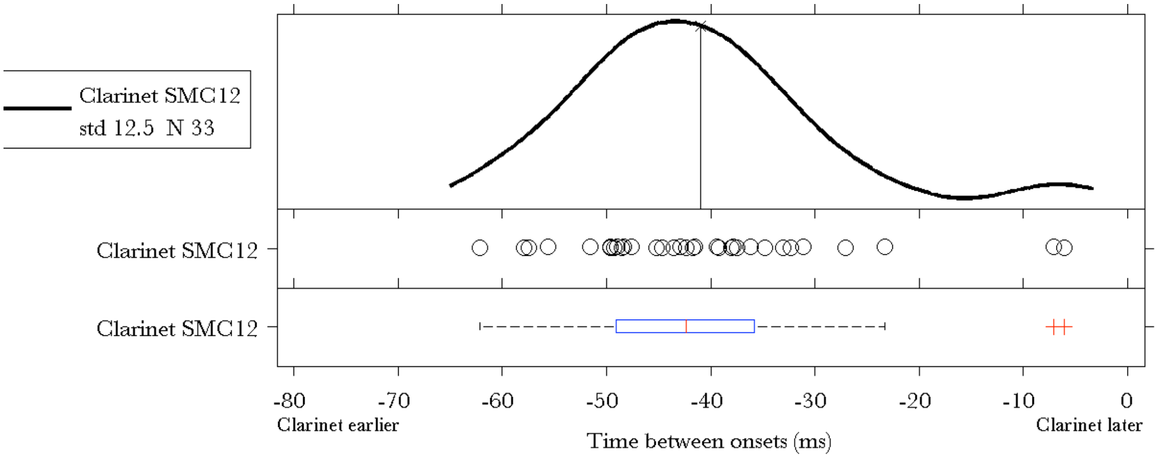


Figure 16: Example results for multiple trials using the same test and reference sounds (*Violin *and* Violin SMC23 *respectively).

The middle plot is a one-dimensional scatter plot of all responses (with slight jitter added to the Y axis for visual clarity). On the bottom is a standard box plot of the same data. The top plot is one estimate of the underlying probability distribution using nonparametric kernel density estimation: a Gaussian distribution of fixed variance is placed around each data point and these are all summed together. The vertical line in the top plot shows the mean.

If we could assume that the reference sound's PAT is exactly zero (i.e., that the mean and variance of its PAT-pdf are both zero), then we would interpret the top plot of Figure 16 as the intrinsic PAT-pdf of the test sound.[74] But if we assume that the reference sound itself has an unknown intrinsic PAT-pdf, then Figure 16 is the shape of the pdf for $D_{T,R}$ and we cannot directly measure the PAT-pdf of the test or reference sounds. Informally, if the test and reference sounds' PATs are both ambiguous, then what we measure will be doubly ambiguous.

3.4.4.1 Statistical Independence of T and R

If T and R are independent random variables, then the pdf of their difference $D_{T,R} = T - R$ is the cross-correlation[75] of the pdfs for T and R and the variances add: $\mathrm{var}(D_{T,R}) = \mathrm{var}(T) + \mathrm{var}(R)$.

There might be cases where these two random variables are not independent:

> I think an interesting case to consider is a sound whose attack, on close inspection, looks like a 'click-whoosh.' That is, there is an impulsive attack, followed closely by a more spread out portion. The level of the impulse can be reduced until it and the following "whoosh" are equally likely to be heard as the attack. If the listener attends to the impulse, the distribution is tight; if the impulse is disregarded and the "whoosh" is considered the "main event" of the attack, then the distribution is broader. The result is a bimodal, nonsymmetric distribution.
>
> Now consider two such sounds played in alternation to form a regular beat - we can now consider how the uncertainties should be combined for this case. In thinking about this, it occurs to me that an important effect is 'conditional listening.' That is, however you decide to listen to sound A generally affects how you listen to sound B. For example, if you attend to the impulse at the beginning of sound A, you are more likely to attend to a leading impulse in sound B, and so on.
>
> (Julius Smith, personal communication, 10 August 2007).

If T and R are not independent then

$$\mathrm{var}\,(T\text{-}R) = \mathrm{var}(T) + \mathrm{var}(R) - 2\ \mathrm{covariance}(T,R)$$

For simplicity we will assume that these are independent random variables.

[74] Again, this is why I chose the sign convention that D=T-R rather than the equally reasonable D=R-T.

[75] Cross-correlation is equivalent to convolution with a left/right flip of one input. I will follow the notational convention that $a\star b$ means "the cross-correlation of a with b." (http://ccrma.stanford.edu/~jos/mdft/Cross_Correlation.html)

3.4.4.2 Penalty Term for Each Pair of Sounds

Some sound pairs are inherently more difficult to align with each other due to auditory streaming effects that Section 3.6 (page 62) will discuss at length. Rather than modeling this effect with covariance, we will assume that the random variables are independent and instead add an extra "penalty" for each pair of sounds that adds additional uncertainty (i.e., variance). We'll represent this with an extra noise term:

$$D_{T,R} = (T\text{-}R) + Penalty_{T,R}.$$[76]

We'll use an abbreviation to notate the variance of these penalties:

$$\sigma^2_{T,R} \triangleq \mathrm{var}(Penalty_{T,R})$$

We make the following assumptions about this penalty term:

- The penalty is a random variable drawn from a zero-mean distribution. (There should be no reason that difficulty in hearing two sound's relative time alignment would lead to results that are biased in either direction.)
- $Penalty_{A,A} \triangleq 0$ for any A. This is because two copies of the same sound should always be in the same auditory stream and hence there should be no added difficulty in perceiving their relative time alignment. (See Section 3.6.)
- $Penalty_{A,B} = Penalty_{B,A}$. In other words, the extra difficulty in aligning any given pair of sounds should not depend on which is the test and which is the reference.

Even with these assumptions the penalty term gives our model too many degrees of freedom, because we can explain the observed variance of any $D_{T,R}$ with any combination of variance in T and R and variance in $Penalty_{T,R}$. For example, one extreme would be to assume that every intrinsic PAT has zero variance (i.e., that any sound's PAT is a single moment in time, not a probability distribution) and to attribute all of the observed variances to the penalty terms. Our model doesn't quite allow this since $Penalty_{A,A} \triangleq 0$, but as long as each I_A has enough variance to account for the variance in $D_{A,A}$ then we can account for the variance in any $D_{A,B}$ with $Penalty_{A,B}$.

In other words, if trials involving sound A always result in a wide range of results, is it because sound A has a broad range of acceptable PAT times, or because sound A happens to be difficult to align against every other sound?

[76] We will sometimes use $\varepsilon_{T,R}$ as a synonym for $Penalty_{T,R}$.

3.5 Getting from PAT Measurements to Intrinsic PATs

We are able to sample a random variable $D_{T,R}$ equal to the difference between the random variables for the intrinsic PAT-pdfs of the test and reference sounds (T and R respectively), plus a zero-mean noise term $Penalty_{T,R}$ representing the additional difficulty in aligning the given pair of sounds:

$$D_{T,R} \triangleq T - R + Penalty_{T,R}$$
$$mean(Penalty_{T,R}) = 0$$

We want to know T and R, the intrinsic PAT-pdfs of each sound, and would like a way to infer them from the data we are able to measure. There are many statistical techniques for estimating the distribution of the pdf of $D_{T,R}$ from the observed data. Unfortunately the characteristics of the pdf of $D_{T,R}$ tell us very little about T or R. For example, knowing that $D_{T,R}$ has a mean of about -41 ms (as in Figure 16) could mean that T has a mean of 20 ms and R has a mean of 61 ms, or that T has a mean of 120 ms and R has a mean of 161 ms. Even knowing that $D_{T,R}$ fits a Gaussian distribution is no guarantee that T and R are Gaussian.[77]

By assuming that T and R are independent, we can use these elementary properties of differences of independent random variables:

If $D_{T,R} = T\text{-}R$, then

$mean(D_{T,R}) = mean(T)\text{-}mean(R)$

$var(D_{T,R}) = var(T)+var(R)$

If $D_{T,R} = (T\text{-}R)+Penalty_{T,R}$ and $mean(Penalty_{T,R}) = 0$, then

$mean(D_{T,R}) = mean(T)\text{-}mean(R)$

$var(D_{T,R}) = var(T) + var(R) + var(Penalty_{T,R})$

3.5.1 Comparing Multiple Pairs Drawn from the Same Set of Sounds

The only way to get more information about the shapes of each sound's intrinsic PAT-pdf distributions is to perform experimental trials with multiple pairs of sounds drawn from the same set.

[77] In fact, the central limit theorem tells us that as we add together more probability distributions the result will eventually become Gaussian even if the addends are not, so in general D will be "more" Gaussian than T or R.

One prediction of this model so far, without making any assumption about shapes of distributions or statistical independence, is that

$$mean(D_{A,C}) = \mu_A - \mu_C = \mu_A - \mu_C + (\mu_B - \mu_B) = (\mu_A - \mu_B) + (\mu_B - \mu_C) = mean(D_{A,B}) + mean(D_{B,C})$$

In other words, if we have any trio of sounds that were all compared against each other, the means should "add up" as we'd intuitively expect. (This is the result of assuming that $Penalty_{T,R}$ has a mean of zero.) Section 4.3.10 (page 103) checks this prediction for the results of the listening experiment described in Chapter 4.

Unfortunately it is often not practical to carry out full-factorial experimental design comparing every possible pair of sounds drawn from a given set, especially if we need dozens of trials with each pair A and B to produce a robust estimate of the distribution of $D_{A,B}$. Suppose there are n sounds in total, and let us define $N_{A,B} \geq 0$ as the number of trials that used sound A as the test and sound B as the reference. For full generality we must consider that $N_{A,B}$ may be zero or too small to allow us to estimate the shape of $D_{A,B}$ for any given A and B, so $D_{A,B}$ will be unknown for some pairs of sounds.

3.5.2 A Weak Upper Bound On Intrinsic Variance

Since variance is always non-negative, we can put a weak upper bound on each sound's intrinsic variance:

$$\begin{aligned}
&\mathrm{var}(D_{A,B}) = \mathrm{var}(I_A) + \mathrm{var}(I_B) + \mathrm{var}(Penalty_{T,R}) \\
&\mathrm{var}(D_{A,B}) \geq \mathrm{var}(I_A) \\
&\min_B \left(\mathrm{var}(D_{A,B})\right) \geq \mathrm{var}(I_A)
\end{aligned}$$

3.5.3 What We Can Learn from Trials Using the Same Sound as Test and Reference

Because S and S are the same sound, the mean of $D_{S,S}$ should be zero, and tells us nothing about the mean of I_S.

The distribution of D should be symmetric when the test and reference sounds are the same sound. This fits common sense: for any given delay time between the two copies of the sound that makes the result sound synchronous or isochronous (depending on the task), the opposite of that delay time will produce an acoustically identical stimulus.

We can estimate of the variance of each sound's intrinsic PAT-pdf by looking at the sample variance of all trials aligning that sound against a second copy of itself:

$var(D_{S,S}) = 2var(I_S) + var(Penalty_{S,S})$

$var(I_S) = (var(D_{S,S}) - var(Penalty_{S,S}))/2$

The penalty term when aligning two copies of the same sound should be very small, so

$var(I_S) \approx var(D_{S,S}) / 2$

One way to test this assumption is to check whether

$$\mathrm{var}(D_{A,B}) \geq \left(\mathrm{var}(D_{A,A}) + \mathrm{var}(D_{B,B})\right)\Big/2$$

holds for all A and B. (See Table 13 on page 101.)

We can also use this assumption to estimate[78] the variance of $Penalty_{A,B}$ for any distinct pair of sounds A and B:

$$\hat{\sigma}^2_{A,B} = \mathrm{var}(D_{A,B}) - \frac{\mathrm{var}(D_{A,A}) + \mathrm{var}(D_{B,B})}{2}$$

3.5.4 Normal model

If we assume that all intrinsic PAT-pdf distributions are normal (i.e., Gaussian) then the situation becomes more tractable. For one thing, a Gaussian distribution is completely characterized by its mean and variance, so we can ignore all other details of our measurements. Also, the difference of two Gaussian random variables is another Gaussian random variable, and we have an analytic solution relating the means and variances:

If

$R \sim N(\mu_R, \sigma_R^2)$ R is normal with mean μ_R and variance σ_R^2

$T \sim N(\mu_T, \sigma_T^2)$ T is normal with mean μ_T and variance σ_T^2

$Penalty_{T,R} \sim N(0, \sigma_{T,R}^2)$ $Penalty_{T,R}$ is normal with zero mean and variance σ_P^2

$D_{T,R} = T - R + Penalty_{T,R}$

[78] We'll follow the convention of notating our estimates of statistical parameters with "hats", so that $\hat{\sigma}^2_X$ represents our estimate of σ^2_X .

T, *R*, and *Penalty*$_{T,R}$ are independent

then

$D_{T,R} \sim \mathcal{N}(\mu_D=\mu_T-\mu_R,\ \sigma_D^2=\sigma_T^2+\sigma_R^2+\sigma_P^2)$

So for a set of *n* sounds, our model has $n(n+3)/2$ parameters: the $2n$ intrinsic means and variances $\sigma_1^2,\sigma_2^2,...,\sigma_n^2,\mu_1,\mu_2,...,\mu_n$ and the $n(n-1)/2$ alignment penalty variances $\sigma_{2,1}^2,\sigma_{3,1}^2,\sigma_{3,2,}^2,...,\sigma_{n,1,}^2...,\sigma_{n,n-1}^2$. (Remember that $\sigma_{i,i}^2=0$ and $\sigma_{i,j}^2=\sigma_{j,i}^2$.)

There are infinitely many choices for these parameters that yield the same results, because we can shift all of the intrinsic means by any constant and still end up with the same results for all $D_{A,B}$. We somewhat arbitrarily choose this constant so that *min(mean(I*$_A$*))=0.*

3.5.4.1 <u>Finding PAT-pdf by Comparison to a Sound with Known PAT-pdf</u>

If we already have estimates $\hat{\mu}_R$ and $\hat{\sigma}_R^2$ of the mean and variance of the intrinsic PAT-pdf of a reference sound *R*, and we have the sample mean and variance μ_D and σ_D^2 from a series of trials comparing test sound *T* to reference sound *R*, then we can estimate the parameters of *T* with

$$\hat{\mu}_T=\hat{\mu}_R+\mu_D$$
$$\hat{\sigma}_T^2=\sigma_D^2-\hat{\sigma}_R^2-\hat{\sigma}_{P_{T,R}}^2$$

Note that this formula allows $\hat{\sigma}_T^2$ to be negative if $\sigma_D^2<\hat{\sigma}_R^2$, which is impossible because variance must always be nonnegative. If this occurs than our estimate $\hat{\sigma}_R^2$ is clearly wrong and must be revised so that $\hat{\sigma}_R^2\le\sigma_D^2$.

It also might be the case that we already have estimates $\hat{\mu}_A$, $\hat{\sigma}_A^2$, $\hat{\mu}_B$, and $\hat{\sigma}_B^2$ of the intrinsic means and variances for two reference sounds *A* and *B*, both of which were compared to a third sound *C*. In general we can expect that

$$\hat{\mu}_A+\mu_{D_{A,C}}\neq\hat{\mu}_B+\mu_{D_{B,C}}$$

In other words, the two reference sounds *A* and *B* might provide us with conflicting estimates of the intrinsic mean of *C*.

3.5.4.2 An Optimal Least-Squares Solution For All Relative PAT Times

Marcus formulated the problem of converting a matrix of pairwise PAT alignment results into a single (relative) intrinsic PAT value for each individual sound in terms of a least squared error optimization (Marcus 1981, 248).[79] His desired answer was a single scalar value p_s for each sound s, with the understanding that the true intrinsic PAT for each sound s was p_s+c, with c forever unknown but constant for all sounds. (For now we will put aside the idea of PAT-pdf and concentrate on the problem of finding intrinsic PAT, thinking of PAT temporarily as a single instant.)

His assumption was that the result of each trial was a time offset amount equal to the difference between the intrinsic PATs of the two sounds, plus a noise term that he assumed to be Gaussian with zero mean and a measurable variance attributed to subject's inconsistency "in reproducing his own chosen set of offsets" (Marcus 1981, 248). In my model the variance is the sum of the intrinsic variances of the two sounds plus the penalty term. Since the sum of Gaussian random variables is itself a Gaussian random variable, we can apply his method without committing to an interpretation of the source of the variance. Likewise, we can interpret the output of his method as the mean of each sound's intrinsic PAT-pdf, without necessarily subscribing to the notion that a sound's PAT is a single discrete instant.

In his experiment, each subject aligned each possible pair of (nine) sounds exactly once (not including each sound against itself) with the isochrony method, so his expression for the total error has one term for each of $n(n-1)$ pairs of different sounds. I will extend this method to the case where there might be a different number of trials for each pair of sounds, possibly zero, with each trial weighted equally in the result.[80] Let n be the number of sounds. Let $\mathcal{N}_{i,j}$ be the number of trials aligning sounds i and j, regardless of which was test and which was reference, so that $\mathcal{N}_{i,j}=\mathcal{N}_{j,i}$. This algorithm does not use any results from trials comparing a sound with itself, so for convenience let $\mathcal{N}_{i,i} \triangleq 0$. Let each $d_{i,j}(k)$ be the result of one trial aligning test sound i against reference sound j, or the opposite of the result of one trial aligning test sound j against reference

[79] I have taken the liberty of translating Marcus' ideas into my own terminology, using "intrinsic PAT" where he used "p-center," and renaming his variables for consistency.

[80] This could be considered a bias towards sounds and pairs of sounds for which more trials were performed. We could instead weight each trial by the reciprocal of the number of trials of that type; this would weight each pair of sounds equally by introducing a different kind of bias, making each trial count more for pairs with few trials and making each trial count less for pairs with large numbers of trials. One could also trade off these two kinds of bias by weighting each trial comparing sounds i and j by $(\mathcal{N}_{i,j})^x$ where $-1 \le x \le 0$.

sound *i*, with $1 \le k \le \mathcal{N}_{i,j}$. The order of the trials is unimportant.[81] For convenience let $d_{i,j}(k)$=-$d_{j,i}(k)$.

We will find *p*, a (column) vector containing the relative intrinsic PAT for each sound, by minimizing the following sum-of-squared-error cost function:

$$\mathcal{J}(p) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{i-1}\sum_{k=1}^{\mathcal{N}_{i,j}}\left(d_{i,j}(k) - p_i + p_j\right)^2$$

We find the partial derivative of $\mathcal{J}$ with respect to each p_s:

$$\begin{aligned}
&\frac{\partial}{\partial p_s}\mathcal{J}(p) = \frac{\partial}{\partial p_s}\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{i-1}\sum_{k=1}^{\mathcal{N}_{i,j}}\left(d_{i,j}(k) - p_i + p_j\right)^2 \\
&= \frac{1}{2}\frac{\partial}{\partial p_s}\left(\sum_{j=1}^{s-1}\sum_{k=1}^{\mathcal{N}_{s,j}}\left(d_{s,j}(k) - p_s + p_j\right)^2 + \sum_{i=s+1}^{n}\sum_{k=1}^{\mathcal{N}_{i,s}}\left(d_{i,s}(k) - p_i + p_s\right)^2\right) \\
&= \frac{1}{2}\frac{\partial}{\partial p_s}\left(\sum_{i=1}^{s-1}\sum_{k=1}^{\mathcal{N}_{i,s}}\left(-d_{i,s}(k) - p_s + p_i\right)^2 + \sum_{i=s+1}^{n}\sum_{k=1}^{\mathcal{N}_{i,s}}\left(d_{i,s}(k) - p_i + p_s\right)^2\right) \\
&= \frac{1}{2}\frac{\partial}{\partial p_s}\left(\sum_{i=1}^{s-1}\sum_{k=1}^{\mathcal{N}_{i,s}}\left(p_s + \left(d_{i,s}(k) - p_i\right)\right)^2 + \sum_{i=s+1}^{n}\sum_{k=1}^{\mathcal{N}_{i,s}}\left(p_s + \left(d_{i,s}(k) - p_i\right)\right)^2\right) \\
&= \frac{1}{2}\frac{\partial}{\partial p_s}\sum_{i=1}^{n}\sum_{k=1}^{\mathcal{N}_{i,s}}\left(p_s + \left(d_{i,s}(k) - p_i\right)\right)^2 \\
&= \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{\mathcal{N}_{i,s}}\frac{\partial}{\partial p_s}\left(p_s + \left(d_{i,s}(k) - p_i\right)\right)^2 \\
&= \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{\mathcal{N}_{i,s}}\frac{\partial}{\partial p_s}\left(p_s^{\,2} + 2p_s\left(d_{i,s}(k) - p_i\right) + \left(d_{i,s}(k) - p_i\right)^2\right) \\
&= \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{\mathcal{N}_{i,s}}\left(2p_s + 2\left(d_{i,s}(k) - p_i\right)\right) \\
&= \sum_{i=1}^{n}\sum_{k=1}^{\mathcal{N}_{i,s}}\left(p_s + d_{i,s}(k) - p_i\right) \\
&= \left(p_s\sum_{i=1}^{n}\mathcal{N}_{i,s}\right) - \left(\sum_{i=1}^{n}\mathcal{N}_{i,s}p_i\right) + \left(\sum_{i=1}^{n}\sum_{k=1}^{\mathcal{N}_{i,s}}d_{i,s}(k)\right) \\
&= -\mathcal{N}_{1,s}p_1 - \mathcal{N}_{2,s}p_2 - \ldots + \left(\sum_{i=1}^{n}\mathcal{N}_{i,s}\right)p_s - \ldots - \mathcal{N}_{n,s}p_n + \left(\sum_{i=1}^{n}\sum_{k=1}^{\mathcal{N}_{i,s}}d_{i,s}(k)\right)
\end{aligned}$$

[81] In other words, we treat the set of all trials involving sounds *i* and *j* as a set, not a sequence; the index *k* in $d_{i,j}(k)$ will only appear in summations over all *k*.

To get from the first to the second line we used the fact that each term in the partial derivative will be zero unless *i*=*s* (in which case $1 \le j \le i-1 = s-1$) or *j*=*s* (in which case *i*>*j*=*s*). From the second to the third we renamed *j* to *i* and used $d_{i,j}(k)=-d_{j,i}(k)$ and $\mathcal{N}_{i,j}=\mathcal{N}_{j,i}$. From the fourth to the fifth note that combining the two summations added an extra term for *i*=*s*, but that this is zero since $\mathcal{N}_{s,s}=0$.

Setting $\frac{\partial}{\partial p_s}\mathcal{J}(p)=0$ for each *s* produces a system of *n* linear equations in *n* unknown values of *p*:

$$\begin{pmatrix} \sum_{i=1}^{n}\mathcal{N}_{i,1} & -\mathcal{N}_{2,1} & \cdots & -\mathcal{N}_{n,1} \\ -\mathcal{N}_{1,2} & \sum_{i=1}^{n}\mathcal{N}_{i,2} & \cdots & -\mathcal{N}_{n,2} \\ \vdots & \vdots & \ddots & \vdots \\ -\mathcal{N}_{1,n} & -\mathcal{N}_{2,n} & \cdots & \sum_{i=1}^{n}\mathcal{N}_{i,n} \end{pmatrix} \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_n \end{pmatrix} = \begin{pmatrix} -\sum_{i=1}^{n}\sum_{k=1}^{\mathcal{N}_{i,1}} d_{i,1}(k) \\ -\sum_{i=1}^{n}\sum_{k=1}^{\mathcal{N}_{i,2}} d_{i,2}(k) \\ \vdots \\ -\sum_{i=1}^{n}\sum_{k=1}^{\mathcal{N}_{i,n}} d_{i,n}(k) \end{pmatrix}$$

An arbitrary constant can be added to all p_i and still equally explain the observed experimental results, so the matrix on the left will not be invertible. We can get around this problem by instead computing the pseudo-inverse, or by arbitrarily removing one of the columns and removing the corresponding p_i from the middle column vector.

Marcus chose this constant "such that $\sum p_i = 0$" (Marcus 1981, 248), not attempting to relate PAT to absolute time but only finding relative PAT values for all sounds. Based on my assumption that PAT should never precede physical onset, I instead choose this constant such that $\min(p_i)=0$, so that each resulting p_i is a lower bound on the absolute PAT of sound *i*.

Once we have these estimated intrinsic means, it would be trivial to set all estimated intrinsic variances to zero and every $\text{var}(\mathit{Penalty}_{A,B}) = \text{var}(D_{A,B})$, but of course this tells us nothing about the presumably interesting shapes of each sound's PAT-pdf distributions. We should therefore choose a better estimate for the intrinsic variances.

3.5.4.3 An Optimal Maximum Likelihood Solution For the Entire Model?

Our full model says that for each pair of sounds *T* and *R* we are able to observe samples of a random variable $D_{T,R}$ such that

$$D_{T,R} = I_T - I_R + \varepsilon_{T,R}$$

I was unable to obtain a closed-form maximum likelihood solution analytically for all of the parameters of this model (namely, mean and variance of each sound's intrinsic PAT-pdf, plus variance for the penalty term for each pair of sounds) in terms of a set of observed experimental outcomes; Appendix B (page 170) presents the details. However, if we assume that all the variances are known constants, then we can find a maximum likelihood solution for the means as shown in Section B.1.1 (page 173).

3.5.5 Five Complete Algorithms to Estimate all Normal Model Parameters

With all of the above caveats, warnings, and partial solutions in mind, here are five complementary approaches to estimating all of the intrinsic means μ_i, intrinsic variances σ_i^2, and penalty term variances $\sigma_{i,j}^2$ given only the observed samples of $D_{T,R}$ for an arbitrary subset of n sounds $(1 \le i \le n, 1 \le j \le n)$ and assuming that everything is Gaussian.

Note that the result of any of these algorithms is a statistical model, and we can compute the total likelihood of any such model given all our experimental results as derived in Appendix B:

$$\sum_{i=1}^{n}\sum_{j=1}^{i}\left(\mathcal{N}_{i,j}\log(\frac{1}{\sqrt{2\pi}}) - \mathcal{N}_{i,j}\frac{1}{2}\log\left(\sigma_i^2+\sigma_j^2+\sigma_{i,j}^2\right) - \frac{1}{2\left(\sigma_i^2+\sigma_j^2+\sigma_{i,j}^2\right)}\sum_{k=1}^{\mathcal{N}_{i,j}}\left(d_{i,j}(k)-\mu_i+\mu_j\right)^2 \right)$$

Many of these algorithms work by searching a space of possible model parameters, computing the likelihood of each candidate, and outputting the result with the highest likelihood.

Note also that the variances always appear only as part of the sum $\sigma_i^2+\sigma_j^2+\sigma_{i,j}^2$, so for any given set of variances we can always choose any *i*, select any x $0 \le x < \sigma_i^2$, set $\sigma_i^2{}' = \sigma_i^2 - x$, then set all $\sigma_{i,j}^2{}' = \sigma_{i,j}^2 + x$ to produce another set of parameters with exactly the same likelihood for any possible experimental results. Only the restrictions $\sigma_{i,i}^2 \triangleq 0$ and $\sigma_{i,j}^2 \triangleq \sigma_{i,j}^2, i \ne j$ prevent this.

Another consequence of this relationship among the variances is that, given the observed sample variance for each $D_{T,R}$ and a set of estimated intrinsic variances σ_i^2, we can almost perfectly account for all the observed sample variance by setting $\hat{\sigma}_{i,j}^2 = \mathrm{var}(D_{i,j}) - \hat{\sigma}_i^2 - \hat{\sigma}_j^2$.

3.5.5.1 Shortest-Variance Path from a Chosen Reference

Let us represent the observed results of a PAT alignment experiment as a graph. The nodes are the set of *n* sounds and the edges are the comparisons between sounds, with an edge between sounds *A* and *B* if and only if we have (enough) experimental results from trials aligning *A* and *B*. The distance between nodes *A* and *B* is $\operatorname{var}(D_{A,B})$.

First of all, this graph must be connected in order for *any* algorithm to be able to estimate all of the statistical parameters; otherwise there is no common reference between the disconnected subgraphs.

Suppose we have selected one sound as our reference, and somehow initialized the intrinsic mean and variance for this sound. We apply Dijkstra's well-known shortest path algorithm (Dijkstra 1959) to produce the shortest path tree rooted on our chosen node. Now we proceed from the source node through the tree. At each step we know the intrinsic mean and variance of the parent, so we can use the observed sample mean and variance from trials comparing the parent and child to estimate the intrinsic parameters for the child:

$$\hat{\mu}_{child} = \hat{\mu}_{parent} + \mu_{D_{child,parent}}$$
$$\hat{\sigma}^2_{child} = \sigma^2_{D_{child,parent}} - \hat{\sigma}^2_{parent}$$
$$\hat{\sigma}^2_{P_{T,R}} = 0$$

If at any step in this process $\hat{\sigma}^2_{child} < 0$, then we have a problem. Let $x = -\hat{\sigma}^2_{child}$. We can arbitrarily pick some small minimum variance v_{min} such as 0.1 ms^2 and set $\hat{\sigma}^2_{child} = v_{\min}$. Then we must subtract $x+v_{min}$ from $\hat{\sigma}^2_{parent}$, add $x+v_{min}$ to $\sigma^2_{D_{parent,grandparent}}$ (unless *parent* is the root of the tree), and recompute the variances of any other children of *parent*.

How do we select the reference and its intrinsic mean and variance? One option is to perform a grid search over all possible reference sounds and over a range of plausible intrinsic variances, run the algorithm on each option, and choose the parameters with the highest likelihood. There is no need to search over a range of intrinsic means for the reference because we can always add the same constant to all intrinsic means without changing the likelihood of the model.

Another option would be to select the reference sound and its intrinsic variance (and possibly also mean) based on some *a priori* knowledge. For example, if one of the sounds is the ideal digital impulse (see page 75) then we may have reason to use that as the reference. We might also choose the reference's intrinsic variance according to $\hat{\sigma}^2_{reference} = \sigma^2_{D_{reference,reference}} / 2$.

One motivation for this algorithm is that in general, the derivation of I_X that results in the lowest variance gives us the most information about I_X. Specfically,

a. If var($D_{A,B}$) < var($D_{A,C}$), then B is the better reference sound for finding I_A because $var(Penalty_{A,B})+var(I_B) < var(Penalty_{A,C})+var(I_C)$.

b. If var($D_{A,C}$) > var($D_{A,B}$) + var($D_{B,C}$) then $D_{A,B}$ and $D_{B,C}$ together tell us more about A's intrinsic PAT-pdf than $D_{A,C}$:

   $var(I_A) + var(I_C) + var(Penalty_{A,C}) > var(I_A) + var(I_B) + var(Penalty_{A,B}) + var(I_B) + var(I_C) + var(Penalty_{B,C})$

   $var(Penalty_{A,C}) > + 2var(I_B) + var(Penalty_{A,B}) + var(Penalty_{B,C})$

   In other words, aligning A with C is so difficult that we get a more accurate understanding of the relationship between A and C by aligning A with B and then aligning B with C.

3.5.5.2 <u>Greedy Search for the Next Sound with the Smallest Intrinsic Variance.</u>

This method is a slight variant of the shortest-path method just described, motivated by an attempt to avoid the backtracking step.

Again we start by selecting one sound as our reference (or by trying every possible sound as a reference with grid search) and perform a greedy algorithm that estimates the intrinsic mean and variance for one new sound on each iteration. This time, however, we choose the next sound that has the lowest variance. Here is pseudo-code for the algorithm:

```
until every sound has an estimated intrinsic mean and variance
  for each sound s without estimates for $\hat{\mu}_s$ and $\hat{\sigma}^2_s$
    lowest_intrinsic_variance(s) = Infinity
      parent(s) = NULL
      for each sound r with an estimate for $\hat{\mu}_s$ and $\hat{\sigma}^2_s$
      $guessed\_\hat{\sigma}^2_s = \max(\sigma^2_{\min}, \sigma^2_{D_{s,r}} - \hat{\sigma}^2_r)$
      if $guessed\_\hat{\sigma}^2_s <$ lowest_intrinsic_variance(s)
```

$$\mathit{lowest\_intrinsic\_variance}(\mathrm{s}) = \mathit{guessed}\_\hat{\sigma}^2_s$$

$$\mathit{parent}(\mathrm{s}) = \mathrm{r}$$

next_estimate = s with the lowest *lowest_intrinsic_variance*(s)

$$\hat{\sigma}^2_{next\_estimate} = \mathit{lowest\_intrinsic\_variance}(\mathrm{s})$$

$$\hat{\mu}_{next\_estimate} = mean(D_{next\_estimate, parent(next\_estimate)}) + \hat{\mu}_{parent(next\_estimate)}$$

3.5.5.3 Batch Estimation from Known Estimates

Here is one final variation on the theme of successively deriving all intrinsic means and variances from a given reference sound *R* whose intrinsic variance is taken as given. On each iteration of the algorithm we will estimate all possible intrinsic means and variances based on the estimates we have so far. We choose a minimum intrinsic variance v_{min} to be a very low number such as 0.01 ms^2.

On the first step we consider every sound *S* for which $N_{S,R}$ is sufficient to estimate $var(D_{S,R})$:

$$\hat{\mu}_S = \hat{\mu}_R + \mu_{D_{S,R}}$$
$$\hat{\sigma}^2_S = \max(v_{\min}, \sigma^2_{D_{S,R}} - \hat{\sigma}^2_R)$$

If we have (enough) trials comparing every sound to *R* then we're done. Otherwise we iterate the following until we have estimated all of the intrinsic means and variances:

For each sound *T* for which there is at least one sound *S* such that $N_{T,S}$ is sufficient to estimate $var(D_{T,S})$ and we have already estimated $\hat{\mu}_S$ and $\hat{\sigma}^2_S$, estimate *T*'s parameters from taking the mean over all sounds *S*:

$$\hat{\mu}_T = mean(\hat{\mu}_{S1} + \mu_{D_{T,S1}}, \hat{\mu}_{S2} + \mu_{D_{T,S2}}, \ldots)$$
$$\hat{\sigma}^2_T = \max(v_{\min}, mean(\sigma^2_{D_{T,S1}} - \hat{\sigma}^2_{S1}, \sigma^2_{D_{T,S2}} - \hat{\sigma}^2_{S2}, \ldots)$$

One variant would be to replace the mean with a weighted average weighted by each $N_{T,S}$.

The advantage of this algorithm is that the source of the estimates for each sound is either the chosen reference *R* or an equal weighting of all the sounds that are "as close as possible" to *R* according to the interpretation of the experimental results as a graph. A potential problem is that each estimate might be derived from a different number of reference sounds, which could be a source of inconsistency.

3.5.5.4 Variances from Trials Against Self, Maximum Likelihood Means

This solution starts by finding the intrinsic variance for each sound from trials using two copies of that sound, according to

$$var(I_S) \triangleq var(D_{S,S}) \ / \ 2$$

For any sound S for which $N_{S,S}$ is too small to estimate $var(D_{S,S})$, we need an alternate method of estimating intrinsic variance. Given the intrinsic variances that we were able to estimate, we can find an upper bound on the remaining variances as follows:

$$\begin{aligned}
&\mathrm{var}(D_{A,B}) = \mathrm{var}(I_A) + \mathrm{var}(I_B) + \mathrm{var}(Penalty_{T,R}) \\
&\mathrm{var}(D_{A,B}) - \mathrm{var}(I_A) - \mathrm{var}(I_B)) \geq 0 \\
&\mathrm{var}(D_{A,B}) - \mathrm{var}(I_B)) \geq \mathrm{var}(I_A) \\
&\min_B \left(\mathrm{var}(D_{A,B}) - \mathrm{var}(I_B)\right) \geq \mathrm{var}(I_A)
\end{aligned}$$

Since $var(D_{S,S}) \ / \ 2$ is likely to be somewhat of an underestimate for $var(I_S)$ due to comb filtering effects, for consistency we should choose a similarly low estimate for $var(I_X)$. We can arbitrarily pick some α ($0<\alpha<1$) and multiply each intrinsic variance's upper bound by α to determine the estimate.

Once we have estimated the intrinsic variance for each sound, we can estimate the variance of all the penalty terms with

$$var(D_{T,R}) = var(T) + var(R) + var(Penalty_{T,R})$$

$$\max(0, var(D_{T,R}) - var(T) - var(R)) = var(Penalty_{T,R})$$

In other words, we choose the penalty for any pair of sounds T and R to account exactly for all of the observed variance in $D_{T,R}$ that our estimated intrinsic variances do not explain.

Now that we have estimates for all of the variances of our model, we can use the maximum likelihood solution in Appendix B to solve for the intrinsic means. The results are determined except for an arbitrary constant c that can be added to all means to produce the same result; we choose c such that $\min(\mu)+c>=0$, and if we have extremely impulsive sounds, so that $\min(\mu)+c=0$.

3.5.5.5 Least Squares Means, Variances from Trials Against Self

A variant of the previous algorithm is to compute all the means according to "An Optimal Least-Squares Solution For All Relative PAT Times" (Section 3.5.4.2, page 53), and then independently find the variances as in the previous section. The only difference is that the least squares solution for the means does not depend on variance while the maximum likelihood solution uses the estimated variances essentially to down-weight results from pairs of sounds that were more difficult to align, as Section B.1.1 (page 173) suggests.

3.6 Choice of a Reference Sound

How then should one select the reference sound? One criterion is that the reference sound's PAT should be maximally clear and unambiguous; in other words, its probability distribution should be very tall and narrow. Consider the opposite extreme: suppose the reference sound were white noise that faded in and out very slowly over 10 seconds. Obviously this sound's own intrinsic PAT would be a very wide probability distribution. Consequently, no matter what sound we tried to measure against this highly ambiguous reference, the resulting distributions would always be very wide. In other words, our measurements would not tell us very much about the sounds we were trying to measure.

This argues for using a short impulsive percussive sound as the reference, because we expect such sounds to have narrow and definite PAT distributions. Gordon tried this: his experiments I and II measured the PAT of a set of synthetic orchestral tones against a reference tone drawn from the same set, while his experiment III used a sample of a conga slap as the reference for measuring PAT of the same orchestral tones. Since the conga slap clearly had a more definite PAT one would expect his results for experiment III to be "better," that is for the measured probability distributions to be taller and narrower than in experiments I and II. In fact, he found the opposite: the standard deviations when testing against the conga slap were about 10-16 ms, while the same measurements against the synthetic clarinet or bassoon had standard deviations of only about 6-12 ms, as shown in Table 3.

Isochrony vs. Clarinet	~8-15 ms
Synchrony vs. Clarinet or Bassoon	~6-12 ms
Synchrony vs. 'Cello	~6-17 ms
Synchrony vs. Conga	~10-16 ms

Table 3: Standard deviations of PAT measurements from (Gordon 1987, 92)

Why was the drum a worse reference? "[A]cuity of temporal order was an important factor. Confusion was probably enhanced by the inherently different attack characteristics of the drum standard and the other 16 stimuli; the drum sound's rise time, as measured from physical onset to maximum amplitude, was less than 10 ms, whereas the quickest rise times among the other 16 stimuli were 45-50 ms. Some subjects reported simply giving up on several of the trials" (Gordon 1987, 94). In other words, Gordon is making a sort of "apples and oranges" argument that the simple difference in attack times made it more difficult for subjects.

I do not believe this explanation: in my experience playing in musical ensembles, there does not appear to be added difficulty synchronizing fast-attacking instruments with slower-attacking instruments, and in fact the overall rhythmic synchronization of an ensemble tends to be easier when percussive instruments are present. Gordon acknowledges this as well: "After all, the drum is the standard rhythm-setting instrument to which all other instruments synchronize" (Gordon 1987, 94).

I believe that Gordon's results in experiment III are due to auditory streaming effects.[82] Research has shown that listeners' ability to detect the temporal order of two sound events is much lower when they are further apart in pitch or spectral content: "It seems that accurate judgments of order require sounds to be in the same stream and that sounds with grossly different timbres resist being assigned to the same stream" (Bregman 1990, 94). Although Bregman is not referring specifically to perceptual attack time, a difficulty in discerning the order of sounds will obviously lead to a difficulty in discerning whether or not two sounds are perceptually simultaneous.

So what makes sounds have "different timbres"? The ANSI standard for "Psychoacoustical terminology" defines timbre as "that attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar" (ANSI 1973). This negative definition explains what timbre is *not* rather than what timbre *is*, and is therefore very broad. Which aspects of timbre are responsible for auditory streaming, and to what degree? Unfortunately, "much of the existing research on the grouping of timbres… [does] not ask what the dimensions of timbre [are], or the relative potency of different kinds of timbre differences" (Bregman 1990, 95). Bregman goes on to review many studies showing a strong effect of spectral content on auditory grouping; this is clearly an important factor.

Later he considers the temporal shape of the beginnings of sounds' amplitude envelopes (that is, "rise time"), first pointing out that the shape of the amplitude envelope determines the overall magnitude frequency spectrum,[83] and in particular that very fast onsets sound like broad

[82] I am deeply grateful to David Wessel for suggesting this interpretation.

[83] Considering that the definition of an amplitude envelope is a signal that is multiplied in the time domain by some underlying signal, and that multiplication in the time domain corresponds to convolution in the frequency domain, it is obvious that the spectral content of the amplitude envelope will have a strong effect on the spectral content of the

frequency clicks that separate perceptually from the rest of the tone (Bregman 1990, 114). In other words, the timbral difference supposedly caused by the shape of the amplitude envelope can also be explained as a difference in the frequency magnitude spectrum. He reports on some "informal observations" he made, in which he used "complex tones that are rich in harmonics" in an attempt to compensate for the spectral effect of differences in the rise time of amplitude envelopes: "It seemed to me that the difference in onset suddenness did make it more likely that the… tones would segregate from each other, but the effect was not strong" (Bregman 1990, 114-115).

Let us return to Gordon's Experiment III and the difficulty he found in measuring synthetic orchestral tones' PAT using a recorded conga slap as a reference. It is safe to say that the difference between the conga slap's overall magnitude frequency spectrum and that of any of the orchestral tones was probably much larger than the difference between any pair of tones.[84] What if the overall spectrum of the conga slap had somehow been close enough to the overall spectrum of the orchestral tones for them to be in the same auditory stream? This question motivates the next section.

3.7 Spectrally Matched Click Synthesis

We now have two criteria for a good reference sound to measure a test sound's PAT:

1. The reference sound should have a low-variance PAT-pdf, that is, its PAT should be maximally definite and unambiguous.
2. The reference sound should be close enough to the test sound in overall magnitude frequency spectrum that they will be perceived in the same auditory stream.

These criteria are somewhat contradictory. The extreme case to satisfy the first criterion would be a pure impulse, which would in theory have infinitely short duration. Our closest practical approximation to this would be the ideal digital impulse, that is, a digital recording consisting of a single sample with nonzero amplitude followed by a series of zero amplitude samples. But the ideal digital impulse has a totally flat magnitude frequency spectrum, that is, it contains equal energy at all frequencies, and so it will fail to meet the second criterion for any musical sound whose PAT one might want to measure.

resulting sound. In particular, if the amplitude envelope starts abruptly then its spectrum will be broad in frequency; see http://ccrma.stanford.edu/~jos/sasp/Relation_Smoothness_Roll_Off_Rate.html

[84] The original recording of the conga slap is now lost, so it is no longer possible to quantify this difference.

The extreme case to satisfy the second criterion would be to use a second copy of the test sound as the reference sound. This would guarantee that the spectra are identical and that the sounds would be in the same auditory stream, but we would expect the distribution of subjects' responses to have a mean of zero, since the PAT of the identical test and reference sounds would of course be equal. Therefore aligning a sound with another copy of itself cannot tell us anything about the center time of the PAT.[85]

Another case that would completely satisfy the second criterion while better satisfying the first criterion would be to use a minimum-phase version of the test sound as the reference sound (Smith 2007a).[86] This minimum-phase signal has a magnitude frequency spectrum identical to the test sound, but the phases would be changed so as to maximally concentrate the energy towards time zero. That does not mean that the PAT of the minimum-phase signal would be zero, but it would be closer to zero than the original, and as close to zero as possible for any sound with the exact same magnitude frequency spectrum.

What if we relax the second criterion somewhat, to allow reference sounds whose spectra are similar, but not necessarily identical, to the test sound? I call this *spectrally matched click synthesis ("SMC"):* given any test sound, create a short-duration[87] reference sound whose frequency magnitude spectrum is as close as possible to the original. This can be formulated as a finite impulse response (FIR) filter design problem, as shown in Table 4.

[85] However, if we repeatedly align a sound with itself, we can interpret the *variance* in the distribution as telling us something about the width of the sound's intrinsic PAT distribution; see Section 3.5.3 on page 50. For example, we would expect that the variance in the results of aligning short clicks with copies of themselves would be smaller than the variance in the results of aligning slow-attacking "mushy" sounds with copies of themselves.

[86] http://ccrma.stanford.edu/~jos/filters/Minimum_Phase_Means_Fastest.html

[87] This desire for short duration sounds is based on the assumption that all else equal they must have narrower PAT distributions. Another way of thinking about this is that a sound's perceptual attack time should occur during the window of time during which the amplitude is nonzero.

Spectrally Matched Click Synthesis	***FIR Filter Design***
Spectrum of test sound	Desired magnitude frequency response
Maximum duration of reference click	FIR filter order
Outputted reference click	Outputted filter impulse response
Play the click as a sampled sound	Filter incoming sound by convolving it with the filter's impulse response.

Table 4: Correspondence between spectrally matched click synthesis and FIR filter design

3.7.1 Methods for Spectrally Matched Click Synthesis

Finite impulse response (FIR) filter design is a rich and well-established field and there are many techniques for producing FIR filters from various specifications (Rabiner and Gold 1975; Smith 1983, 2007d; Williams and Taylor 2006).[88] I have not made any contribution to this field except to discover a new application for these techniques.

For spectrally matched click synthesis, the format of the filter design specification is a nonparametric sampled representation of the desired magnitude filter response, produced by the magnitude of the discrete Fourier transform (Smith 2007b) of the input test sound.

A good method for shortening the eventual impulse response (no matter what filter design strategy is used) is to apply *critical band smoothing* to this spectrum as a pre-processing step before filter design (Smith 1982, 1983). In general any form of smoothing that removes fine detail from the desired magnitude frequency response of the filter will tend to reduce the order of the filter and hence the duration of the click. Critical-band smoothing is a specific instance in which the spectrum is smoothed with a moving average filter whose width is in perceptual units of *Equal Rectangular Bandwidths* (Moore and Glasberg 1996)[89] In other words, instead of the moving average always encompassing a fixed bandwith in Hertz, the width of the moving average adapts in a nonlinear way matched to human auditory perception, so that a wider range of frequencies are averaged together in the high frequencies where the perceptual effect of smoothing by a fixed linear frequency bandwidth is less audible.

Matlab (Natick, MA: The MathWorks, mathworks.com)[90] is a numerical computing environment and programming language, and it comes with optional "toolboxes" that implement many filter-

[88] See http://ccrma.stanford.edu/~jos/sasp/Optimal_FIR_Digital_Filter.html

[89] http://ccrma.stanford.edu/~jos/bbt/Equivalent_Rectangular_Bandwidth.html

[90] See http://ccrma.stanford.edu/~jos/matlab for a description of Matlab and free alternatives such as GNU's Octave (www.octave.org).

design techniques. I chose their "fir2" filter design procedure,[91] which designs an FIR filter with an arbitrary magnitude frequency response by interpolating "the desired frequency response… onto a dense, evenly spaced grid of length *npt* (512 by default)... The filter coefficients are obtained by applying an inverse fast Fourier transform to the grid and multiplying by a window" (according to the Help browser's reference page for "fir2" in Matlab version 7.5.0.338). I used the default Hamming window.

The "fir2" procedure always returns a linear-phase filter, in other words, a filter with a symmetric impulse response. We use the filter as an SMC by playing the filter impulse response directly as an audio signal, not to filter other signals, so there is no advantage to having the filter be linear-phase. Instead, the goal is to make the sharpest possible attack, so that the PAT will be maximally definite; therefore I converted the output of "fir2" to a minimum-phase signal "by computing the cepstrum and converting anti-causal exponentials to causal exponentials"(Smith 2007a).[92] After converting the filter to minimum phase, there is often very little energy in the later portion of the impulse response, and so often the SMC may be further truncated by brute rectangular windowing with no audible consequences.

3.7.2 Other Applications of Spectrally Matched Click Synthesis

SMC's ability to produce arbitrarily short and impulsive sounds matching any input sound has applications beyond the production of reference sounds for PAT measurement experiments.

First of all, simply converting any signal to minimum phase makes it maximally percussive while retaining the exact magnitude frequency spectrum, so this can be used to make "drum-like" sampled sounds from arbitrary input material.

[91] I also experimented with some of Matlab's other FIR filter design procedures, many optimal in some sense but ultimately unsuited for this application. Most of these are for building specific "classical" filter types such as lowpass and are therefore not useful for SMC, where we need to specify an arbitrary sampled shape for the desired magnitude frequency response. The "firlpnorm" procedure (in the Filter Design Toolbox) performs "least P-norm optimal FIR filter design", but was not practical for producing filters clicks of longer than a few dozen samples in duration, because the running time appears to be exponential in the order of the filter. The "firpm" procedure (in the Signal Processing Toolbox) performs "Parks-McClellan optimal equiripple FIR filter design" but is not well-suited to SMC synthesis because the specification requires non-zero-width *don't care* regions between each pair of desired line segments; for SMC synthesis every region of the frequency spectrum might be important in keeping the SMC in the same auditory stream as the test sound. Matlab contains a whole family of filter design functions with the same problem, including "fircband" (in the Filter Design Toolbox) for "Constrained-band equiripple FIR filter design," "firgr" (in the Filter Design Toolbox) for "Generalized Remez FIR filter design," and "firls" (in the Signal Processing Toolbox) for "Linear-phase FIR filter design using least-squares error minimization." Note that the free Octave provides a version of the fir2 function and many of Matlab's other filter design implementations.

[92] See http://ccrma.stanford.edu/~jos/filters/Creating_Minimum_Phase_Filters.html for a description of this technique, and in particular http://ccrma.stanford.edu/~jos/filters/Matlab_listing_mps_m_test.html for Prof. Smith's freely-available and quite elegant one-line Matlab implementation:

```
        sm = exp( fft( fold( ifft( log( clipdb(s,-100) )))));
```

Creating a series of SMCs with different durations from a single input sound results in a sort of "morph" (timbral and temporal interpolation) between the original sound all the way to the ideal digital impulse, smoothly becoming shorter, more percussive, and more broad in frequency.

The result of mixing an SMC back in with an original sound, aligned so that the PATs are equal, often fuses into a single perceptual event that sounds just like the original but with a stronger attack.[93] By controlling the relative volume of the SMC and the original sound, and by varying the duration of the SMC, it's possible to increase the "attackiness" of any sound. As early as 1979 Wessel pointed out that sharpness of attack is one of the main perceptual dimensions of musical timbre and suggested that "both the fine tuning of rhythm in music and psychoacoustic research will benefit greatly if the control software of our synthesis systems allows easy and flexible adjustment of [attack characteristics] in complex musical contexts" (Wessel 1979). Almost every synthesis system does indeed provide easy and flexible adjustment of sharpness of attack *for synthetic sound* through features such as amplitude envelopes. The benefit of controlling sharpness of attack through this method of mixing in an SMC is that it applies to any arbitrary sampled sound, not just synthesized sound.

The method described above for converting SMCs to minimum phase requires computing the spectrum of the signal. To do this properly it is important to use sufficient zero-padding to produce enough frequency resolution to avoid time aliasing.[94] As a compositional effect, however, time-aliased minimum-phase signals produce a strange form of periodicity, with an impulsive burst of energy at the beginning of the signal, then a weaker and less distinct second attack at exactly the midpoint of the resulting signal. This generates a form of quasi-periodicty that could be used to advantage in the synthesis of rhythmic material.

Producing longer SMCs without converting to minimum phase can also produce some musically interesting results. At a duration of about 10ms or longer, the symmetric quality of the linear-phase impulse response manifests as a clear fade in and fade out around a central point. It might be interesting to explore what is the PAT of these symmetric linear-phase signals.

Finally, spectrally matched clicks can be useful for the common task of adding multiple copies of a new sound on top of an existing music recording to hear the output of an algorithm such as an onset detector or pulse tracker. In this case one generally wants a percussive sound that will clearly mark the instants output by the algorithm. Such a sound should be enough like the existing

[93] Because of the spectral matching, the SMC is much more likely to fuse perceptually with the original sound than an unrelated sound with the same sharp attack.

[94] http://ccrma.stanford.edu/~jos/sasp/Example_2_Time_Domain.html

recording that it will be easy to hear the relative timing of the algorithm's output against the music recording, but distinct enough that it will be clearly perceived as something added to the original recording. SMC's arbitrary tradeoff of spectral similarity for percussiveness makes it good for creating a click sound that blends with or stands out from the original recording in the desired amount.

3.7.3 Spectrally Matched Click Synthesis Future Work

Currently I isolate the sound to be matched by slicing it out in the time domain. For the SMCs described in the next chapter I was able to take the source sounds out of full music recordings because in each case I was able to find a segment of time containing only the desired sound event, with no other sounds playing. However, in general it would be desirable to be able to make SMCs matched to sound events taken out of arbitrary polyphonic mixes. The extremely difficult problem of extracting an individual sound event from a polyphonic mix[95] is much easier when the end result is an SMC, because all that is required is the approximate magnitude frequency spectrum, which may survive intact even through artifacts and other imperfections of the polyphonic source separation process. It would also be possible to compute the magnitude frequency spectrum *S* of the entire mix during the time span of the desired event, and then not even bother trying to recreate the desired event on its own, but instead partition *S*'s energy arbitrarily into the desired spectrum of the SMC and a residual spectrum representing all of the other sounds and noise being removed. For example, suppose we want an SMC matching a snare drum from a recording, but that snare drum always plays in unison with a hi-hat cymbal. If we can find a segment where the hi-hat cymbal plays a note alone, we can use it to approximate the spectrum of that instrument. Then we can find another segment where the snare drum and hi-hat play a note together, take the spectrum, subtract the spectrum of the hi-hat, and use the remainder to create an SMC of the snare drum.

As mentioned, there are very many algorithms for FIR filter design, some of which will probably perform better than "fir2" plus "mps" as described above. In this case we can define "better" as "producing a (perceptually) closer approximation to the desired magnitude frequency response for a given filter order." It may be the case that different techniques may be optimal depending on the duration of the SMC. It is also likely that an algorithm specifically designed to produce minimum-phase filters will perform better than the combination of designing a linear-phase filter and then converting the impulse response to minimum phase in a second step. Finally, it might be

[95] See, for example, (Master 2006).

advantageous to use a filter design method in which the minimized error between the desired magnitude frequency spectrum and the filter’s magnitude frequency spectrum is weighted perceptually by frequency region.

Chapter 4 Listening Experiment

4.1 Introduction

I carried out an Internet-based listening experiment in which a total of 57 subjects downloaded custom-built software for measuring perceptual attack time (PAT) using the synchrony method, in other words, adjusting the relative timing of various pairs of sounds until the attacks were perceived as synchronous.

4.1.1 Hypotheses

- Subjects will not exactly replicate their response for repetitions of the same trial, but instead will fit a probability distribution.
- The shapes of these probability distributions will vary based on the sharpness of attack and other characteristics of the musical material.
- These probability distributions will be narrower (i.e., subjects will repeat the same results more accurately) when the reference sound is spectrally more similar to the sound being tested.
- Subjects will be more accurate when the musical material establishes an understandable and predictable rhythmic context.

In addition to testing these hypotheses, another goal of the experiment was to make statistical models of various sounds' perceptual attack time.

4.2 Materials and Methods

4.2.1 Online paradigm

To measure the shape of the probability density function (pdf) for the difference between the PAT of a pair of sounds requires a large number of trials, since each trial provides only a single data point that can be interpreted as a sample drawn from the unknown pdf being measured. In other words, it takes many trials with the same condition to estimate the PAT-pdf for that condition. Comparing PAT-pdfs for the same test sound against multiple reference sounds further multiplies the number of trials required. To do this for more than a few sounds requires a large total number of trials, so I decided to use a web-based online experimental paradigm (Disley, Howard, and Hunt 2006; Honing and Ladinig 2008; Reips 2002).

My inspiration for using this paradigm was a study in which Honing found 162 subjects online for a test on tempo-specific timing in piano recordings of music by Bach, Beethoven, Chopin, and Schumann (Honing 2006). Another group of researchers used a similar paradigm to have 59 subjects rate twelve musical instrument samples on 15 scales labeled with timbral adjectives such as "clear," "ringing," and "nasal" (Disley 2006). In what is probably the largest experiment dealing with sound perception, Cox found 130,000 subjects to rate sounds on a six-point scale ("not horrible", "bad", "really bad", "awful", "really awful", and "horrible") in a "hunt for the worst sound in the world" (Cox 2007). In a similar spirit, but collecting huge amounts of data from people without enrolling them formally as experimental subjects, Slaney and White used almost 1.5 million jazz song ratings from 380,000 users to compute a similarity metric (Slaney and White 2007). What all this prior work has in common is that subjects listen to a collection of fixed sounds, and then answer multiple-choice questions about their perception of the sounds. My experiment required a much greater degree of interactivity, so it had to be administered by custom software (described in detail in Appendix A, page 145), not a simple web form.

4.2.2 Subjects

Professor Jonathan Berger kindly supported the pilot study for this experiment by incorporating it into Stanford's Music 151 course ("Psychophysics and Cognitive Psychology for Musicians") during Spring 2007. 17 of his students took the experiment in an early form, providing invaluable feedback about technical challenges, user interface design, and other aspects of the experiment. Although I had to discard about half these trials due to a problem with the sounds, I incorporated the remaining 1797 trials from the pilot study into the final analysis.

Another 40 subjects took the final version of the experiment. Although I took pains to allow subjects to participate anonymously in the experiment (as described in Section A.1.3 on page 149), none actually did so, so I know at least the email address of all these people. I recruited subjects by sending out an email to selected members of my personal lists of musician and computer music researcher contacts, to my colleagues at CCRMA and CNMAT, and also to appropriate mailing lists such as "AUDITORY" (researchers in auditory perception) and "sAmBiStAs!" ("discussion of performance of the music of Brazil"). Approximately half these subjects were people I knew personally.

Of the 57 total subjects, 38 said they were male, 10 female, and 5 declined to state. Subjects ranged in age from 19 to 62; Figure 17 is a stem-and-leaf plot (Tufte 2001, 140)[96] of the subjects' reported ages.[97]

```
1|99
2|0123345566677778889
3|001112344455688
4|1223389
5|
6|22
```

Figure 17: Stem-and-leaf plot of subjects' reported ages.

(An additional 7 subjects did not state their ages.)

4.2.3 Apparatus

Each subject used his or her own computer and sound system to take the experiment.[98] The software for the experiment runs on Windows or Macintosh computers[99] and is described in detail in Appendix A. Subjects needed to have an Internet connection to download the software and then again at the end of the experiment to email the results, but did not need to be connected to the Internet while taking the experiment.[100]

A major weakness of this Internet-based method of performing listening experiments is a lack of control over the audio hardware used by subjects (Disley, Howard, and Hunt 2006). An optional question at the end of each trial asked the subject "how are you listening to the sound from the computer";

Table 5 shows the frequency of each response.

[96] A stem and leaf plot is like a histogram, except that instead of simply showing the number of instances in the range corresponding to each bin it shows each individual value. In this case the bin width is ten years, so that the bin number is the first digit of the subject's age (base 10) and the numeral inside each bin is the second digit.

[97] For ethical reasons Stanford's Institutional Review Board required me to restrict participation in this experiment to volunteers at least 18 years of age.

[98] Some subjects in the pilot study took the experiment using computers at CCRMA.

[99] At least one would-be subject was not able to participate due to the lack of Linux support.

[100] At least one subject took the experiment through headphones on a long airplane flight.

Response	*Number of trials*
"Good headphones"	600
"Decent headphones"	239
"Bad headphones"	0
"Good speakers"	104
"Decent speakers"	218
"Built-in speakers"	180
"Bad speakers"	0
[no response]	2138

Table 5: Sound equipment subjects reported using.

These are the menu options available to answer the question "How are you listening to the sound from the computer?" as described in Section A.2.3 (page 164).

4.2.4 Stimuli

Table 6 lists the 20 sound files that were used in this experiment. Each of these 20 files is a sound example, whose name is given in the "Name" column, but with spaces replaced by underscores. (For example, the sound *Brazil loop* is the sound example named *Brazil_loop*.)

Most of these sounds were derived from John Grey's historic collection of short orchestral tones (Grey 1975), resynthesized sinusoidal models of short notes played by a variety of orchestral instruments and equalized perceptually for pitch (always E^b above middle C), duration (about 300 ms) and loudness, with no partial's frequency ever exceeding 10 kHz. I chose these tones because they were the subject of a pioneering study of PAT (Gordon 1987) as well as many studies of timbre (Gordon and Grey 1978; Grey 1975, 1977; Grey and Moorer 1977; Wessel 1979), and since I have both the original additive synthesis data (courtesy of David Wessel) and some of the results of Gordon's study (see Section 4.3.5.1 on page 87). I selected the trumpet, clarinet, and violin tones, because they are varied in terms of both their overall magnitude spectra and the timing of their attacks, and because this allowed me to replicate three of the conditions of Gordon's earlier study as described in Section 4.3.5 (page 87). Each "Grey tone" exists in the form of breakpoint function envelopes for amplitude and frequency trajectories for a small collection of sinusoids. I did not have access to the original resynthesized time-domain audio samples that Gordon used but instead synthesized them in Matlab. I linearly interpolated each amplitude and frequency trajectory, and I started each sinusoid's phase at zero and let instantaneous phase be simply the integral of instantaneous frequency (Wright and Smith 2005).

For each of these three tones I also synthesized three spectrally matched clicks ("SMC," as described in section 3.7, page 63) with durations of 1024, 512, and 256 samples (corresponding to about 23.3, 11.6, and 5.8 ms). I chose these durations according to my personal subjective sense of the closeness of the spectral match to the original. In each case, the 23-ms SMC sounded like good match, the 12-ms SMC sounded like a decent match, and the 6-ms SMC sounded like a click with just some of the characteristics of the original.

Name	***Description***	***Duration***
Brazil loop	One-bar loop of Brazilian *Maracatu* drumming (Crook 2005, 145-166) by members of *Maracatu Nação Estrela Brilhante* during an informal warm-up before a parade in February 2007, recorded and looped by me.	1959.2 ms
Brazil SMC23	Spectrally matched click ("SMC") made from an isolated note played in unison by *gonguê* (bell), *tarol* (snare drum) and *abê* (beaded gourd), taken from *Brazil loop*.	23 ms (1024)
Clarinet	Grey's E-flat Clarinet tone ("ecq716")	330 ms
Clarinet SMC23	Spectrally matched click made from *Clarinet*	23 ms (1024)
Clarinet SMC12	Spectrally matched click made from *Clarinet*	12 ms (512)
Clarinet SMC6	Spectrally matched click made from *Clarinet*	6 ms (256)
Funk loop	One-bar loop of James Brown's famous drum break from the song *Funky Drummer* (Greenwald 2002, 261-263; McGuiness 2005, 62-73; Stewart 2000, 304-305), originally performed by Clyde Stubblefield and looped by me.	2365.2 ms
Snare	Isolated single snare drum note taken from *Funk Loop*	170.1 ms
Snare SMC3	Spectrally matched click made from *Snare*	3 ms (398)
Ideal impulse	Single digital "1" embedded in a stream of "0" samples, aka the "Kronecker delta function," the "unit impulse", or "ideal digital impulse."	0.02 ms (1 sample)
Mauritania loop	One-bar loop from the metered portion of the instrumental introduction to *The Tortoise's Song (Ishteeb Laggatri)* by Khalifa Ould Eide and Dimi Mint Abba (World Circuit WCD 019, 1990), looped by me.	2053.7 ms
Mauritania SMC12	Spectrally matched click made from an isolated bass drum note from *Mauritania loop*.	12 ms (512)
Trumpet	Grey's Trumpet tone ("tpq642")	360 ms
Trumpet SMC23	Spectrally matched click made from *Trumpet*	23 ms (1024)
Trumpet SMC12	Spectrally matched click made from *Trumpet*	12 ms (512)
Trumpet SMC6	Spectrally matched click made from *Trumpet*	6 ms (256)
Violin	Grey's Violin tone ("vcq526")	344 ms
Violin SMC23	Spectrally matched click made from *Violin*	23 ms (1024)
Violin SMC12	Spectrally matched click made from *Violin*	12 ms (512)
Violin SMC6	Spectrally matched click made from *Violin*	6 ms (256)

Table 6: Sound files used in the experiment.

Parenthesized number in "duration" column is duration in samples.

I also selected three one-bar loop excerpts of rhythmic music. These were for a different kind of task, marking the times of individual events within the loop rather than simply aligning a pair of isolated events. (See Section 4.2.5.2 "Tasks" below.) For each loop I isolated one example instance of the sound the subject was supposed to mark, then made an SMC from the isolated sound. In the case of the *Funk loop* (for which the isolated sound was a single snare drum hit), I used the original isolated sound in addition to the SMC.

Finally, I also included the ideal impulse[101] so that I would have a reference whose absolute time location is known to within a single digital audio sample.

4.2.5 Procedure

4.2.5.1 Preliminaries

Each volunteer subject began by visiting the web page for the experiment.[102] That page contained links to download the software to run the experiment, installation instructions (which were simply to open the downloaded archive and then double-click the appropriate file), and an image showing the software's opening screen. It also addressed the popular question "how long will this experiment take?" Near the top was a link titled "click here to learn what this research is about and why it's important" that went to a separate page defining PAT, demonstrating it with some simple sound examples, and addressing the questions "why is perceptual attack time important?", "what will these experiments find out?", and "how do these experiments fit into the overall research?" Finally there was an email address allowing potential subjects to contact me with any questions.[103]

The software for the experiment (described in detail in Appendix A) began with a series of screens guiding the subject through various preliminaries:

- **Agreement** (Figure 66): Make sure subjects agree to participate in the experiment, offering a chance to quit.
- **What's the point?** (Figure 67): Provide some context for the experiment and its user interface.[104]
- **Email test** (Figure 85): Create a test message and make sure it opens properly in the subject's email program.

[101] http://ccrma.stanford.edu/~jos/filters/Impulse_Response_Representation.html

[102] http://ccrma.stanford.edu/~matt/together

[103] Some subjects asked me questions by email before beginning the experiment; "how long will it take?" was by far the most popular.

[104] Thanks to Michelle Logan for pointing out the need for this screen.

- **Personal information** (Figure 68). Ask subjects for name, age, gender, level of musical experience, and description of musical training, all optional. This additional data might become the subject of future research; also, asking these kinds of "filter questions… at the beginning of an experiment encourage[s] serious and complete responses" (Reips 2002, 254).
- **Volume adjustment**: Starting from zero, have subjects gradually increase the volume to a comfortable listening level (which then becomes the default initial volume for all trials). This screen also has a troubleshooting area that diagnoses a few ways that the software might not make a sound (such as incomplete download of the software or misconfigured audio settings in MSP or the operating system) and offers solutions.
- **Tap input method selection and calibration**: Ask subjects to choose whether they will enter taps (for multiple-tap trials) via the QWERTY keyboard or audio input, as described in Section A.2.2.
- **Explanation of the user interface** (Figure 73): A diagram showing the spatial layout of the keys used in the interface (which the user could view again at any time by pressing the question mark key).
- **Example trials** (Figure 70, Figure 71, and Figure 72): Three screens guiding the subject step by step through every aspect of the interface in the context of performing two example trials.

4.2.5.2 Tasks

Each task used the synchrony method to measure the relative PAT of two sounds. The trials came in a sequence of blocks with the same task, as shown in Table 7. Note that the blocks alternated between tasks based on aligning a single pair of isolated sounds (tasks beginning with "synchronize") and tasks based on entering times for multiple sounds against a repeating rhythmic musical example (tasks beginning with "mark each").[105]

[105] In retrospect, since many subjects did not complete the full 75 trials (see Figure 18), it might have been good to randomize the order of the trial blocks to avoid the bias of more total trials for the earlier tasks.

Trial Block	***Trial numbers***	***"Your task for this trial"***	***Fixed repeating sound ("loop")***	***Sound adjustable by the subject ("click")***
1	1-11	"Synchronize two synthetic tones"	*Clarinet, Trumpet,* or *Violin*	*Clarinet, Trumpet,* or *Violin*
2	12-14	"Mark each note of the snare drum"	*Funk loop*	*Clarinet*, *Snare*, *Snare SMC3*, *Brazil SMC23*, or *ideal impulse*.
3	15-41	"Synchronize click with tone"	*Clarinet, Trumpet,* or *Violin*	*Clarinet SMC23, Trumpet SMC23,* or *Violin SMC23*, or *SMC12* or *SMC6* matching the fixed sound, or *ideal impulse*
4	42-44	"Mark each note of the bell"	*Brazil loop*	*Brazil SMC23* (which captured the timbre of the bell), *Snare SMC3*, *Clarinet*, or *ideal impulse*.
5	45-72	"Synchronize two clicks"	A click sound (*ideal impulse, Snare* or any *SMC*)	Another click sound (*ideal impulse, Snare*, or any *SMC*)
6	73-75	"Mark each note of the bass (lowest pitched) drum"	*Mauritania loop*	*Mauritania SMC12* (which matched the timbre of the bass drum, *Clarinet*, or *ideal impulse*.
	Above 75	At this point each trial was selected randomly from one of the above.		

Table 7: Blocks of trials in the listening experiment.

With only three synthetic tones in my study it was easy to present all 9 possible ordered combinations, considering each tone against itself and each of the others and also swapping the fixed/adjustable roles (trial block 1). This would not have been practical in trial block 5 with the 14 click sounds (ideal impulse, SMCs from three sounds isolated from looped excerpts, and SMCs of three different durations from all three synthetic tones, plus the *Snare* sample that isn't technically a click), so the experiment used only 22 of the 196 possible ordered pairs. Nine of these were all of the possible ordered pairs from the subset consisting of *Ideal impulse, Snare*, and *Snare SMC3*.

For tasks based on aligning a single pair of isolated sounds, the fixed reference sound repeated every 600 ms. The movable test sound started in a random initial temporal relationship to the reference sound. For the "synchronize two synthetic tones" trials the initial offset (i.e., the delay between the physical onset of the two sounds) was chosen uniformly from the range ± 340 ms. For blocks in which one of the sounds was a short click, the initial offset was chosen uniformly from the range ± 113 ms. The subject could adjust the relative timing of the sounds either with the QWERTYUIOP keys (as shown in Figure 73) or an on-screen slider (as shown in Figure 69). Each pixel of the slider corresponded to about 1.2ms, whereas each key moved the test sound forward or backward by about 10 ms, about 5 ms, about 1 ms, about ¼ ms, or 1/44.1 ms (one audio sample) with respect to the reference.

For tasks based on marking multiple notes in a looping example, the repetition period was the length of the loop, as shown in Table 6. The subject first had to enter an initial time for each test sound by tapping on the keyboard or the audio input, then select each test sound (which would mute all the other test sounds) and fine-tune its temporal placement with respect to the loop.

For all trials the subject had complete control of both the total volume and the relative volume of the reference and test sounds, either via on-screen sliders (in increments of about 0.6 dB per pixel) or using the arrow keys (each of which changed relative or overall gain by about 1.2 dB per keypress). The software made no attempt to achieve a standard presentation level.

The subject could pause the experiment at any time, and was forced to do so (and encouraged to take a break) for 60 seconds every 15 minutes. Each trial lasted until the subject was satisfied with the resulting temporal alignment between test and reference sounds. At the end of each trial the subject had the opportunity to provide optional additional information about the trial (as shown in Figure 82) and the option to continue with another trial or pause and email the results so far. After 75 trials the software reminded the subject that only 75 trials were requested, but encouraged the subject to continue if desired (as shown in Figure 75). When the subject quit the software it would save the personal information and number of trials completed into text files; if the subject opened the software again it would go through the same series of opening screens (to re-confirm subject's agreement to participate, to re-configure the necessary settings, and to remind the subject of the user interface) and then resume from the next trial in the sequence.

4.3 Analysis

4.3.1 Removal of Bogus Trials

Because the software recorded the entire time sequence of user actions for each trial, it was possible to examine what the user did for each trial. In a few cases the user "accepted" the random initial time offset between two sounds in less than one second; I judged these kinds of trials (which generally tended to be extreme outliers) to be user errors due to lack of experience with the interface and removed them.

I also removed all trials from one pilot study subject who did the experiment in a noisy computer lab at CCRMA listening through the relatively quiet built-in speaker on a Mac mini computer; a huge proportion of these trials were outliers.

Also, I manually examined each trial whose result was more than three standard deviations from the mean. In each case I listened to the two sounds with the subject's final time offset for that trial (in other words, the final alignment that supposedly sounded "synchronized" to the subject), and if the result sounded obviously wrong (in other words, if it sounded like two distinct attacks separated in time) I labeled it as an outlier and removed it.

4.3.2 Subjects' Seriousness: Dropout, Time per Trial, and Trial Self-Ratings

One issue that must be addressed with online experiments is "to distinguish between serious and unserious responses" (Honing and Ladinig 2008, 5). The typical method for this is to measure *dropout*, the situation in which a subject begins but does not complete the experiment (Reips 2002).

```
 0|12225
 1|0111244455
 2|
 3|0046669
 4|2
 5|17
 6|8
 7|0244456
 8|18
 9|7788
10|00000001
11|
12|24
13|
14|
15|
16|
17|
18|
19|
20|
21|4
22|
23|
24|
25|
26|3
27|3
```

Figure 18: Stem-and-leaf plot of the number of trials completed by each subject.

Note the two modes at 75 and 100 trials (the number of trials requested for the full experiment and the pilot study respectively). Many subjects dropped out before completing 75 trials, and five subjects kept going with extra trials after "finishing" the experiment.

I did not keep a log of the number of times the website for the experiment was viewed or the number of times the software was downloaded, so there is no way to know how many people started the experiment (or thought about starting it) but did not email any trials. Figure 18 is a stem-and-leaf plot of the number of trials completed by each subject. Many of the single-digit responses come from students who spent a total of one classroom hour on the experiment.

The software automatically kept track of the total amount of time spent on each trial, including breaks. Of course there's no way to know the portion of this time during which the subject was actually listening and paying attention, though in some cases the log shows that the subject made no adjustments for a long period of time, so it's highly likely that this was a break. Figure 19 is a histogram of the time spent on each trial; the extreme outliers to the right were all from trials in which the subject must have taken a very long break.

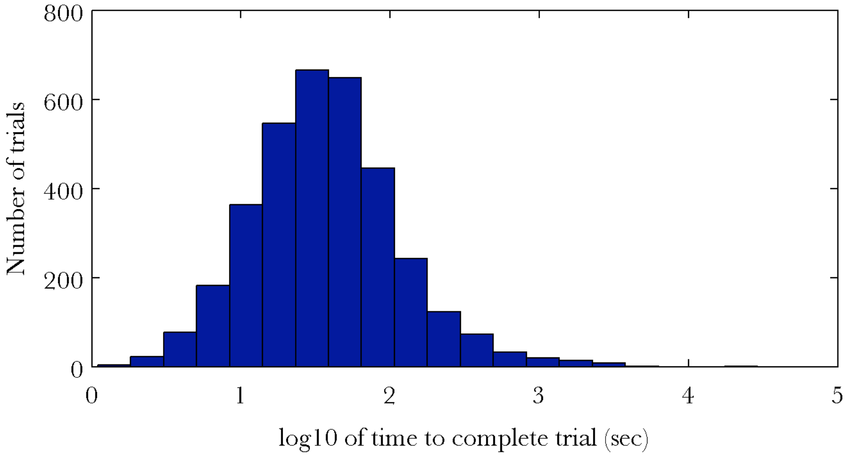


Figure 19: Histogram of (log) amount of time spent on each trial (including breaks). Most trials took about a minute.

An optional question at the end of each trial asked the following question: "On a ten-point scale of accuracy, precision, and how much you care about this trial and the experiment as a whole, how would you rate this trial?" These example responses calibrated the scale:

- Zero: "I don't care, I'm not paying attention, the data I'm giving you is random junk, I have no idea what you want me to do, the sound isn't working, etc."
- Five: "Nothing sounds exactly right to me, but this answer is more or less OK"
- Ten: "I listened as carefully as I could for as long as necessary to come up with my definitive best final answer."

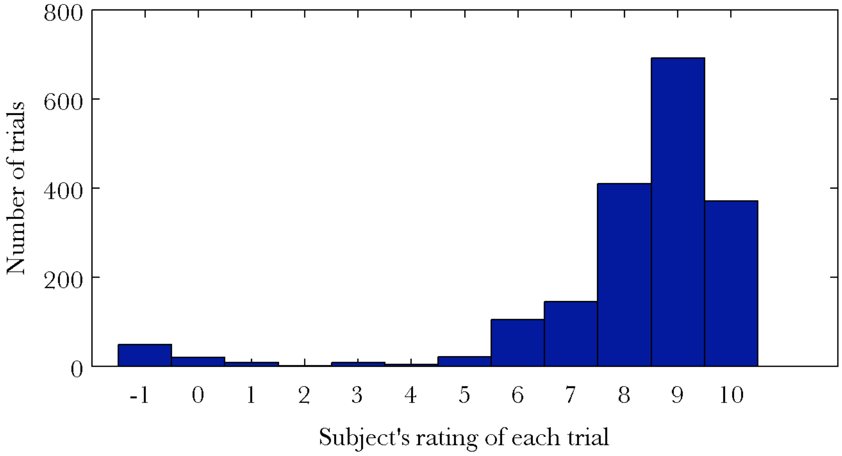


Figure 20: Histogram of subjects' holistic rating of each trial "on a ten-point scale of accuracy, precision, and how much you care about this trial and the experiment as a whole." (-1 indicates no response.)

Figure 20 shows the histogram of the frequency of each of the 11 possible responses to this question. There is a huge mode at the response "9," with the vast majority of trials (> 85%) having a response of 8, 9, or 10, and the frequencies falling off steadily towards 5. 97.9% of responses had a rating of at least five ("this answer is more or less OK"). Of the 19 trials that the subject scored as zero, all but one were tasks of the "mark each" variety, which was more difficult to understand and to perform than the tasks aligning pairs of isolated sounds.[106]

As one might expect, for the 1703 trials for which the subject gave a rating, there was a small but significant correlation ($r=0.12$, $p < 10^{-5}$) between the log of the time spent and the trial's rating.[107] Figure 21 compares these two variables in a scatter plot.

[106] On the other hand, one subject made the following comment after the second trial involving a pair of isolated tones: "Make it more interesting than just one repeated sound. Something more musical."

[107] To compute this correlation I removed trials that the subject did not rate. However, the unrated trials appear (encoded with a rating of -1) in Figure 21.

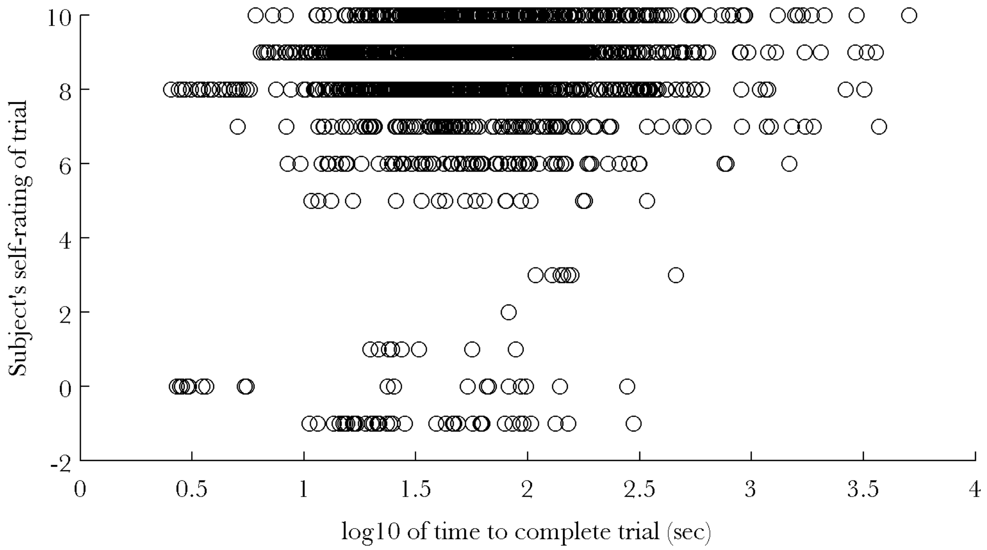


Figure 21: Scatter plot of subject's holistic rating of each trial versus the time spent on the trial. A "rating" of -1 indicates no response.

All in all, I am confident that the vast majority of these data come from trials in which the subject was taking the experiment seriously.

4.3.3 Method of Plotting Results Comparing Pairs of Sounds

Throughout the rest of this chapter I will display results graphically for PAT comparisons of various pairs of sounds, for example Figure 22. Each such figure contains two or three plots sharing the same X axis. The X axis is always relative time in milliseconds between physical onsets of the two sounds, which is equal to the relative time in milliseconds between the sounds' PATs. Labels on the two extremes of the X axis indicate the meaning of positive versus negative values of X; for example in Figure 22 the negative portion of the X axis is labeled "Trumpet earlier" while the positive portion is labeled "Clarinet earlier."

The bottom area contains one or more *box plots*, also known as "box and whisker" plots (Cleveland 1993, 25-27). Each displays the following:

- The median (the short red vertical line in the middle of each box)
- The positions of the 25th and 75th percentiles (the left and right edges of the box)
- The extent of the rest of the non-outlier data (the "whiskers," plotted as dotted lines extending outwards from the box), and
- The outliers (plotted individually with "+" and defined as data lying more than 1.5 times the interquartile range away from the median).

Above the box plots is an area containing the same number of one-dimensional scatter plots in the same vertical order, each of which simply shows every data point for the given pair of sounds. I added slight random jitter to the vertical axis to enhance readability.

Finally, for some plots (including Figure 22) the top area consists of nonparametric kernel density estimates of the shapes of the underlying probability density functions. A Gaussian distribution of fixed variance is placed around each data point and these are all summed together. The trick with these estimates is in choosing the standard deviation of each Gaussian, which in the context of kernel density estimation is called the "window width" *h*. The value of *h* that minimizes the asymptotic mean integrated squared error ("MISE") between the estimate and the true (unknown) underlying distribution is

$$h = (4/3)^{1/5} \sigma n^{-1/5} \approx 1.06 \sigma n^{-1/5}$$

where *n* is the number of data points and σ is an estimate of the standard deviation of the underlying distribution, the smaller of the sample standard deviation or 0.7418 times the interquartile range (Martinez and Martinez 2002, 280-285).

4.3.4 Interchangeability of Fixed and Moveable Sounds

For trials involving the synchronization of two isolated sound events (trial blocks 1, 3, and 5), the task was to "synchronize" the sounds by adjusting the physical onset time of the moveable one to make it "sound like it's lined up exactly" with the one that was repeating at a fixed rate. In other words, the task was to make the PATs of two sounds be equal to each other. By the symmetric property of equality, one would assume that subjects would arrive at the same the relative timing of a given pair of sounds regardless of whether the task was to move A with respect to B or to move B with respect to A. After all, what the subject hears in either case is an identical mixture of A and B. On the other hand, the fact that one sound is fixed in time while the subject can adjust the second sound might have some impact.[108]

There were six pairs of sounds for which there were trials using both possibilities of which was fixed and which was adjustable by the subject: all three pairs of Grey's *Clarinet*, *Trumpet*, and *Violin* tones, and all three pairs of the *Snare* sample, its spectrally matched click *Snare SMC3*, and the *Ideal*

[108] As an analogy, a harmonic series synthesized by additive synthesis will tend to fuse perceptually into a single tone, but if one of the partials moves independently in frequency then we will tend to hear it a separate sound object. Even if that partial's frequency ends up in the "correct" place in the harmonic series, the memory of its previous independent motion will often cause us to keep hearing it as a separate sound object. (See Bregman's description of his "Old-Plus-New Heuristic" (Bregman 1990, 222-227).) Perhaps there is some equivalent effect in the PAT alignment case.

impulse. The following six figures (Figure 22 through Figure 27) compare the results of swapping the test and reference sounds for each of these six pairs of sounds. (Of course it's always necessary to reverse the sign of each result when changing the sense of which was test and which was reference.)

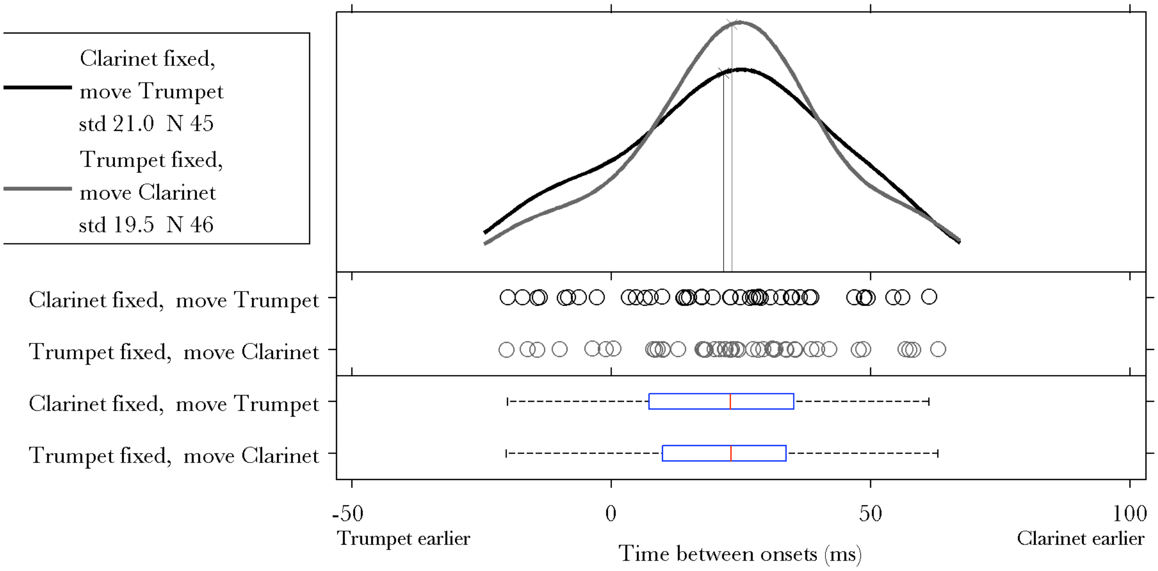


Figure 22: Comparison of trials with Trumpet and Clarinet depending on which was fixed and which was moveable.

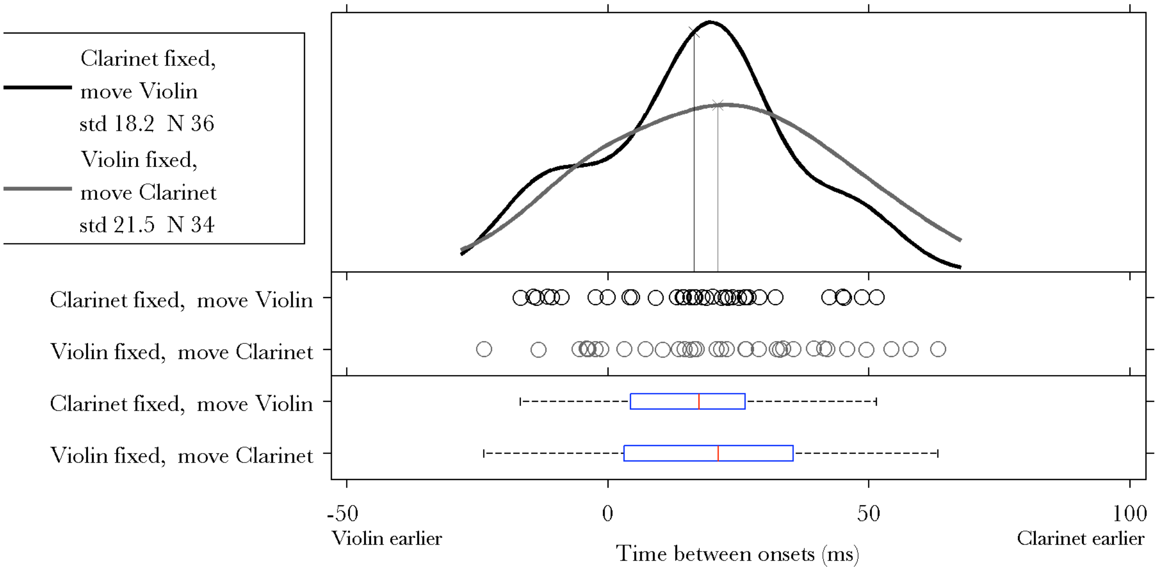


Figure 23: Comparison of trials with Violin and Clarinet depending on which was fixed and which was moveable.

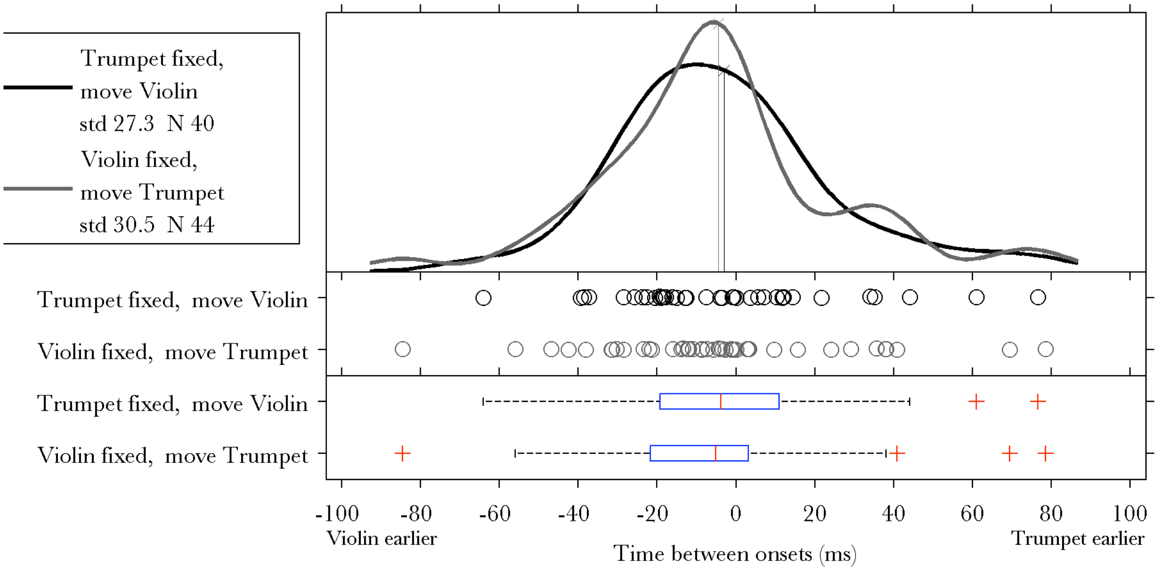


Figure 24: Comparison of trials with Trumpet and Violin depending on which was fixed and which was moveable.

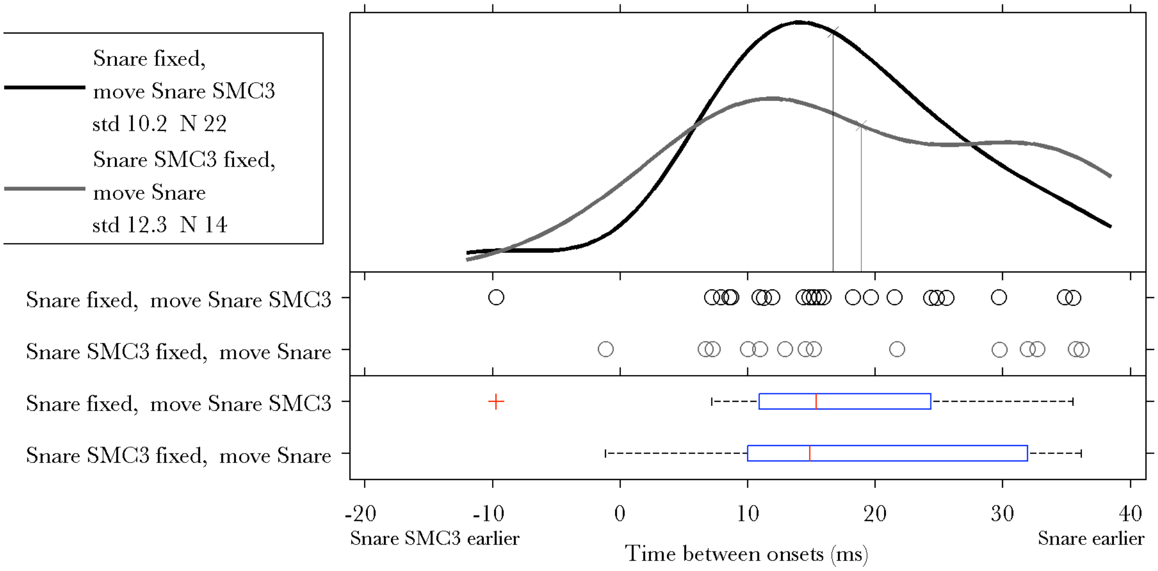


Figure 25: Comparison of trials with Snare and Snare SMC3 (Spectrally Matched Click created from Snare sample) depending on which was fixed and which was moveable.

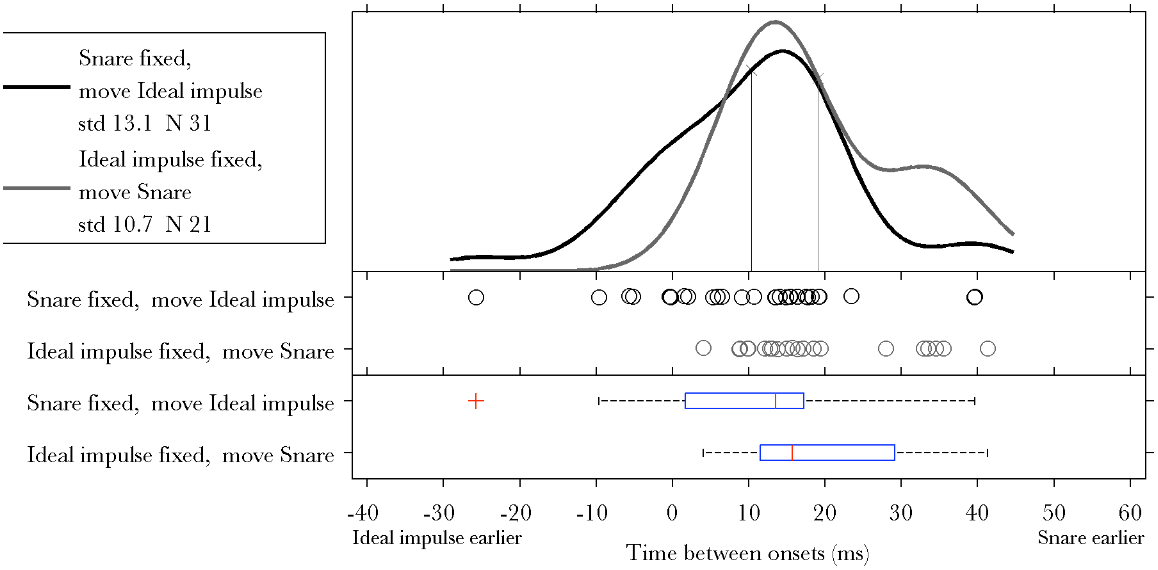


Figure 26: Comparison of trials with Snare and Ideal impulse depending on which was fixed and which was moveable.

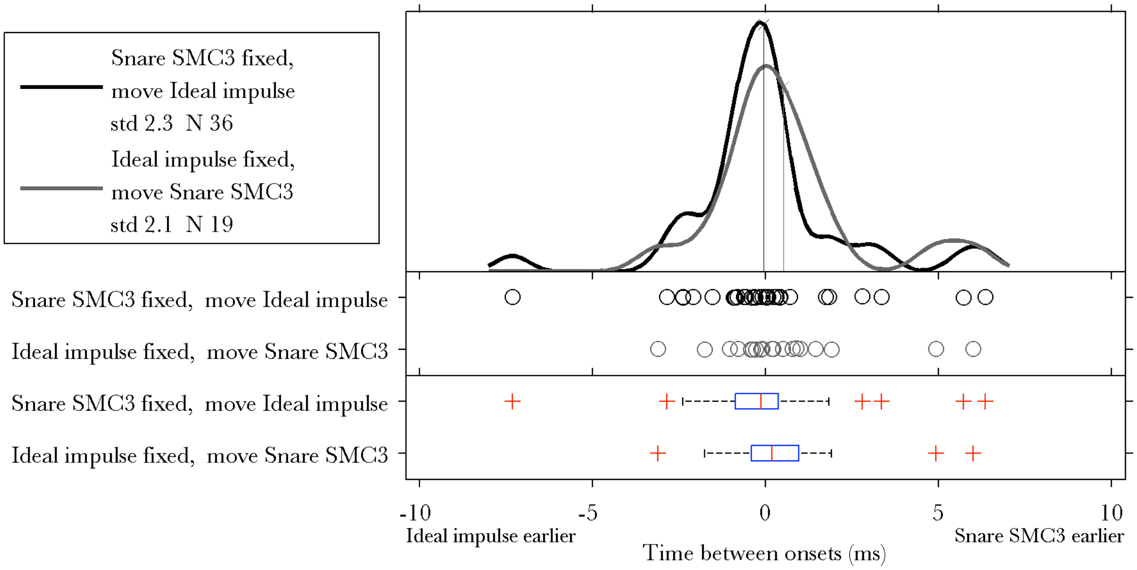


Figure 27: Comparison of trials with Ideal impulse and Snare SMC3 (Spectrally Matched Click created from Snare sample) depending on which was fixed and which was moveable.

For each of these pairs of sounds I also computed the two-sample Kolmogorov-Smirnov goodness-of-fit hypothesis test[109] to see whether the results seem to be drawn from the same underlying distribution. The null hypothesis of this test is that they are from the same

[109] I used the `kstest2` procedure from Matlab's Statistics Toolbox.

distribution, and in no case was the test able to reject the null hypothesis.[110] Table 8 shows the results.

Sound A	***Sound B***	$N_{A,B}$	$N_{B,A}$	***Result***	***pval***
Clarinet	Trumpet	46	45	same	0.73465
Clarinet	Violin	34	36	same	0.35733
Snare	Snare SMC3	14	22	same	0.51087
Snare	Ideal impulse	21	31	same	0.086796
Snare SMC3	Ideal impulse	19	36	same	0.49088
Trumpet	Violin	44	40	same	0.81767

Table 8: For the six pairs of sounds for which there were trials with each sound in both the fixed and moveable role, the result of the two-sample Kolmogorov-Smirnov goodness-of-fit hypothesis test of whether the samples seem to come from the same underlying distribution.

$N_{A,B}$ ***is the number of trials with A moveable and B fixed;*** $N_{B,A}$ ***is the number of trials with B moveable and A fixed. In every case the result is "same," meaning that this test was unable to reject the null hypothesis that the two samples were drawn from the same underlying probability distribution.***

I conclude that the distribution of results does not in fact depend on which sound was fixed and which was moveable, so for subsequent analysis I combined the results from both conditions for each pair of sounds.

4.3.5 Reproduction of Gordon's 1985 Results

As mentioned previously, John Gordon's pioneering PAT measurement work (Gordon 1987; Gordon 1984) was an important inspiration for this study, and I was fortunate to be able to re-use the same synthesized orchestral tones (Grey 1975) over twenty years later in this experiment. This section compares the old and new results for the conditions that I replicated.

4.3.5.1 Resurrecting Gordon's Results

Although Dr. Gordon was kindly willing to share his data with me, he no longer had a copy of any of the results of his experiment, and an attempt to resurrect his circa-1984 files from CCRMA backups indicated that this would not be easy, and likely impossible.[111] It appears that the only

[110] Lack of proof that the distributions are different is not proof that they are the same, but it's better than nothing.

[111] I am grateful to Fernando Lopez-Lezcano and Bill Schottstaedt for their intrepid assistance in this endeavor.

surviving results of his experiment are the published figures (Gordon 1987; Gordon 1984). Therefore I reverse-engineered the figures from a PDF file of a scan of Gordon's paper. (See Figure 12 [page 40] for an example copied [with permission] from Gordon's paper.) Specifically, I copied each graph as a digital image, being very careful to orient the lower left corner of the "copy" rectangle exactly on the origin of each graph, and pasted it into an image file, which I then cleaned up in Adobe Photoshop (e.g., to get rid of random stray pixels presumably added by the process of scanning the original document to create the PDF, and also to remove the numerals and vertical lines that had been added on top of the graphs), and then finally brought into Matlab, where I computed the vertical centroid of the "ink" (i.e., degree of blackness of each pixel) to generate a Y data point for each X value on my axis.

The scaling of the Y axis is immaterial, since it represents relative probability. What about the scaling of the X axis (representing time)? I needed a function to map each horizontal pixel position to a time in milliseconds. Luckily for me, each of Gordon's graphs has the same scaling for the X axis and tick marks every 10 ms, so I only had to do the following operation once. First I copied and pasted just the X-axis tick marks to an image file. Since each of these tick marks had a horizontal width of two or more pixels (black in the middle and grey on the ends), I visually estimated a fractional pixel value to be the "center" of each tick mark. I then performed a linear regression through my table of tick times (-50 to +50 in increments of 10) versus estimated fractional pixel positions to find the desired function.

4.3.5.2 Similarities and Differences Between the Two Experiments

By including the Clarinet, Trumpet, and Violin sounds in this experiment and including trials synchronizing each of the three to the Clarinet sound, I was able to give subjects the same sounds and the same tasks as three of Gordon's conditions, namely, measuring PAT of each with the synchrony method against the Clarinet reference.

One subtlety about the stimuli has to do with the additive synthesis necessary to recreate these three sounds. I was not able to reuse the exact synthesized time-domain samples from Gordon's experiment, but instead resynthesized them from the additive synthesis breakpoint functions as described in Section 4.2.4 (page 73).[112] Gordon also used linear interpolation of amplitude and frequency trajectories and also started each oscillator's phase at zero (John Gordon, personal communication, Jan 24 2008).

[112] I probably would have resynthesized these sounds anyway, even if I had the originals, because of differences in sampling rates.

In terms of apparatus, each of Gordon's subjects performed the experiment in the same "quiet listening environment" listening to sound through the same "single loudspeaker having a wide, flat frequency response and situated about 8-10 feet directly in front of the subject" (Gordon 1984, 32); my subjects each performed the experiment through different audio hardware and in a different physical location, as described in Section 4.2.3. The digital audio sampling rates were different in the two experiments, but this is immaterial because the additive synthesis data was band-limited.

The software for administering the experiment was similar in the two cases, with different keys changing the relative delay by varying amounts (in both cases down to a single audio sample) and visual feedback of the current delay shifted by a random amount so as to prevent subjects from choosing a value such as zero. My software, however, also provided an on-screen slider (shown in Figure 69) as a second option for changing the relative times. Perhaps more importantly, my software also provided on-screen sliders allowing the subject to adjust the relative volume of the two sounds. Another difference is that in Gordon's experiment "each trial began with the relative onset times ~120 ms apart; the subject was thus forced to synchronize a pair of tones that initially was obviously asynchronous" (Gordon 1984, 69), whereas in my experiment the relative onset times were chosen randomly from a uniform range and might have initially sounded synchronous or close to synchronous in some cases.

All 8 of Gordon's subjects "were experienced in computer music and considered to have well-trained ears" (Gordon 1984, 32), and each performed a total of 270 of these kinds of trials (6 instances each of 15 test tones times 3 reference tones), so his N=48 for each of the three pairs of sounds. My 55 subjects had a wide range of backgrounds, degree of familiarity with computer music, and musicianship, and each performed an average of about 7 trials aligning these particular three sounds.[113] My N is 45 for Clarinet against Clarinet, 91 for Clarinet against Trumpet, and 70 for Clarinet against Violin.

[113] Table 7 indicates that each subject should have performed 11 trials. The actual average number is lower because some subjects dropped out before completing the first trial block (as shown in Figure 18) and because some of these trials were bogus or extreme outliers as described in Section 4.3.1.

4.3.5.3 **Comparison**

The next three figures (Figure 28 through Figure 30) plot an average shifted histogram[114] of my results for each pair of sounds against Gordon's published results for the same pair.[115] In each case they are somewhat similar, with the modes of the two distributions only a few milliseconds apart. For both Clarinet vs. Trumpet and Clarinet vs. Violin the distributions from my experiment extend further to the left (which means further in the direction of starting the Clarinet earlier than the other tone) than Gordon's. Perhaps the real difference is that my results have a higher variance because of the greater variability in my subjects' backgrounds and sound hardware (and the simple fact of my having fewer trials each from a larger number of subjects); the fact that this added width is on the left of these figures may be a coincidence.

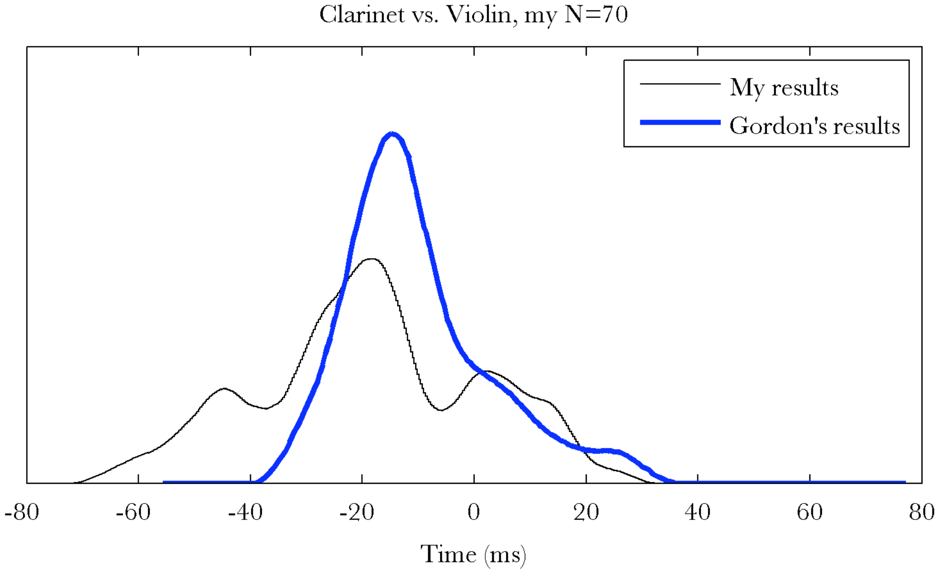


Figure 28: Comparison of Gordon's versus my results synchronizing Clarinet against Violin

[114] To make an average shifted histogram from a set of data points, first make a series of histograms from the same data points and with the same bin widths, but steadily varying the absolute starting position of all bins, then add them all together (Martinez and Martinez 2002, 274-280).

[115] For every figure in this section I scaled the heights of the two curves to make their areas equal.

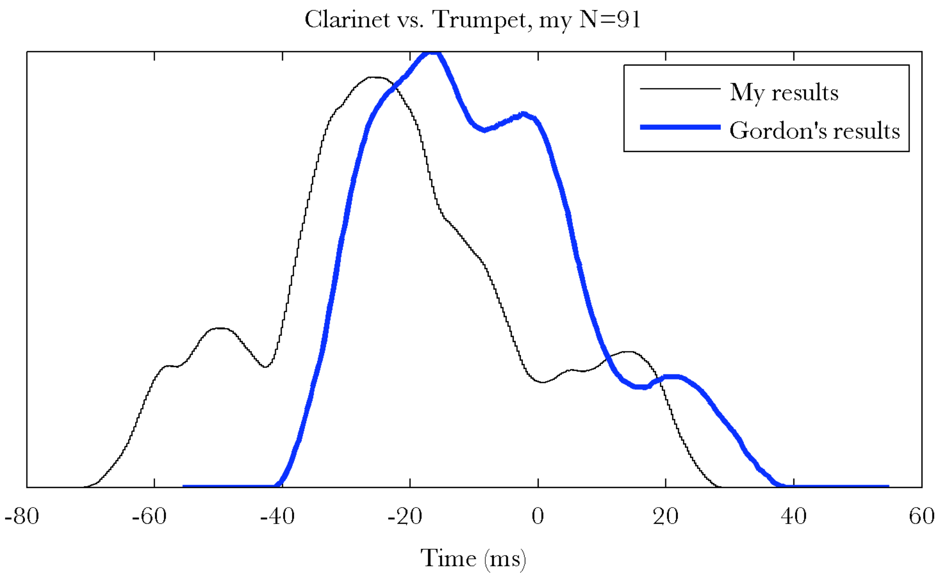


Figure 29: Comparison of Gordon's versus my results synchronizing Clarinet against Trumpet

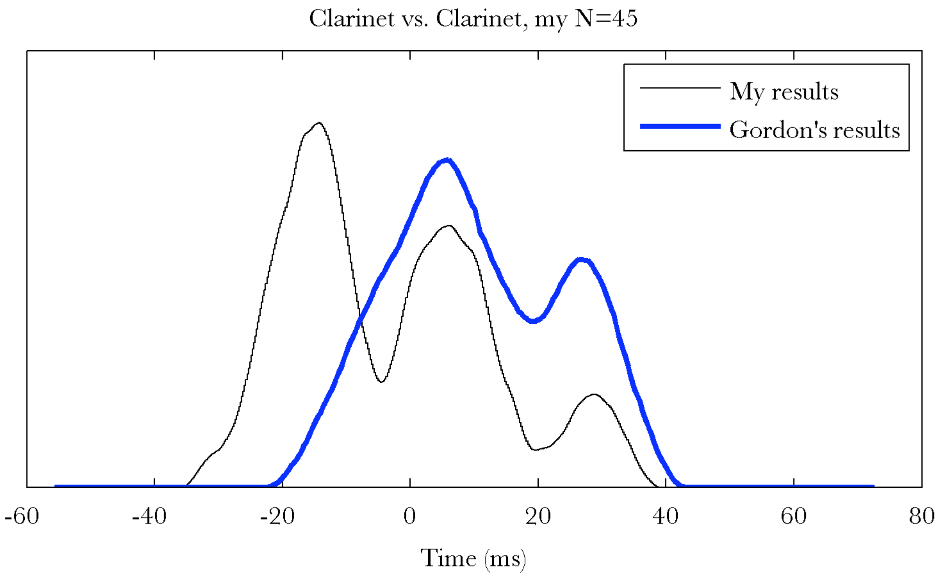


Figure 30: Comparison of Gordon's versus my results synchronizing Clarinet against Clarinet.

The two distributions in Figure 30 are remarkably similar in that both have a mode near time 6 ms and another mode around 21 ms, but quite different because mine also has another mode (and the largest) near time -14 ms, in a region where Gordon's distribution has fallen almost to zero. Since these trials compared two instances of the same Clarinet sound, we would expect the results to be symmetric, as Section 3.5.3 (page 50) suggests. So I artificially made each distribution symmetric by adding a second copy that had been time-reversed around the zero point. Figure 31 compares the resulting pair of symmetric distributions: we see that the formerly distinct structure of the prominent modes has been blurred away, and that they now appear similar only in that they both appear vaguely Gaussian.

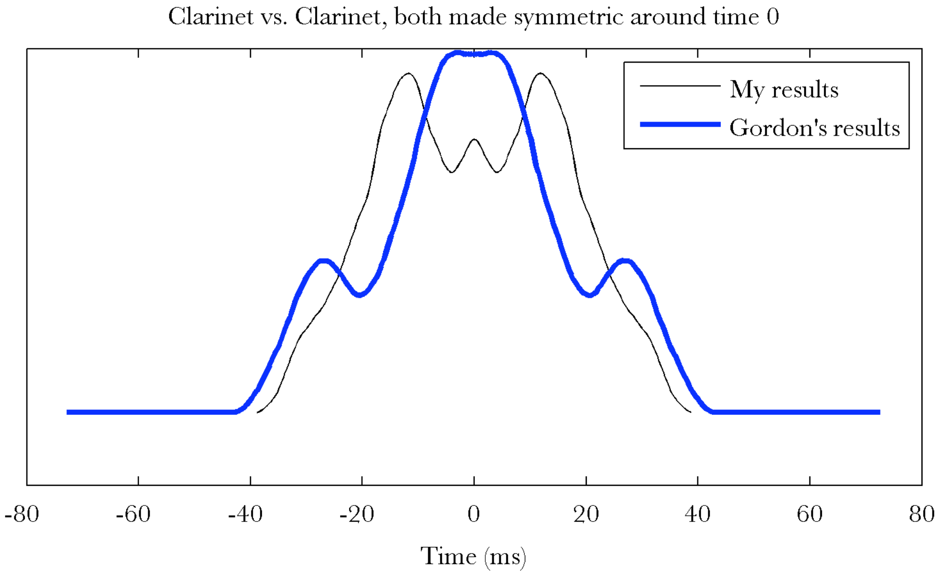


Figure 31: Comparison of Gordon's versus my results synchronizing Clarinet against Clarinet, after artificially forcing both distributions to be symmetric around time zero.

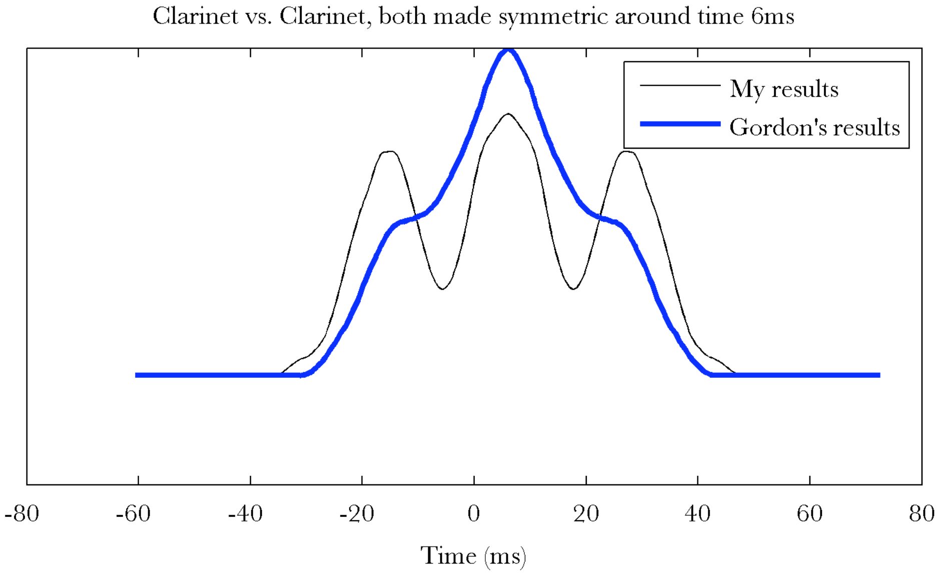


Figure 32: Comparison of Gordon's versus my results synchronizing Clarinet against Clarinet, after artificially forcing both distributions to be symmetric around the 6 millisecond point.

I again forced each distribution to become symmetric, but this time time-reversing around the 6 millisecond point instead of time zero. Figure 32 compares the resulting curves: the central mode is now still distinctly visible (by construction), but in Gordon's results the side mode has been almost smoothed away, while in mine the side mode is almost as prominent as the central mode.

4.3.6 Test of Normality of Results

I used Lilliefors' goodness-of-fit test of composite normality to see whether the results from each pair of sounds fit a Gaussian distribution.[116] The null hypothesis for this test is that the data are normally distributed with unspecified mean and standard deviation. Of the 48 pairs of sounds in this experiment, 12 of the distributions were found to be non-Gaussian, as shown in Table 9, and the remaining 36 were found to be Gaussian, as shown in Table 10.[117] Figure 33 through Figure 36 show histograms of the non-normal distributions superimposed against Gaussian bell curves with the sample mean and sample standard deviation.

Sound A	***Sound B***	***N***	***Skewness***	***Kurtosis***	***Nrml?***	***P-val***	***T.S.***
Snare	self	40	0.0837	3.93	no	<0.001	0.218
Snare SMC3	Ideal impulse	55	0.418	5.98	no	<0.001	0.183
Ideal impulse	Violin SMC6	37	1	5.42	no	<0.001	0.203
Violin SMC6	self	20	1.63	5.74	no	<0.001	0.27
Snare	Ideal impulse	52	-0.121	4.02	no	0.00175	0.161
Trumpet	Violin	84	0.506	4.17	no	0.00255	0.125
Ideal impulse	Maurit. SMC12	32	0.747	4.63	no	0.00974	0.18
Snare SMC3	self	33	1.12	6.43	no	0.0128	0.173
Trumpet	Trumpet SMC6	28	0.747	3.1	no	0.0156	0.184
Clarinet	self	45	0.512	2.42	no	0.0172	0.146
Maurit. SMC12	self	28	1.44	6.02	no	0.0192	0.181
Violin	Violin SMC6	36	0.708	2.85	no	0.0233	0.158

Table 9: Results of Lilliefors' goodness-of-fit test of composite normality for the pairs of sounds whose distributions are not normal.

The sample skewness measures a distribution's asymmetry.[118] The sample kurtosis is a measure of how much a distribution is "peaked"; it would be exactly 3 for a perfectly normal distribution. "Nrml?" stands for "Normal?"; all of the distributions in this table were found not be normal. "P-val" is the P-value giving the statistical significance of this result, and "T.S." is the test statistic. Results are sorted by P-value.

[116] I used the `lillietest` procedure from Matlab's Statistics Toolbox.

[117] Actually, for these 36 sounds all we know is that Lilliefors' test was unable to reject the null hypothesis that the distributions are normal, which is not quite the same as determining that they are normal.

[118] Skewness is zero for a perfectly symmetric distribution. To help give some intuition for the units of the skewness measure, the skewness of a distribution of points each of which is the *absolute value* of a sample drawn from a zero-mean unit-variance Normal distribution (i.e., a Gaussian bell curve chopped in half exactly at the midpoint) is about 1. In other words, this Matlab expression will always have a value close to one:

Sound A	***Sound B***	***N***	***Skewness***	***Kurtosis***	***Nrml?***	***P-val***	***T.S.***
Ideal impulse	self	91	-0.586	4.1	yes	0.0554	0.0921
Snare	Snare SMC3	36	0.0165	2.7	yes	0.0634	0.142
Violin	Violin SMC23	28	0.936	3.21	yes	0.0874	0.153
Clarinet SMC23	self	27	-0.0618	3.22	yes	0.0921	0.155
Brazil SMC23	Ideal impulse	16	1.6	6.7	yes	0.105	0.194
Violin	Violin SMC12	29	0.471	3	yes	0.125	0.144
Violin SMC23	Violin SMC6	20	0.773	3.62	yes	0.126	0.171
Clarinet SMC6	Violin SMC6	21	-0.323	2.11	yes	0.138	0.165
Trumpet SMC23	Violin	30	-1.03	4.54	yes	0.158	0.136
Trumpet	self	37	0.482	2.48	yes	0.158	0.123
Clarinet SMC23	Violin SMC6	14	-0.0433	1.72	yes	0.221	0.183
Clarinet	Ideal impulse	33	0.0387	1.93	yes	0.276	0.118
Clarinet	Trumpet	91	-0.189	2.57	yes	0.28	0.0723
Clarinet	Clarinet SMC23	38	-0.0475	2.33	yes	0.283	0.11
Clarinet	Clarinet SMC12	33	-1.05	4.45	yes	0.285	0.117
Brazil SMC23	self	24	0.0632	2.38	yes	0.329	0.132
Clarinet	Violin	70	0.0308	2.45	yes	0.382	0.0767
Clarinet	Trumpet SMC23	34	0.0859	2.47	yes	0.388	0.108
Clarinet SMC6	self	25	-0.631	3.18	yes	0.401	0.124
Ideal impulse	Trumpet	34	-0.135	1.95	yes	0.431	0.105
Violin	self	36	0.395	3.24	yes	0.453	0.101
Ideal impulse	Violin	27	-0.101	2.87	yes	0.484	0.113
Clarinet	Clarinet SMC6	33	-0.27	2.51	yes	>0.5	0.0944
Clarinet	Violin SMC23	27	-0.486	2.85	yes	>0.5	0.112
Clarinet SMC23	Clarinet SMC6	16	-0.153	1.94	yes	>0.5	0.119
Clarinet SMC23	Ideal impulse	30	-0.324	2.3	yes	>0.5	0.0924
Clarinet SMC23	Trumpet	28	-0.0174	2.51	yes	>0.5	0.0755
Clarinet SMC23	Violin	36	-0.369	2.51	yes	>0.5	0.0856
Clarinet SMC23	Violin SMC23	16	-0.128	2.54	yes	>0.5	0.131
Clarinet SMC6	Ideal impulse	19	-0.156	2.77	yes	>0.5	0.123
Clarinet SMC6	Violin SMC23	13	-0.284	3.09	yes	>0.5	0.114
Ideal impulse	Violin SMC23	21	0.623	3.62	yes	>0.5	0.119
Trumpet	Trumpet SMC23	32	-0.117	2.56	yes	>0.5	0.0898
Trumpet	Trumpet SMC12	31	0.0437	2.35	yes	>0.5	0.0681
Trumpet	Violin SMC23	30	0.222	3.04	yes	>0.5	0.0857
Violin SMC23	self	24	-0.0746	3.49	yes	>0.5	0.114

Table 10: Results of Lilliefors' goodness-of-fit test of composite normality for the pairs of sounds whose distributions are normal.

Columns are the same as in Table 9.

```
skewness(abs(randn(1,10000)))
```

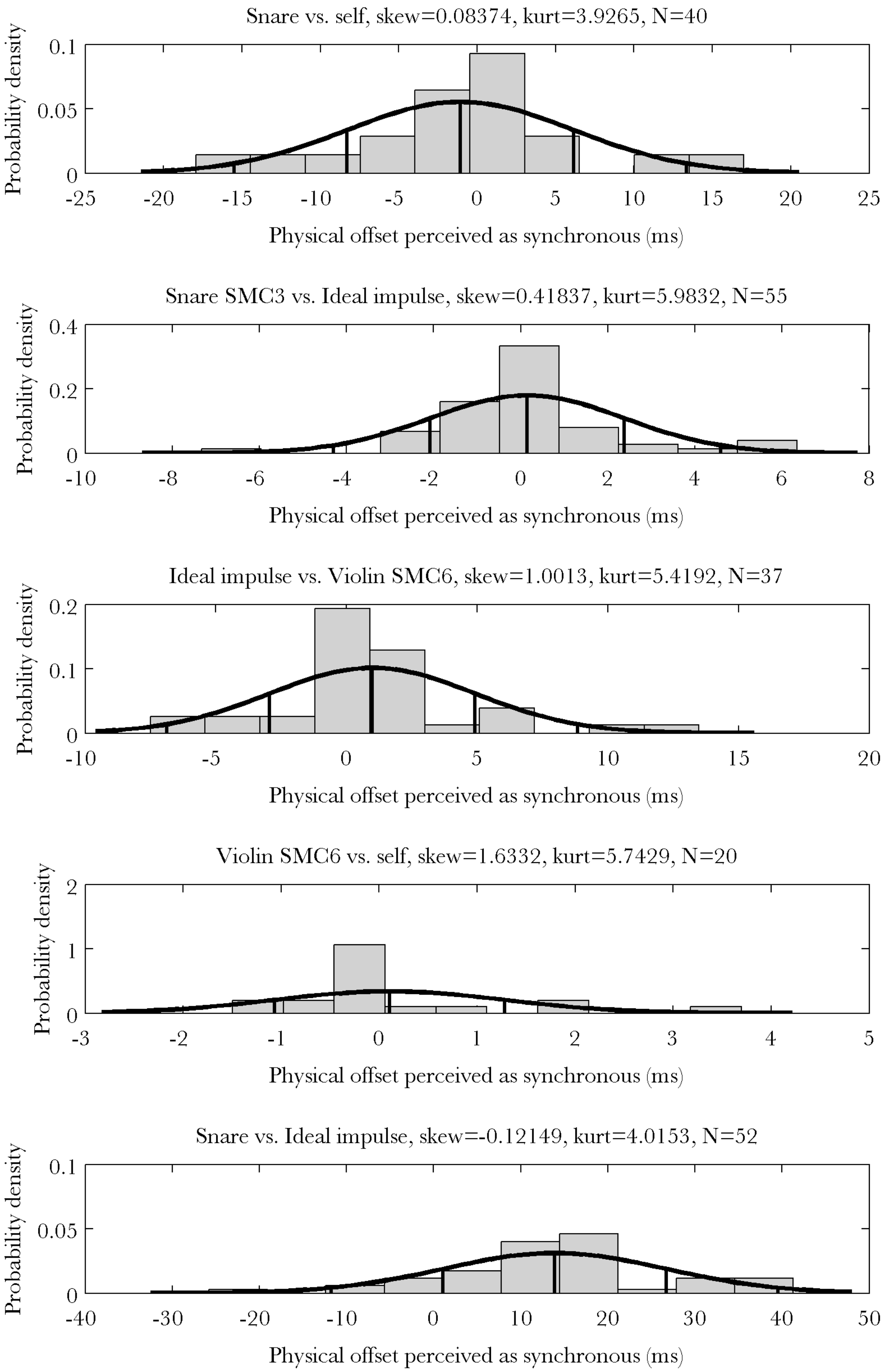


Figure 33: Histograms of the five least likely to be normal distributions from this experiment, with normal curve (of sample mean and sample variance) superimposed. Thick vertical lines indicate the mean and ± 1 and 2 standard deviations from the mean

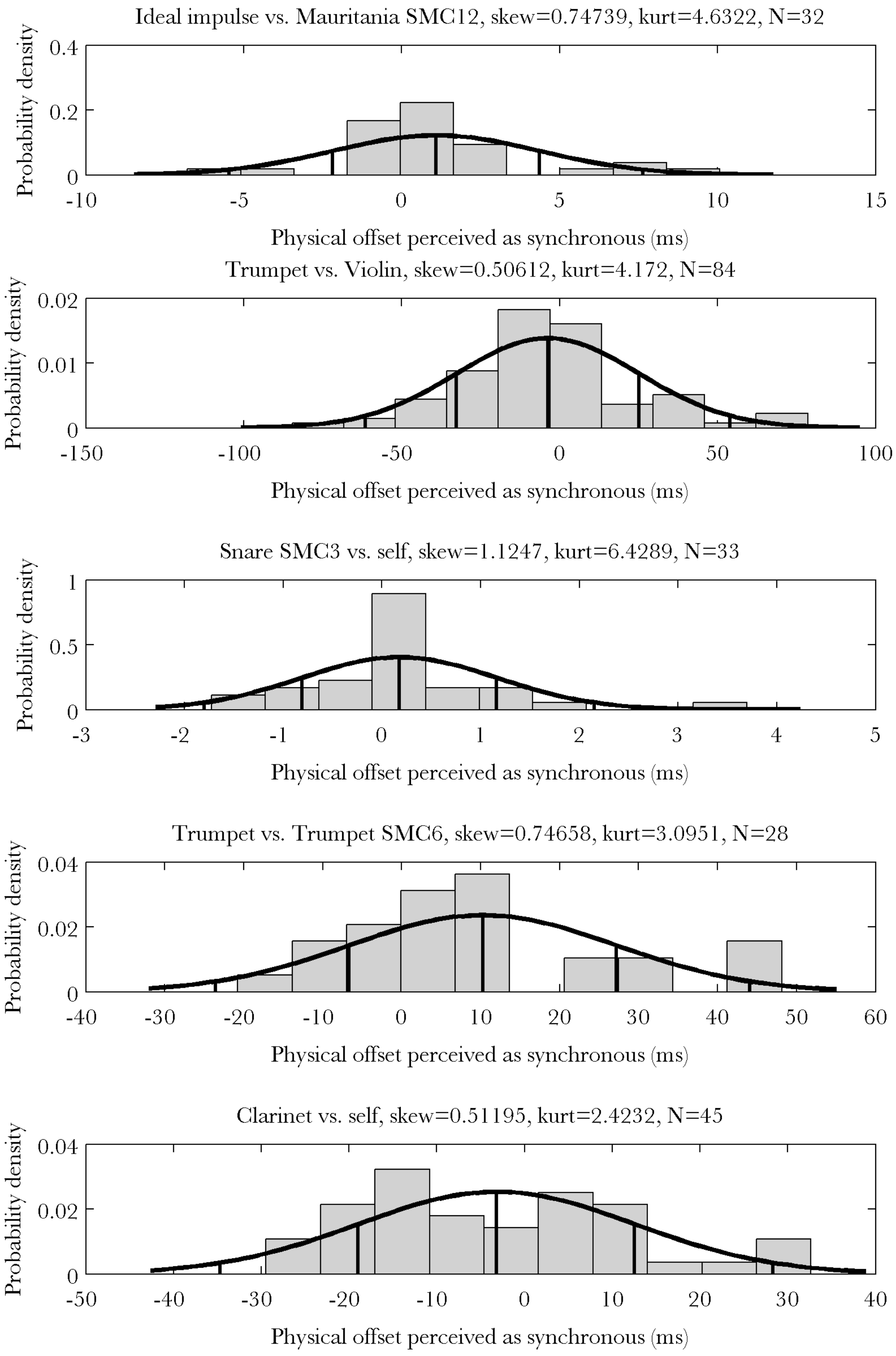
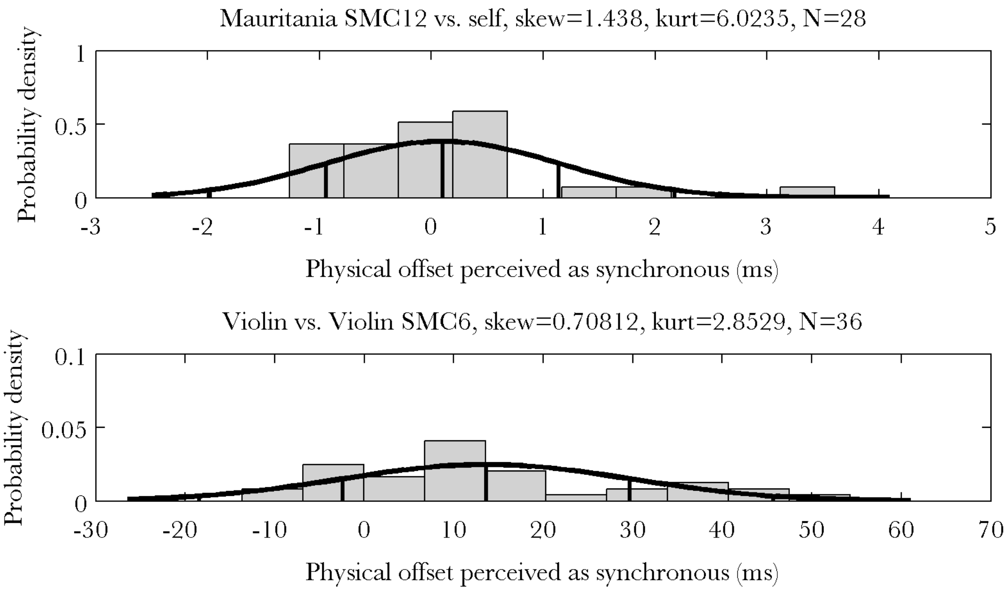


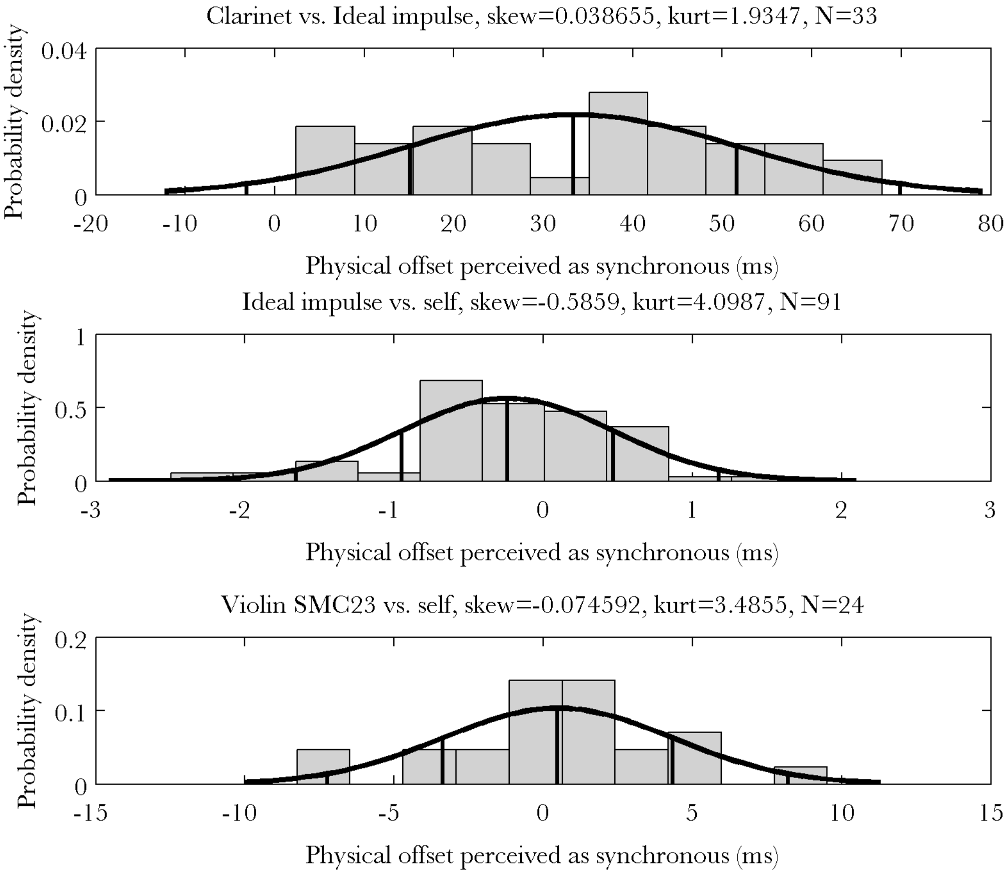
Figure 34: Histograms of the next five least likely to be normal distributions (after those shown in Figure 33). Again a normal curve (of sample mean and sample variance) is superimposed over each.


Mauritania SMC12 vs. self, skew=1.438, kurt=6.0235, N=28

Violin vs. Violin SMC6, skew=0.70812, kurt=2.8529, N=36

Figure 35: Histogram of the final two non-normal distributions (after those shown in Figure 33 and Figure 34), again with normal curves superimposed.

To compare, Figure 36 shows the same kind of plots for a few examples of normal distributions.


Clarinet vs. Ideal impulse, skew=0.038655, kurt=1.9347, N=33

Ideal impulse vs. self, skew=-0.5859, kurt=4.0987, N=91

Violin SMC23 vs. self, skew=-0.074592, kurt=3.4855, N=24

Figure 36: Histograms of three normal distributions with normal bell curves.

4.3.7 Quartiles of Each Distribution

Sound A	***Sound B***	***0%***	***25%***	***50%***	***75%***	***100%***
Brazil SMC23	Brazil SMC23	-1.9	-0.4	0.1	1.1	2.2
Brazil SMC23	Ideal impulse	-2.4	-0.2	0.2	1.1	6.5
Clarinet	Clarinet	-29.5	-15.7	-6.5	7.9	32.7
Clarinet	Clarinet SMC23	17.4	29.8	40.2	43.9	60.3
Clarinet	Clarinet SMC6	4.2	23.6	33.2	42.8	59.6
Clarinet	Clarinet SMC12	6.1	35.9	42.4	49.1	62.2
Clarinet	Ideal impulse	2.4	18.5	35.9	45.1	67.8
Clarinet	Trumpet	-20.3	9.7	23	34.7	62.9
Clarinet	Trumpet SMC23	14.4	25.4	33.9	42.3	59.4
Clarinet	Violin	-23.8	4.1	18.5	32	63.2
Clarinet	Violin SMC23	-0.7	20.6	28.8	36.8	50.8
Clarinet SMC23	Clarinet SMC23	-12.9	-2.4	0.2	1.6	11.1
Clarinet SMC23	Clarinet SMC6	-4	-0.5	1	4.3	6.3
Clarinet SMC23	Ideal impulse	-17.4	-7.3	-0.5	4.7	12.4
Clarinet SMC23	Trumpet	-28.3	-16.2	-7.1	-0.4	13.8
Clarinet SMC23	Violin	-41.1	-22.3	-12	-4	12.1
Clarinet SMC23	Violin SMC23	-12	-2.3	0.7	4.5	12.3
Clarinet SMC23	Violin SMC6	-9.9	-5.2	-1.7	4.8	7.1
Clarinet SMC6	Clarinet SMC6	-6.3	-1.6	-0.4	1	4
Clarinet SMC6	Ideal impulse	-8.8	-2.4	0.6	2.9	8.4
Clarinet SMC6	Violin SMC23	-16.2	-8.8	-5.6	-2.6	3
Clarinet SMC6	Violin SMC6	-7.6	-4.4	-0.5	0.4	4.3
Snare	Snare	-17.9	-4.3	-0.3	0.2	17
Snare	Snare SMC3	-9.8	10.4	15.2	25.2	36.1
Snare	Ideal impulse	-25.8	7.6	14	18.3	41.3
Snare SMC3	Snare SMC3	-1.7	-0.2	0.1	0.5	3.7
Snare SMC3	Ideal impulse	-7.3	-0.8	-0.1	0.7	6.3
Ideal impulse	Ideal impulse	-2.5	-0.6	-0.2	0.2	1.7
Ideal impulse	Maurit. SMC12	-6.8	-0.4	0.2	2.4	10.1
Ideal impulse	Trumpet	-51.8	-25.5	-8.9	9.5	23.7
Ideal impulse	Violin	-67.2	-28.3	-14	-1.7	30.9
Ideal impulse	Violin SMC23	-14.6	-6.3	-3.2	0.7	15.8
Ideal impulse	Violin SMC6	-7.5	-0.6	0.7	1.8	13.5
Maurit. SMC12	Maurit. SMC12	-1.3	-0.6	-0	0.6	3.6
Trumpet	Trumpet	-30.9	-13.8	-0.3	18	46.3
Trumpet	Trumpet SMC23	-14.1	3	8.2	16.1	28.1
Trumpet	Trumpet SMC6	-20.7	-0.4	6.6	17.6	48.1
Trumpet	Trumpet SMC12	-7.3	3.5	8.7	15.9	26
Trumpet	Violin	-84.6	-20	-4.3	8.4	78.5
Trumpet	Violin SMC23	-22.2	-0.7	6.2	18.4	40.8
Trumpet SMC23	Violin	-55.6	-20.1	-12.2	-1.9	13
Violin	Violin	-43.9	-14.6	-7	8	46.3
Violin	Violin SMC23	-2.5	5.7	11.1	23.7	50.7
Violin	Violin SMC6	-13.6	1.3	10.4	20.6	54.3
Violin	Violin SMC12	-3.7	6.7	14.1	18.3	37.9
Violin SMC23	Violin SMC23	-8.3	-1.3	0.5	2.5	9.5
Violin SMC23	Violin SMC6	-3.5	-0.5	0.4	3.6	9
Violin SMC6	Violin SMC6	-1.5	-0.4	-0.2	0.2	3.7

Table 11: Quartiles (in ms) of every distribution from this experiment. Positive values indicate starting Sound B after Sound A.

Table 11 displays the quartiles for all 48 pairs of sounds compared in this experiment. The 0% and 100% values are the furthest outliers in each direction. There is a sound example corresponding to each of these pairs of sounds that plays the two sounds separated by a delay time equal to each of the five quartiles. Each example contains three repetitions of each delay time, then a brief pause before the next delay time. The five delay times appear in the same order as in the table, in other words, the first delay time has Sound A earliest relative to Sound B. Each sound example's name is of the form *quart-Clarinet.vs.Violin*, in other words "*quart-*" followed by one sound's name followed by "*.vs.*" followed by the other sound's name.

4.3.8 Synchronizing Two Instances of the Same Sound

There were twelve sounds that subjects aligned against copies of themselves. As described in Section 3.5.3 (page 50), we expect the mean result to be zero and the distribution to be symmetric in each case. Table 12 shows how well the measured data fits these predictions, showing the number of trials, mean, standard deviation, Z-score[119] of zero, and skewness for each of the twelve sounds. Here the sign convention (for the mean and Z-score of zero) is that the positive direction means the movable copy of the sound ended up starting later than the fixed copy of the sound.

Sound	***N***	***Mean (ms)***	***STD (ms)***	***Z-score of zero***	***Skewness***
Clarinet SMC23	27	-0.2755	5.62	-0.049	-0.06184
Violin SMC6	20	0.1043	1.174	0.089	1.633
Mauritania SMC12	28	0.0988	1.039	0.095	1.438
Trumpet	37	1.88	19.7	0.095	0.4823
Violin SMC23	24	0.48	3.862	0.12	-0.07459
Snare	40	-1.071	7.215	-0.15	0.08374
Violin	36	-3.356	19.3	-0.17	0.3948
Snare SMC3	33	0.1752	0.9869	0.18	1.125
Clarinet	45	-3.207	15.76	-0.2	0.5119
Clarinet SMC6	25	-0.537	2.617	-0.21	-0.6314
Brazil SMC23	24	0.2853	1.03	0.28	0.06315
Ideal impulse	91	-0.24	0.7092	-0.34	-0.5859

Table 12: Summary statistics for trials aligning two copies of the same sound.

N is number of trials. Z-score of zero is the number of standard deviations away from the mean that zero (i.e., perfect synchrony of physical onsets) lies. Since the mean should be zero when aligning two instances of the same sound, we expect the Z-score of zero to be low. Skewness measures asymmetry and is zero for a perfectly symmetric distribution.

[119] A data point's Z-score is its number of standard deviations away from the mean. The Z-score of zero is therefore just the mean divided by the standard deviation.

4.3.9 Pairwise Variance Versus Variance Against Self

We would expect that aligning a sound against a second copy of itself would be the easiest task from the point of view of the auditory streaming effects discussed in Section 3.6 (page 61). In addition, some subjects might attend to spectral comb filtering effects. So we would expect the variance of sound A's intrinsic PAT-pdf to be not much less than half the measured variance from trials aligning sound A against sound A. Also, we would expect in general that the variance from trials aligning sound A against sound B should be at least the sum of the variance of A's and B's intrinsic PAT-pdfs:

$$var(D_{A,B}) \geq 0.5\ (var(D_{A,A})+var(D_{B,B}))$$

Does this inequality hold for the results of this experiment? As Table 13 shows, the inequality holds for 24 of the 29 pairs of sounds.[120] Note that Clarinet SMC23 was involved in four of the five cases (the first five rows of the table) where the inequality did not hold. Removing just one value from each extreme of the results of Clarinet SMC23 against itself would reduce its variance to 22.24 ms^2, which would make the inequality hold within a few percent for all but the comparison against the Trumpet, which is discussed below. Perhaps there is something special about the Clarinet SMC23 sound that makes subjects especially consistent in their results aligning it against other sounds.

In general our prediction is accurate (in other words, "extra" variance is positive) in most cases, and we can interpret the "extra" variance as the specific difficulty of aligning each particular pair of sounds. Note in particular that the ideal was a member of all seven sound pairs with the highest percentage of extra variance (the bottom 7 rows of Table 13), which makes sense because the ideal impulse is spectrally furthest from all of the other sounds.

Surprisingly, however, the next highest percentage penalty is for the Snare sound against its spectrally matched click. This click (like most clicks) has extremely low variance against itself, but the variance of the click against the snare drum is almost twice the variance of the snare drum against itself. Perhaps the click is not sufficiently spectrally matched. (Snare SMC3 is the shortest-duration SMC in this study.) Another difference is that the Snare sound exhibits pronounced comb filtering when aligned against itself (as demonstrated in Sound Example *quart-Snare.vs.Snare*), while there is no such effect when aligning the Snare to the Snare SMC3 (as demonstrated in Sound Example *Snare.vs.Snare_SMC3*).

[120] These are all 29 pairs of single-event sounds that subjects synchronized with each other and also with themselves.

Sound A	***var(AA)***	***Sound B***	***var(BB)***	***var(AB)***	***Extra var.***	***% Extra var***
Clarinet SMC23	31.59	Trumpet	388.2	107.3	-102.6	-48.9
Clarinet SMC23	31.59	Clarinet SMC6	6.85	10.96	-8.262	-43.0
Clarinet	248.4	Clarinet SMC23	31.59	123.8	-16.2	-11.6
Clarinet SMC23	31.59	Violin	372.6	186.5	-15.59	-7.7
Violin	372.6	Violin SMC23	14.91	188.8	-4.912	-2.5
Trumpet	388.2	Violin SMC23	14.91	209.4	7.847	3.9
Violin SMC23	14.91	Violin SMC6	1.379	8.556	0.4104	5.0
Clarinet	248.4	Violin	372.6	394.4	83.97	27.0
Clarinet	248.4	Trumpet	388.2	406.9	88.59	27.8
Violin	372.6	Violin SMC6	1.379	257.4	70.37	37.6
Clarinet	248.4	Clarinet SMC6	6.85	178.7	51.07	40.0
Clarinet	248.4	Violin SMC23	14.91	190	58.34	44.3
Clarinet SMC23	31.59	Violin SMC6	1.379	29.73	13.24	80.3
Clarinet SMC23	31.59	Violin SMC23	14.91	49.78	26.53	114.1
Clarinet SMC6	6.85	Violin SMC23	14.91	23.61	12.73	117.0
Ideal impulse	0.503	Trumpet	388.2	425.2	230.9	118.8
Trumpet	388.2	Violin	372.6	832.6	452.3	118.9
Clarinet	248.4	Ideal impulse	0.503	333.7	209.3	168.2
Ideal impulse	0.503	Violin	372.6	536.2	349.6	187.4
Clarinet SMC6	6.85	Violin SMC6	1.379	12.56	8.449	205.3
Clarinet SMC23	31.59	Ideal impulse	0.503	68.89	52.84	329.3
Snare	52.05	Snare SMC3	0.974	119.8	93.24	351.7
Brazil SMC23	1.062	Ideal impulse	0.503	3.815	3.033	387.6
Clarinet SMC6	6.85	Ideal impulse	0.503	18.79	15.11	410.9
Ideal impulse	0.503	Violin SMC23	14.91	47.69	39.98	518.7
Snare	52.05	Ideal impulse	0.503	164.3	138	525.1
Snare SMC3	0.974	Ideal impulse	0.503	4.949	4.211	570.2
Ideal impulse	0.503	Maurit. SMC12	1.079	10.71	9.918	1253.7
Ideal impulse	0.503	Violin SMC6	1.379	15.48	14.54	1545.3

Table 13: Comparison of variance for each pair of sounds against the variances for each sound against itself.

The variance of Sound A against itself ("var(AA)") and the variance of Sound B against itself ("var(BB") give us estimates for A's and B's intrinsic variance, so the mean of these two values is our predicted lower bound on the variance of aligning the two sounds against each other. The actual sample variance for aligning the two sounds is "var(AB)", so the "extra" variance is var(AB)-mean(var(AA), var(BB)). The rows are sorted by the last column, "Percent extra variance", which expresses the "extra" variance as a percentage of the predicted variance. High values (at the bottom of the table) indicate pairs of sounds that are relatively difficult to align against each other. Negative values indicate unexpected situations where the results aligning the two different sounds are "too accurate". All variance is in units of (ms)^2.

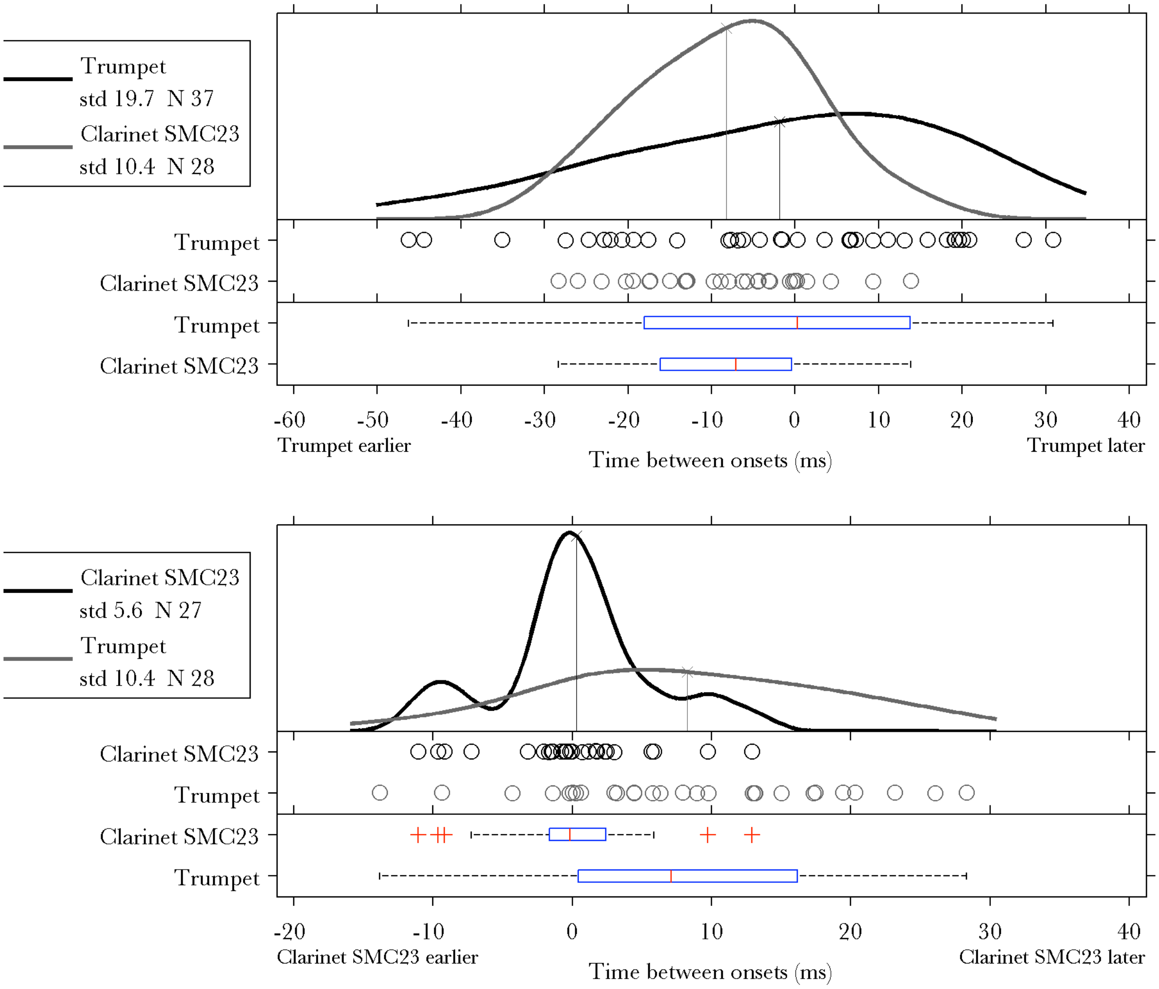


Figure 37: Examination of trials with Trumpet and Clarinet SMC23.

The top plot shows the distributions for both sounds compared to Trumpet; the bottom shows the distributions for both sounds compared to Clarinet SMC23. The variance of aligning these sounds with each other is much lower than the mean of the variances of aligning each sound with itself, as shown in Table 13.

The biggest failure of our prediction is for Trumpet versus Clarinet SMC23, the top row of Table 13. Figure 37 examines the distributions for these sounds against themselves and against each other. One would expect that subjects would be more consistent aligning the Trumpet with the click than the Trumpet against itself, and the upper plot shows that this is indeed the case. Part of the problem could be the five outliers from trials aligning two copies of Clarinet SMC23 (shown as "+" signs in the bottom box plot), but keep in mind that even if the variance of aligning Clarinet SMC23 against itself were zero, the variance of the click against the Trumpet is less than half of the variance of aligning the Trumpet against itself.

4.3.10 Check of Predicted Mean for each Trio of Sounds

Sound A	***Sound B***	***Sound C***	***AB***	***BC***	***AC***	***AB+BC-AC***
Clarinet	Trumpet	Violin	22.4	-3.77	18.7	-0.0451
Ideal impulse	Trumpet	Violin	-10.4	-3.77	-14.2	0.0855
Trumpet	Tpt. SMC23	Violin	9.52	-13.1	-3.77	0.181
Ideal impulse	Trumpet	ViolinSMC23	-10.4	8.01	-2.19	-0.188
Clar. SMC23	Clar. SMC6	Violin SMC6	1.54	-1.73	-0.407	0.218
Clar. SMC23	Ideal impulse	Violin SMC6	-1.15	0.985	-0.407	0.245
Clarinet	Ideal impulse	Violin	33.4	-14.2	18.7	0.449
Clarinet	Ideal impulse	Trumpet	33.4	-10.4	22.4	0.58
Clar. SMC23	Violin	Violin SMC6	-13.2	13.6	-0.407	0.858
Clar. SMC23	Trumpet	Violin	-8.28	-3.77	-13.2	1.12
Clar. SMC23	Trumpet	Violin SMC23	-8.28	8.01	0.887	-1.16
Clarinet	Clar. SMC6	Ideal impulse	31.8	0.329	33.4	-1.28
Clar. SMC23	Violin	Violin SMC23	-13.2	15.4	0.887	1.31
Ideal impulse	Violin	Violin SMC6	-14.2	13.6	0.985	-1.61
Ideal impulse	Violin SMC23	Violin SMC6	-2.19	1.25	0.985	-1.93
Clarinet	Trumpet	Violin SMC23	22.4	8.01	28.3	2.09
Clar. SMC23	Ideal impulse	Violin	-1.15	-14.2	-13.2	-2.23
Clarinet	Clar. SMC6	Violin SMC23	31.8	-5.76	28.3	-2.31
Clarinet	Trumpet	Tpt. SMC23	22.4	9.52	34.3	-2.35
Clarinet	Tpt. SMC23	Violin	34.3	-13.1	18.7	2.49
Clar. SMC23	Violin SMC23	Violin SMC6	0.887	1.25	-0.407	2.54
Clar. SMC6	Violin SMC23	Violin SMC6	-5.76	1.25	-1.73	-2.78
Clarinet	Ideal impulse	ViolinSMC23	33.4	-2.19	28.3	2.86
Clarinet	Clar. SMC23	Ideal impulse	37.5	-1.15	33.4	2.96
Violin	Violin SMC23	Violin SMC6	15.4	1.25	13.6	2.99
Clar. SMC23	Clar. SMC6	Ideal impulse	1.54	0.329	-1.15	3.02
Clar. SMC6	Ideal impulse	Violin SMC6	0.329	0.985	-1.73	3.05
Clar. SMC23	Ideal impulse	Trumpet	-1.15	-10.4	-8.28	-3.26
Ideal impulse	Violin	Violin SMC23	-14.2	15.4	-2.19	3.31
Trumpet	Violin	Violin SMC23	-3.77	15.4	8.01	3.58
Snare	SnareSMC3	Ideal impulse	17.5	0.148	13.9	3.78
Clar. SMC6	Ideal impulse	Violin SMC23	0.329	-2.19	-5.76	3.9
Clar. SMC23	Ideal impulse	Violin SMC23	-1.15	-2.19	0.887	-4.23
Clar. SMC23	Clar. SMC6	Violin SMC23	1.54	-5.76	0.887	-5.1
Clarinet	Clar. SMC23	Violin	37.5	-13.2	18.7	5.63
Clarinet	Violin	Violin SMC23	18.7	15.4	28.3	5.72
Clarinet	Clar. SMC23	Trumpet	37.5	-8.28	22.4	6.8
Clarinet	Clar. SMC23	Clar. SMC6	37.5	1.54	31.8	7.26
Clarinet	Clar. SMC23	Violin SMC23	37.5	0.887	28.3	10

Table 14: Check of prediction of means for trios of sounds.

For each trio of sounds ("A", "B", and "C"), we expect that the mean result of trials aligning A and B ("AB") plus the mean result of aligning B and C ("BC") will be (approximately) equal to the mean result of trials aligning A and C ("AC"). The last column shows AB+BC-AC, which would be zero if the prediction were exactly correct; the table is sorted in increasing order of absolute value of this column (i.e., in decreasing order of the correctness of the prediction) All numeric data are in milliseconds.

This section considers all trios of sounds A, B, and C for which this experiment tested all three pairwise comparisons (A vs. B, A vs. C, and B vs. C). Section 3.5.1 (page 49) suggests that the mean result of trials comparing a pair of sounds will be the difference in means of the intrinsic PAT-pdfs of the two sounds, and therefore predicts the following:

$$mean(D_{A,C}) = \mu_A\text{-}\mu_C = \mu_A\text{-}\mu_C+(\mu_B\text{-}\mu_B) = (\mu_A\text{-}\mu_B)+(\mu_B\text{-}\mu_C) = mean(D_{A,B})+mean(D_{B,C})$$

Table 14 shows how closely the experimental results match this prediction.

4.3.11 What is the Mean of the Intrinsic PAT-pdf for the Ideal Impulse?

The standard deviation of all trials aligning the ideal impulse against itself is only 0.71 ms, so the ideal impulse must have a very narrow PAT-pdf.[121] The ideal impulse has the shortest possible duration of any digital signal, just a single audio sample. (Since the experiment ran entirely at a 44.1 kHz sampling rate, the theoretical duration is about 23 microseconds, though it's likely that reconstruction filters and other aspects of subjects' audio hardware actually produced slightly longer stimuli.)

Sound	***Mean***	***Median***	***STD***	***N***
Clarinet SMC23	-1.15	-0.45	8.30	30
Mauritania SMC12	-1.09	-0.25	3.27	32
Violin SMC6	-0.99	-0.70	3.93	37
Snare SMC3	0.15	-0.09	2.22	55
Ideal impulse	0.24	0.20	0.71	182
Clarinet SMC6	0.33	0.63	4.33	19
Brazil SMC23	0.58	0.19	1.95	16
Violin SMC23	2.19	3.17	6.91	21
Trumpet	10.39	8.88	20.62	34
Snare	13.89	14.05	12.82	52
Violin	14.24	14.01	23.16	27
Clarinet	33.36	35.94	18.27	33

Table 15: Mean, median, standard deviation, and N for the results of each sound compared against the ideal impulse.

Mean, median, and standard deviation are in ms. Negative values for mean indicate that the average final alignment was for the ideal impulse to begin before the other sound.

My study compared almost every sound against the ideal impulse. (All three of the loops and 12 of the 17 single-event sounds.) If the narrow PAT-pdf of the ideal impulse were centered in time around the single audio sample of the ideal impulse, then its mean would be approximately zero.

[121] The sample variance of all trials aligning the ideal impulse against itself is 0.5030 ms^2. We assume this is about twice the variance of the ideal impulse's intrinsic PAT-pdf as described on page 51, which would make the ideal impulse's intrinsic variance about 0.25 ms^2 and its intrinsic standard deviation about 0.50 ms.

We would then expect the mean result for any trial using the ideal impulse to be positive, on the theory that any other sound's PAT should be later in absolute time than the PAT of the ideal impulse. Table 15 and Figure 38 show that this is not quite the case, that for three of the short click sounds (Clarinet SMC23, Mauritania SMC12, and Violin SMC6) the mean response corresponds to starting the ideal impulse *before* the other click.

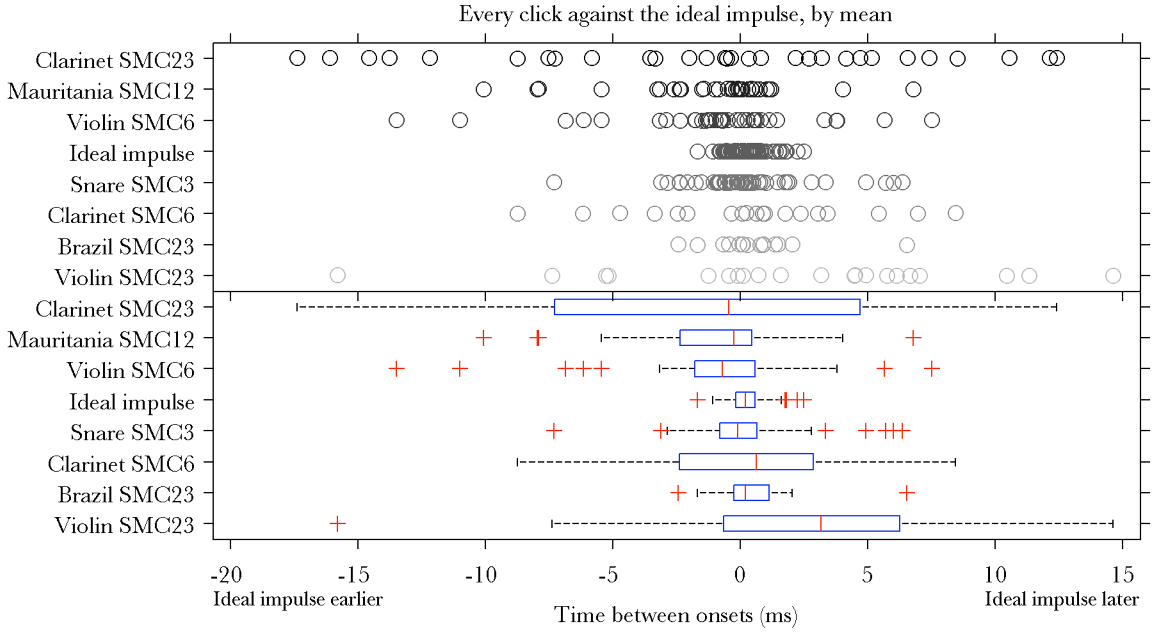


Figure 38: Results of all trials synchronizing any click sound against the ideal impulse, in order of mean.

Figure 45 (page 109) is identical to this one except that it also shows results for the non-click sounds against the ideal impulse.

Table 15 says that the mean of the intrinsic PAT-pdf for Clarinet SMC23 is 1.15 ms *before* the mean of the intrinsic PAT-pdf for the ideal impulse. So if we assume the ideal impulse's PAT is centered on the single nonzero sample, then the PAT of sound Clarinet SMC23 would come 1.15 ms before its physical onset, which should be impossible. To avoid this situation we're forced to place the center of the ideal impulse's intrinsic PAT-pdf at least 1.15 ms after its single nonzero sample.

Listening to the quartiles of the distribution for Clarinet SMC23 against Ideal Impulse (Sound Example *quart-C_SMC23.vs.Ideal*) suggests another interpretation. In nature there are many sounds that begin with a sharp percussive transient followed by a sustained or decaying steady-state portion, such as most sounds produced by plucking or striking. In nature we almost never hear a sound that starts with the sustaining portion followed by a sharp transient. When listening carefully to the outliers in the direction of starting the ideal impulse before Clarinet SMC23 and even the 25% percentile, it is possible to hear that the ideal impulse clearly comes before the longer sound. But they're close enough that it is also possible to hear a single sound event with a sharp attack (the ideal impulse) followed by the more sustaining and pitched portion (the Clarinet SMC23), which is a plausible natural sound.

4.4 Results

4.4.1 Plots of All Data

The following 23 figures (Figure 39 through Figure 61) display the results of all 1640 trials aligning pairs of single-event sounds, grouped so that each individual figure shows all results involving one particular sound. Results comparing two different sounds therefore appear in the figure corresponding to each of the two sounds. (For example, the results for comparisons of Trumpet and Violin appear in Figure 53, since they involve the Trumpet, and also Figure 60, since they involve the Violin.)

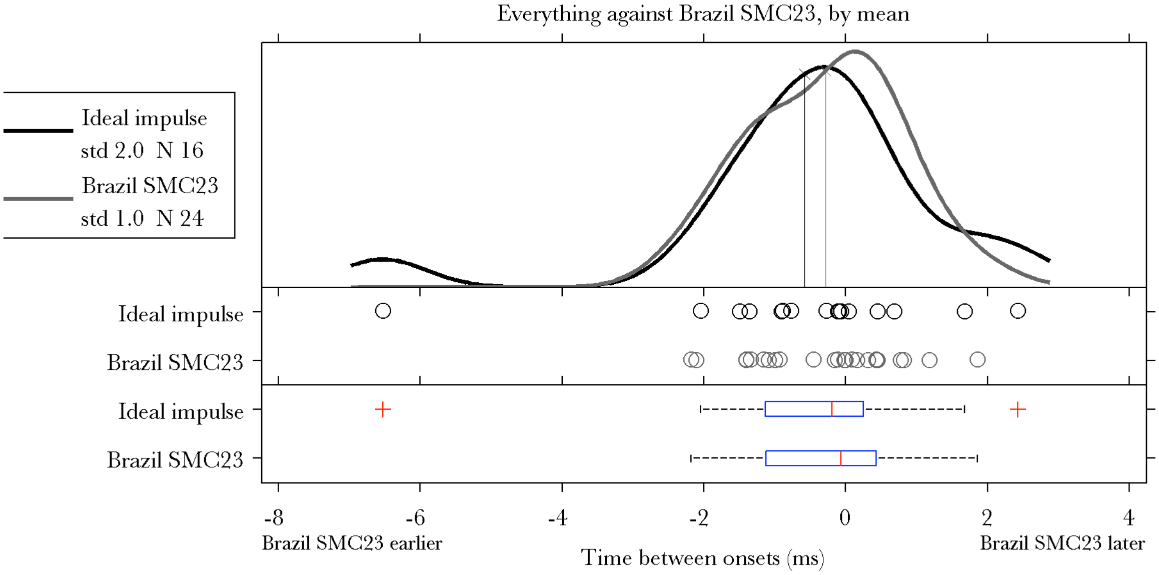


Figure 39: Results of all trials involving Brazil SMC23

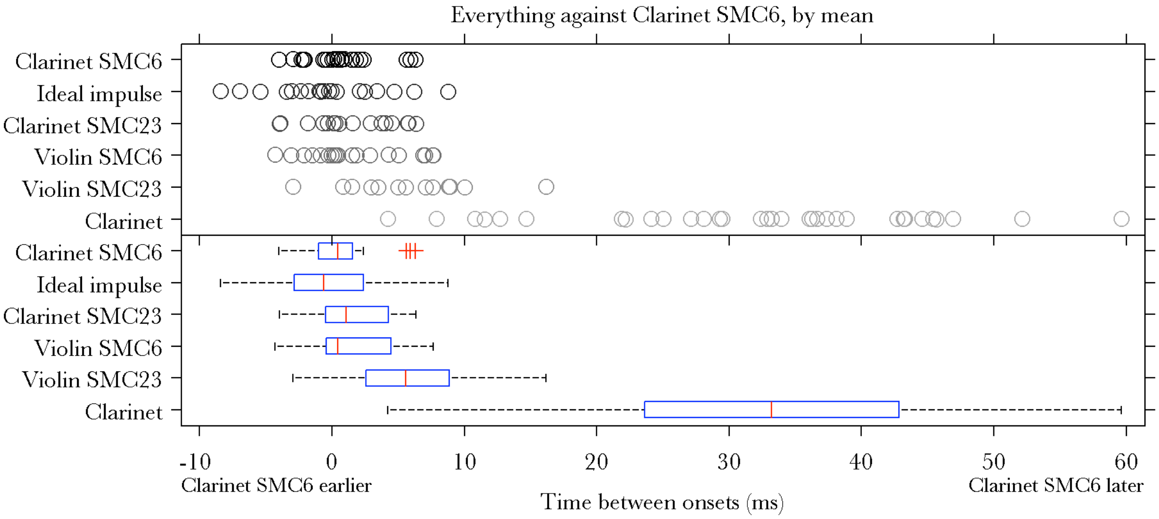


Figure 40: Results of all trials involving Clarinet SMC6

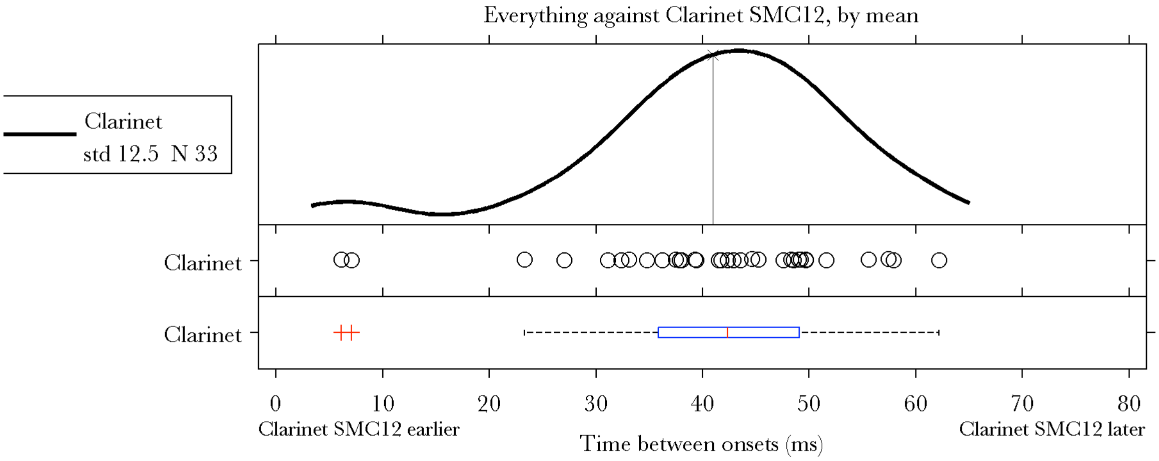


Figure 41: Results of all trials involving Clarinet SMC12

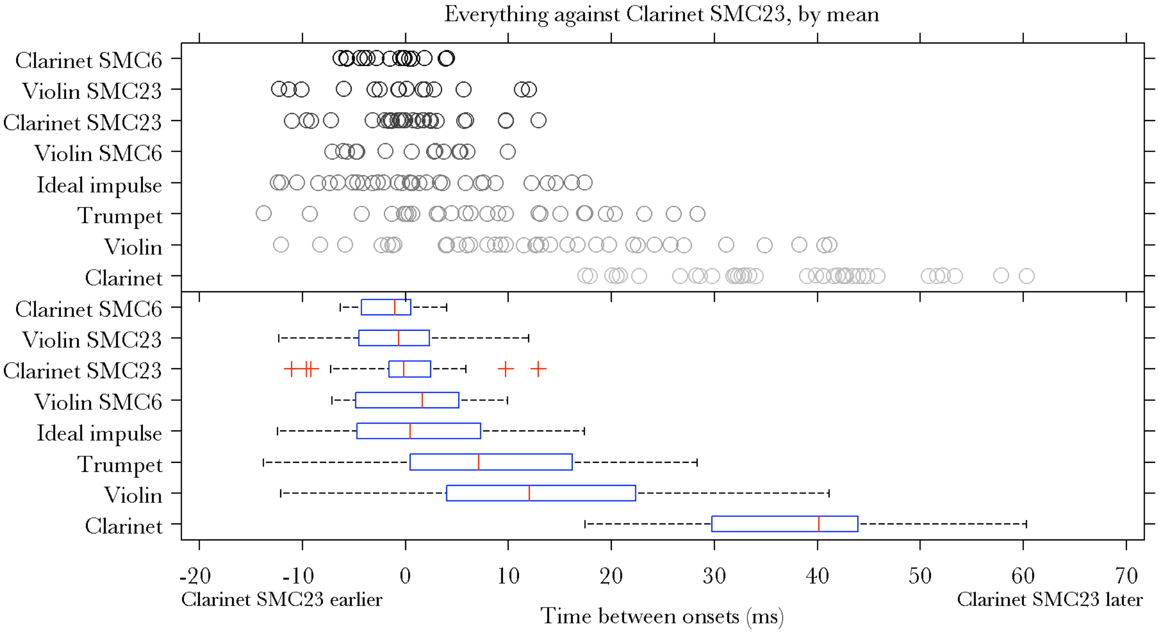


Figure 42: Results of all trials involving Clarinet SMC23

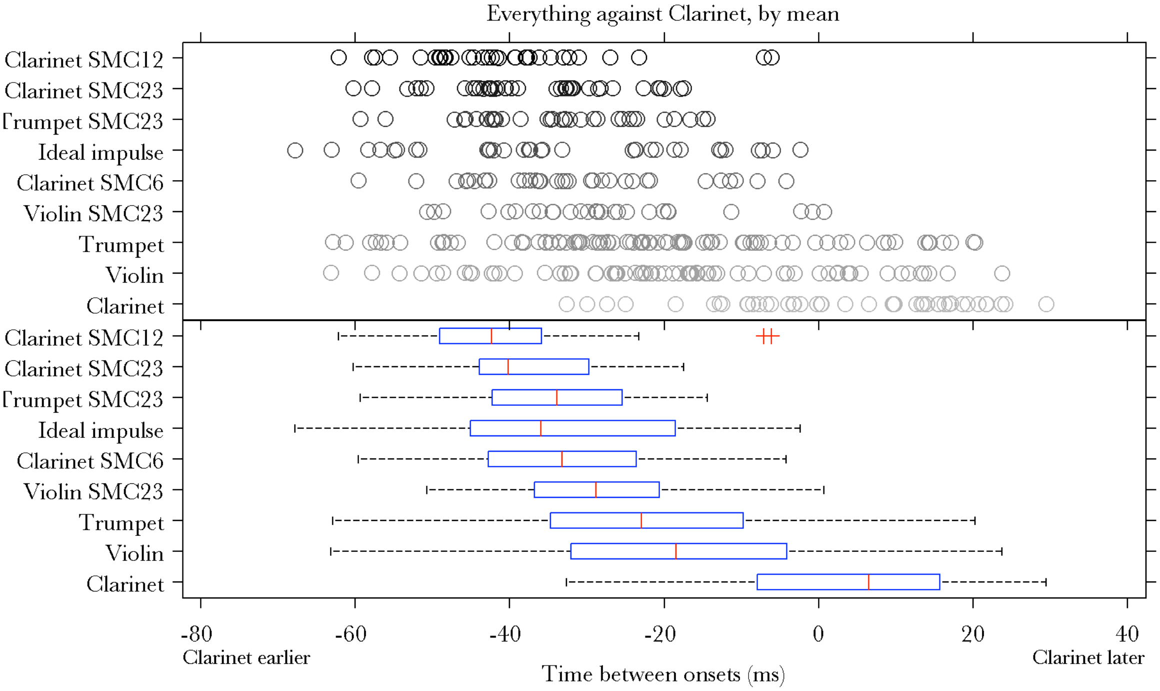


Figure 43: Results of all trials involving Clarinet, in order of mean

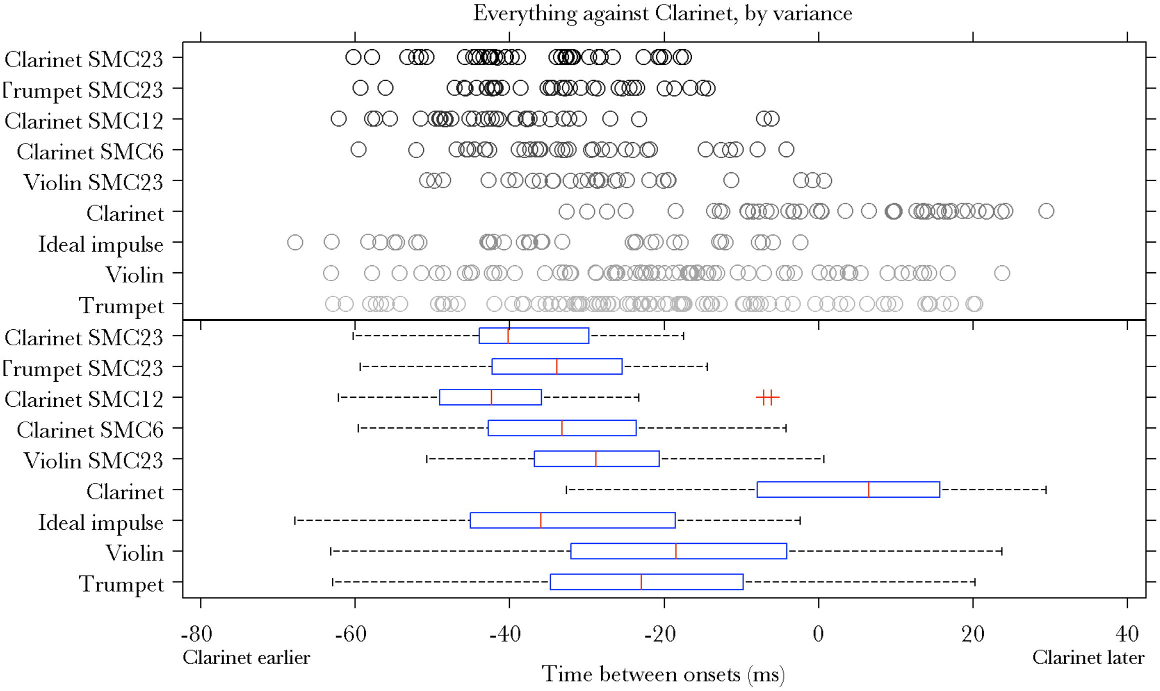


Figure 44: Results of all trials involving Clarinet, in order of variance

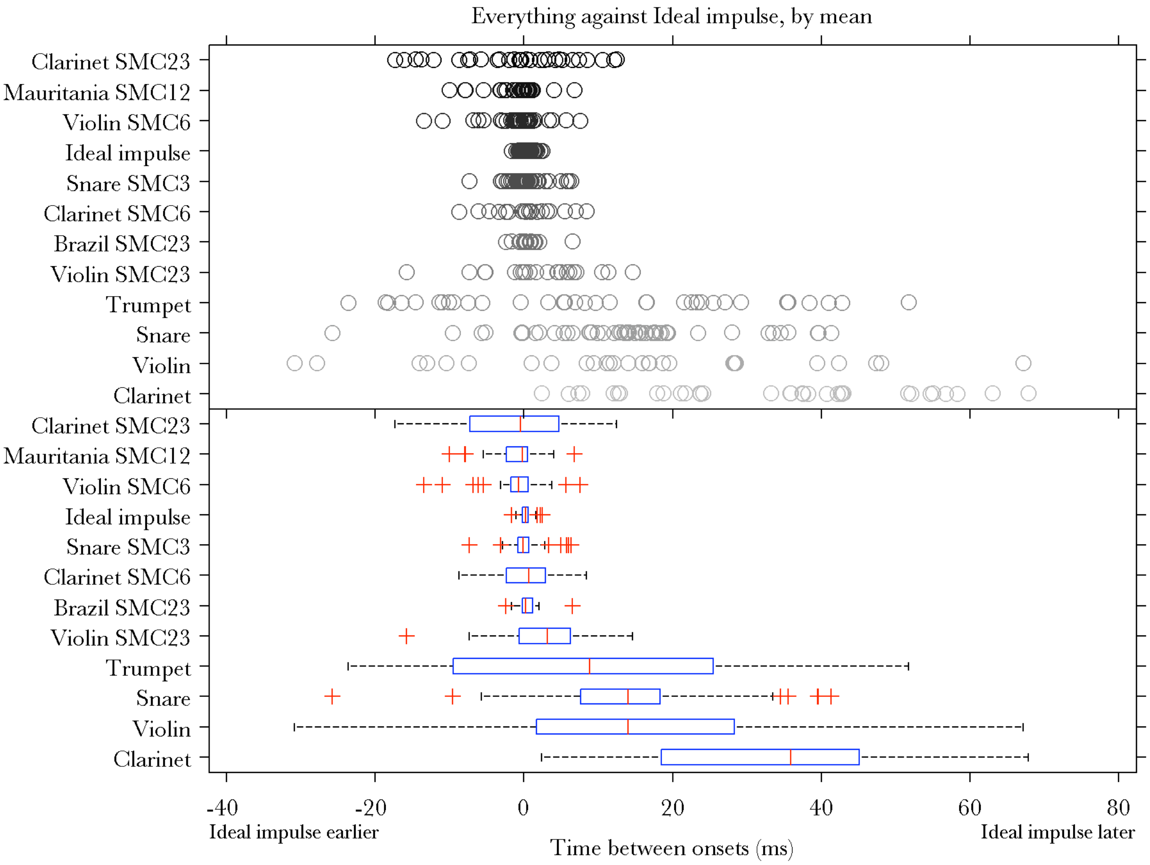


Figure 45:Results of all trials involving the Ideal Impulse, in order of mean

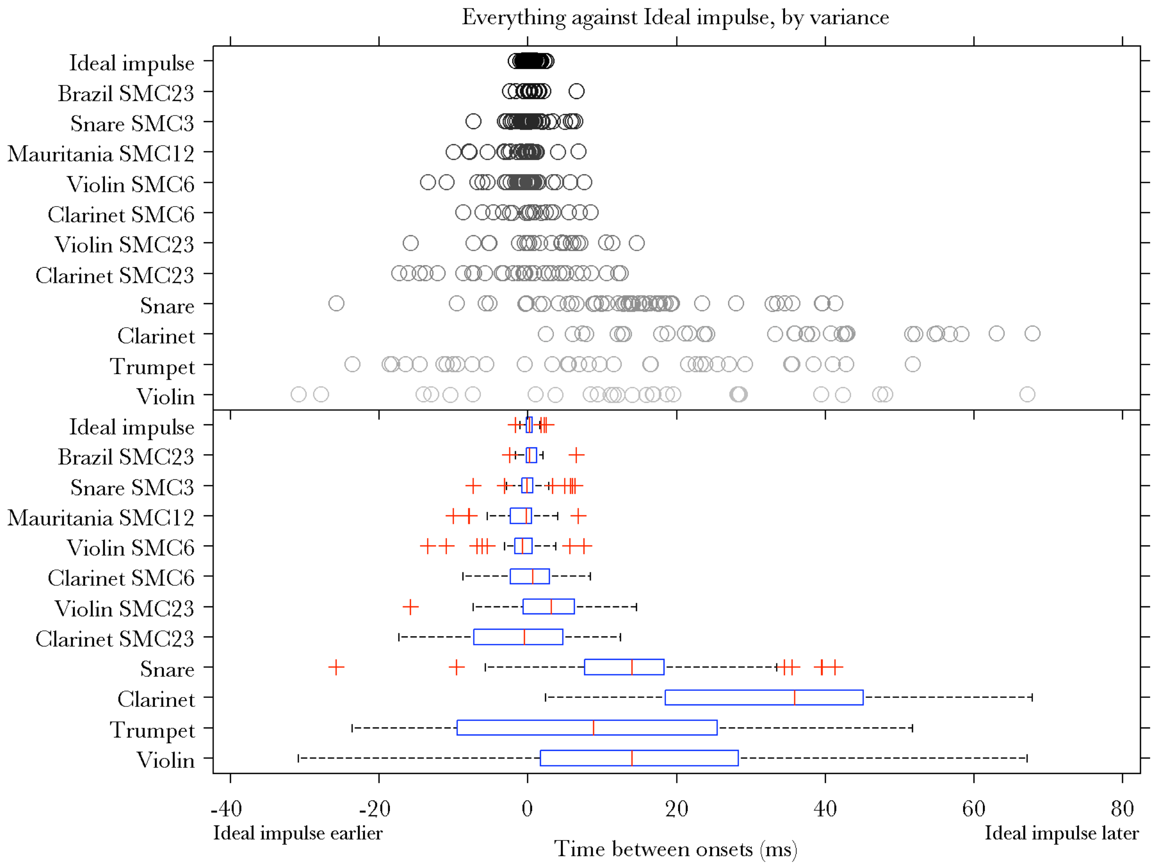


Figure 46: Results of all trials involving the Ideal Impulse, in order of variance

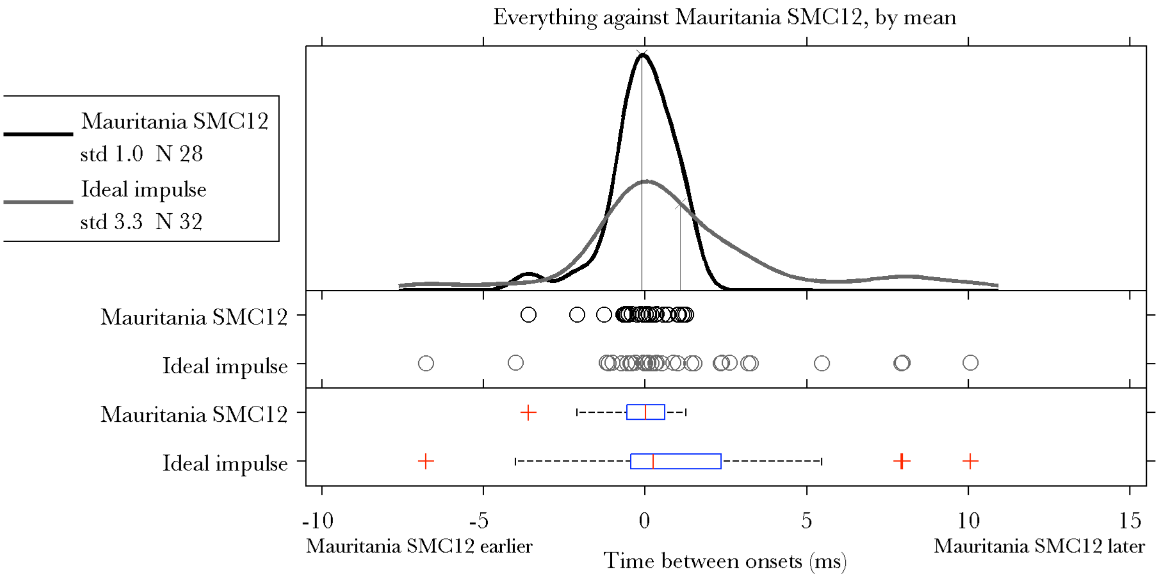


Figure 47: Results of all trials involving Mauritania SMC12

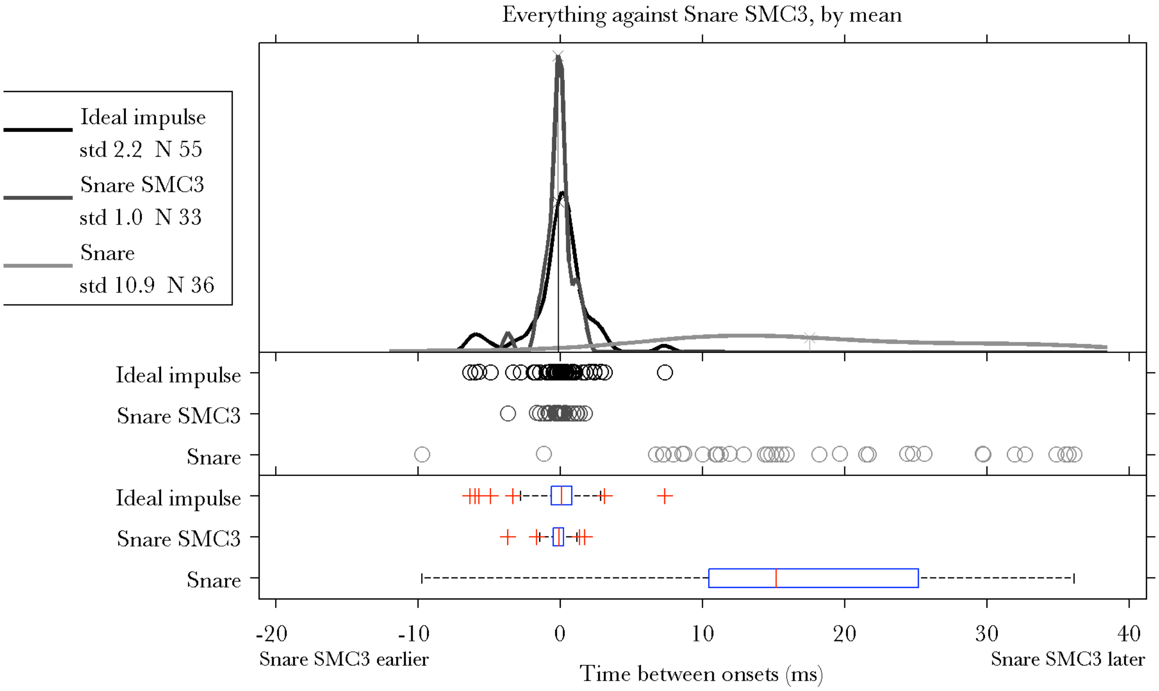


Figure 48: Results of all trials involving Snare SMC3

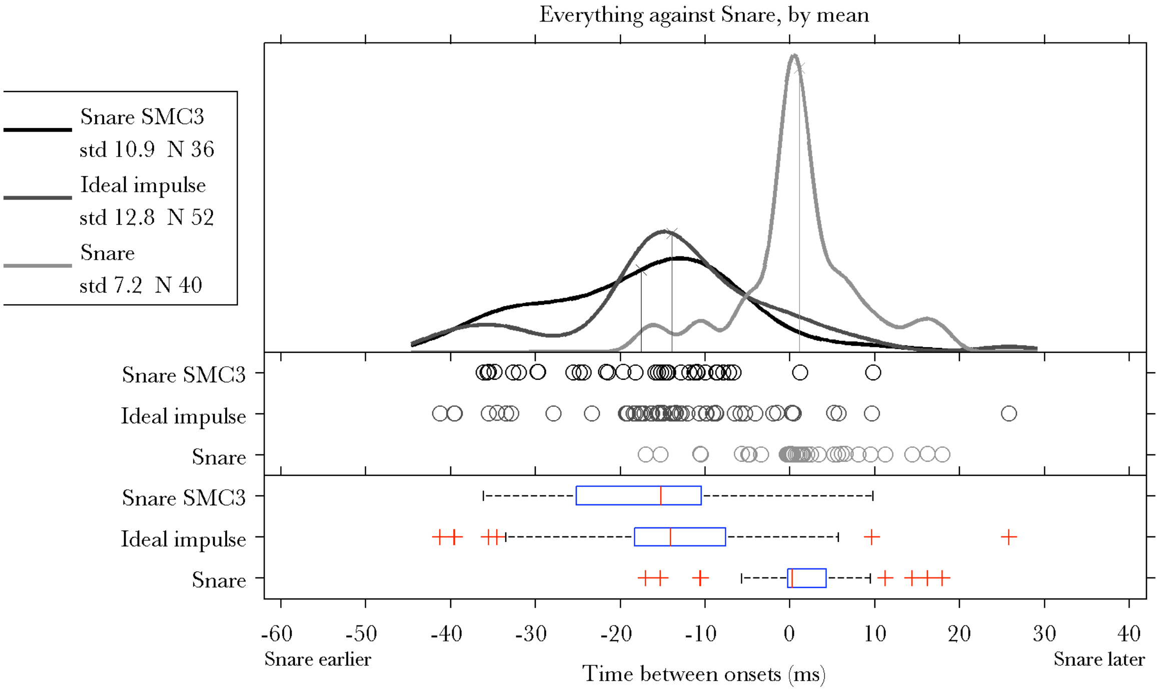


Figure 49: Results of all trials involving Snare

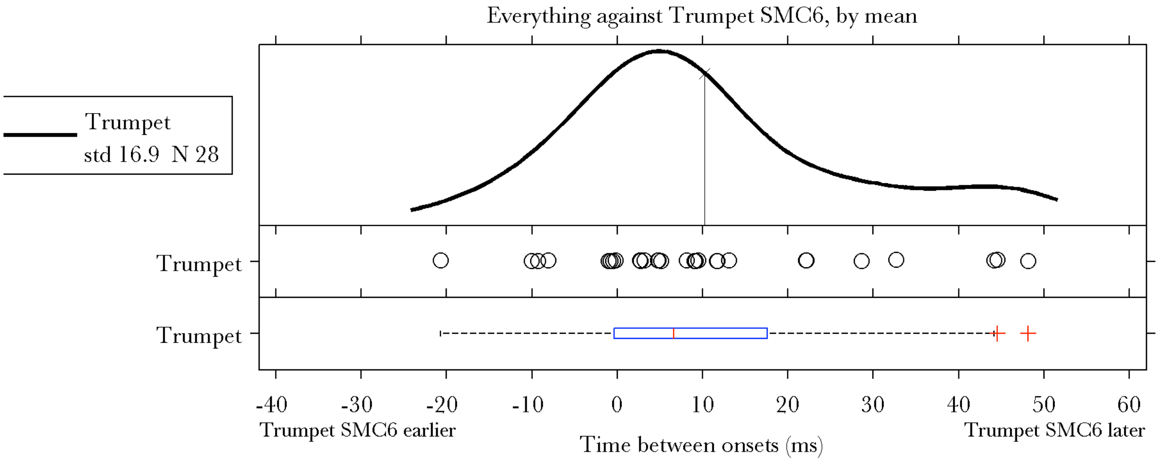


Figure 50: Results of all trials involving Trumpet SMC6

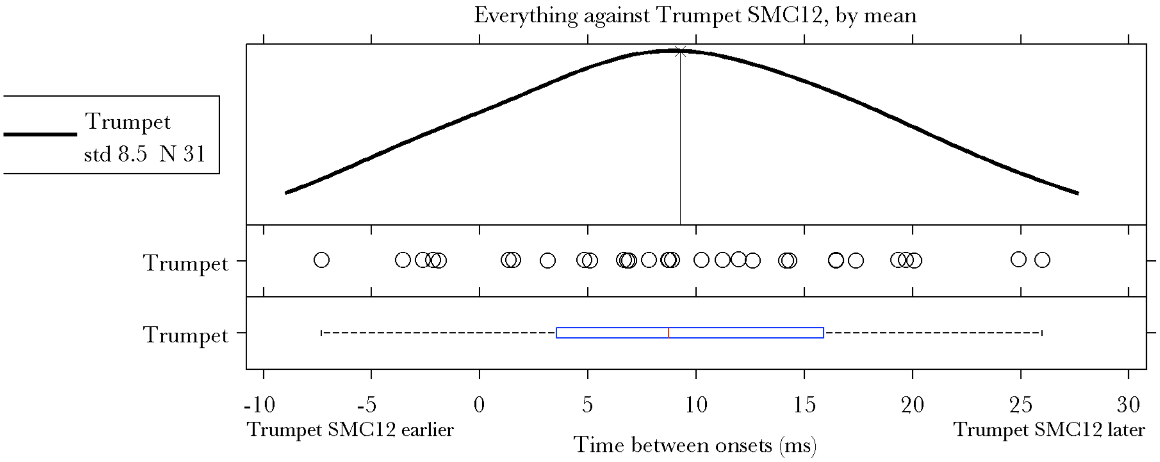


Figure 51: Results of all trials involving Trumpet SMC12

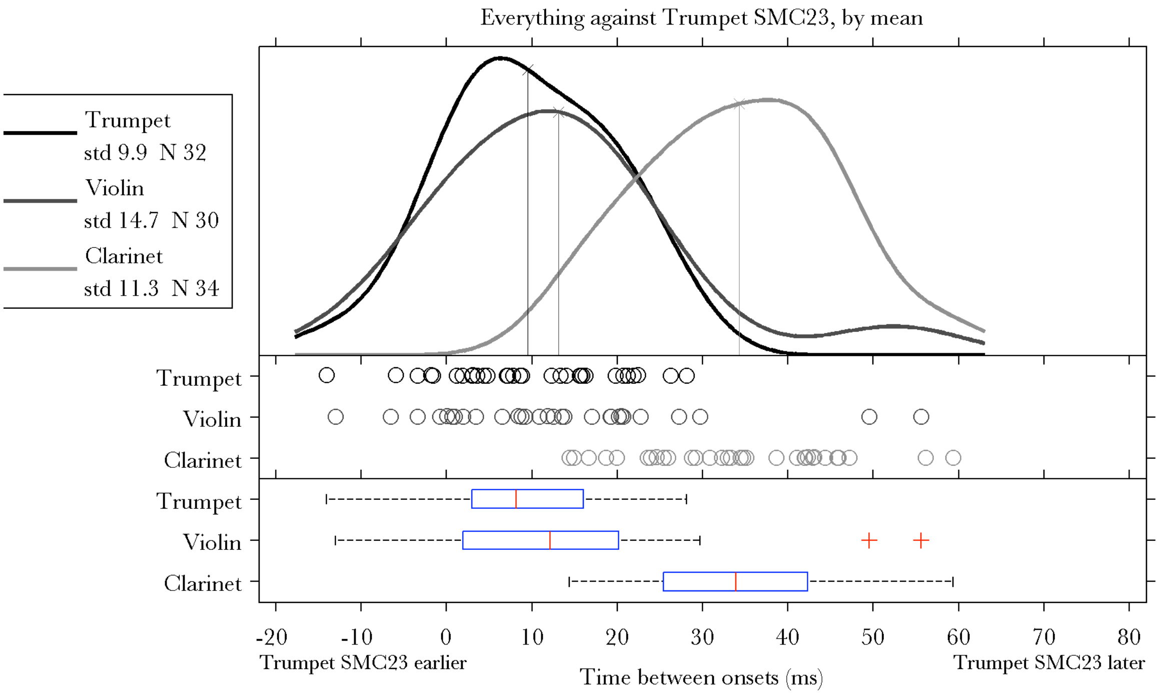


Figure 52: Results of all trials involving Trumpet SMC23

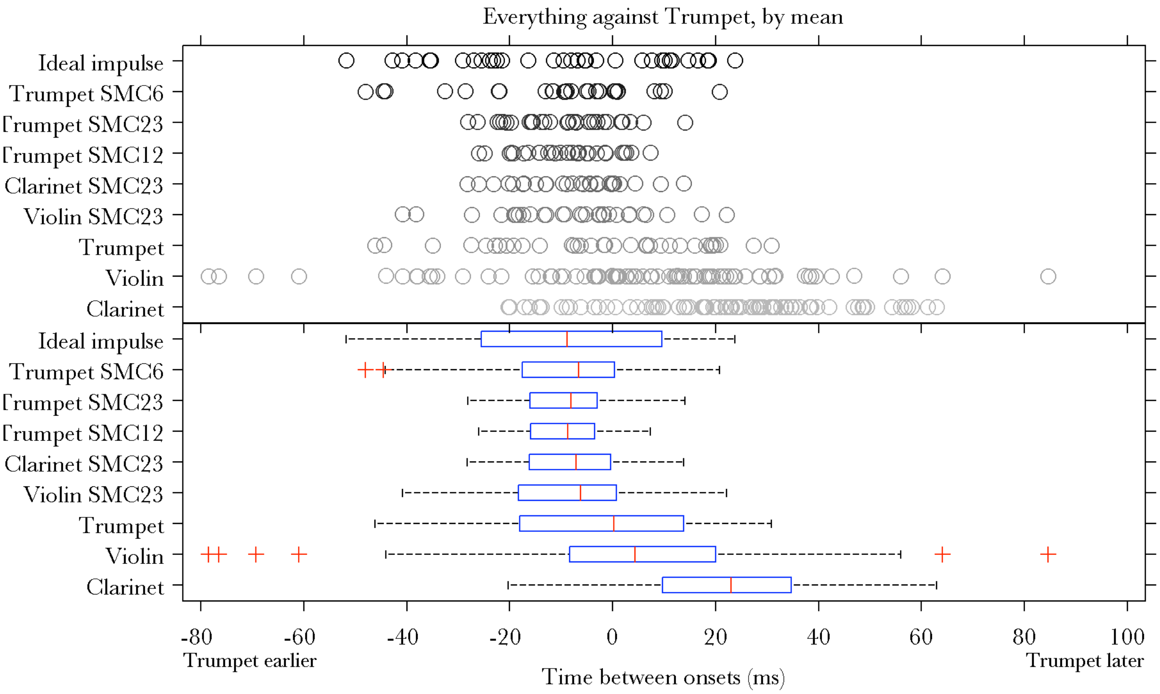


Figure 53: Results of all trials involving Trumpet, in order of mean

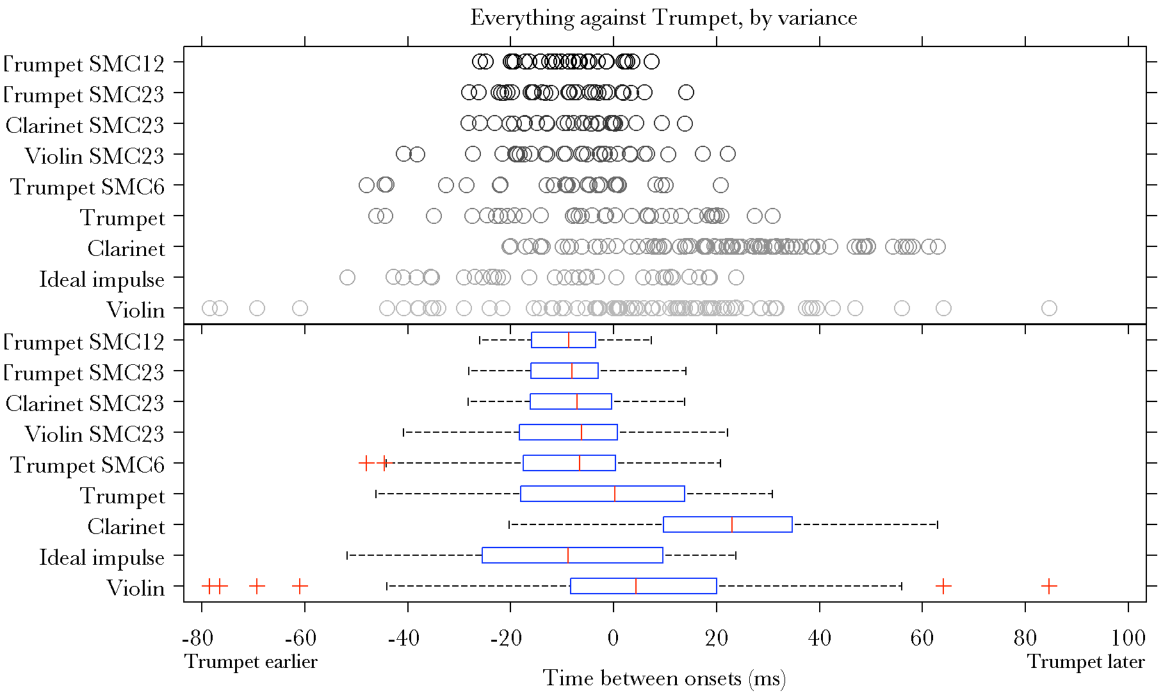


Figure 54: Results of all trials involving Trumpet, in order of variance

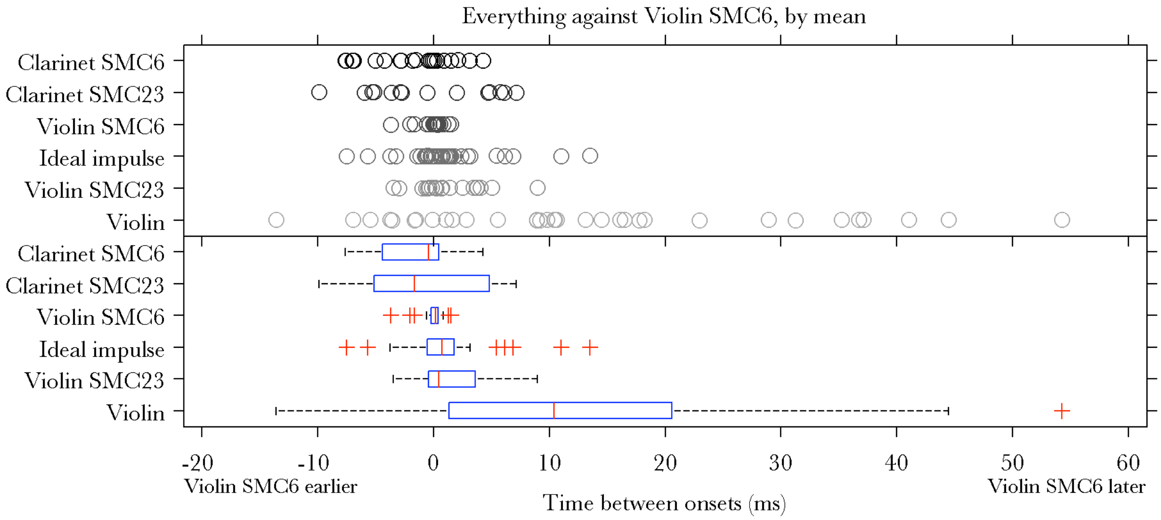


Figure 55: Results of all trials involving Violin SMC6, in order of mean

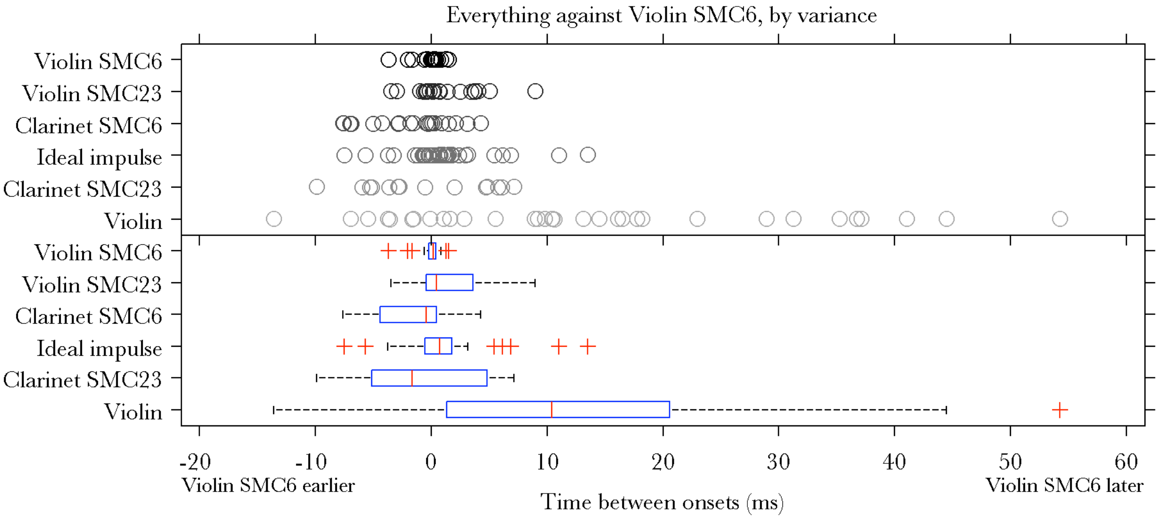


Figure 56: Results of all trials involving Violin SMC6, in order of variance

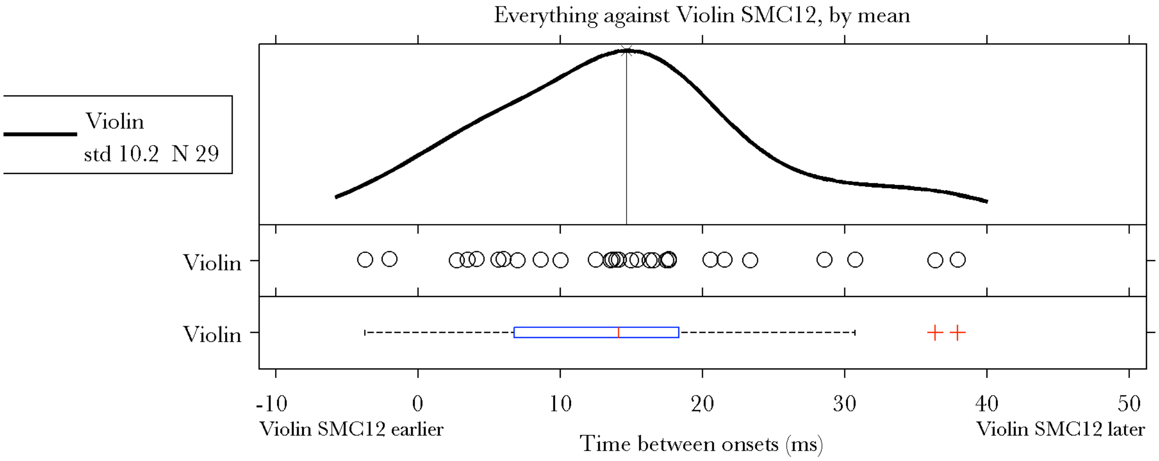


Figure 57: Results of all trials involving Violin SMC12.

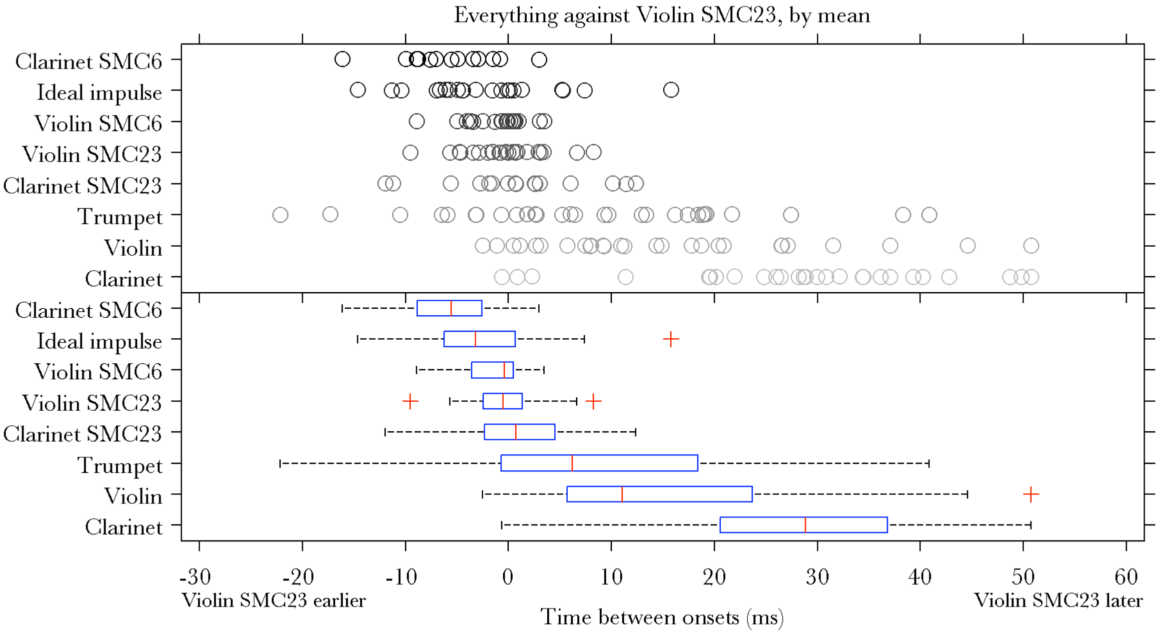


Figure 58: Results of all trials involving Violin SMC23, in order of mean

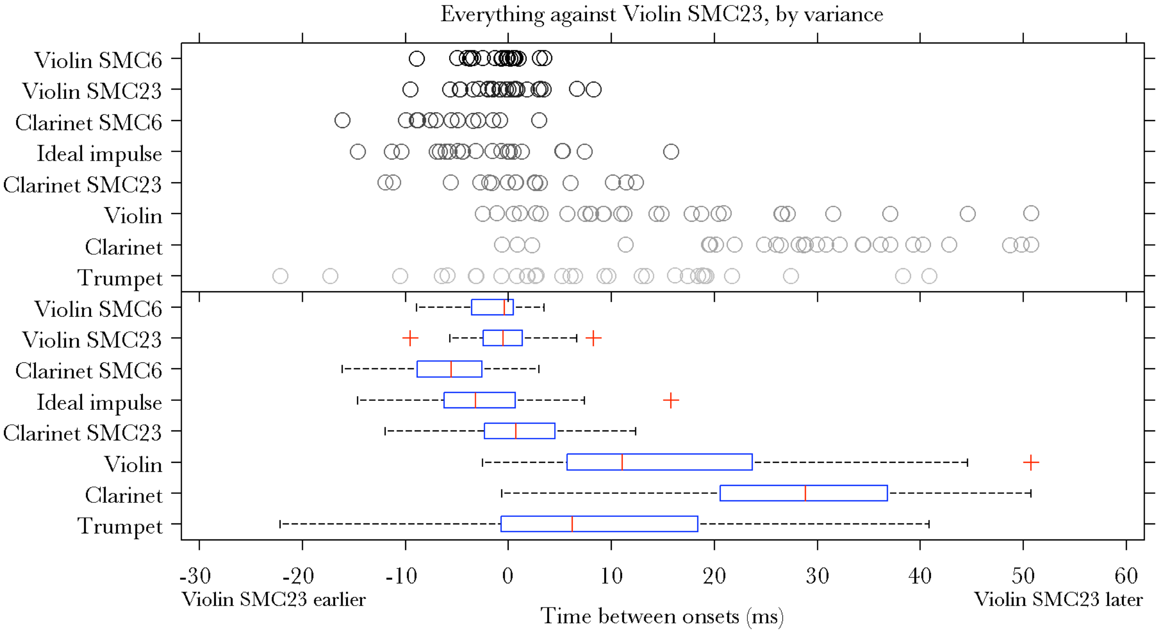


Figure 59: Results of all trials involving Violin SMC23, in order of variance

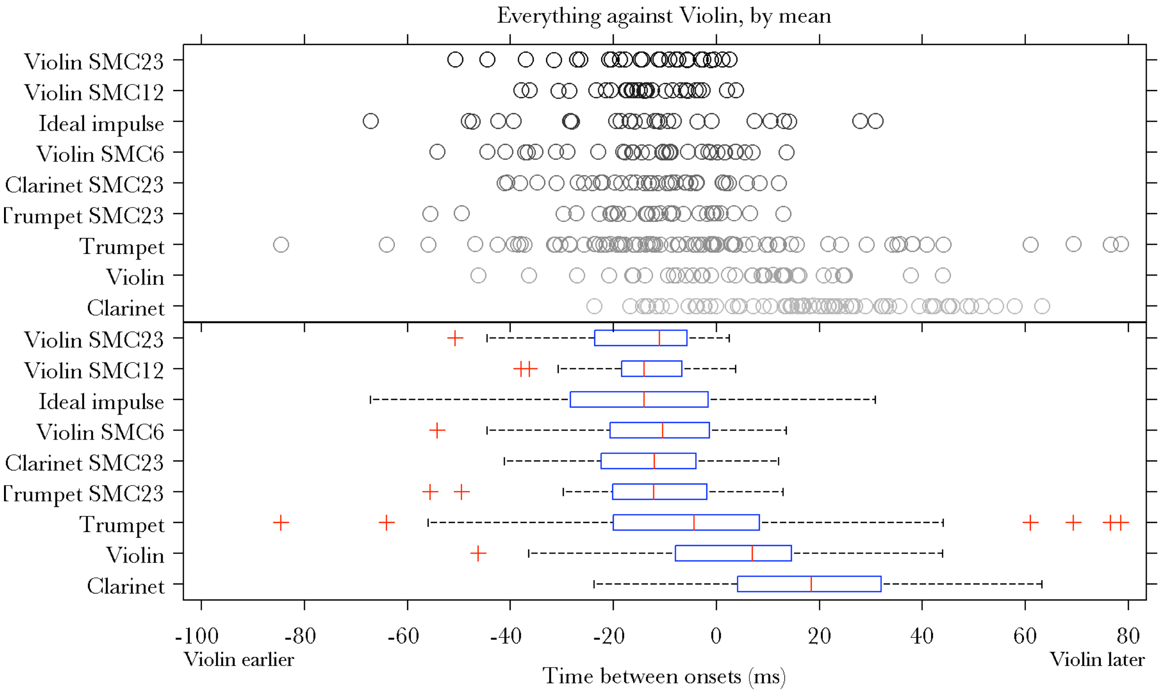


Figure 60: Results of all trials involving Violin, in order of mean

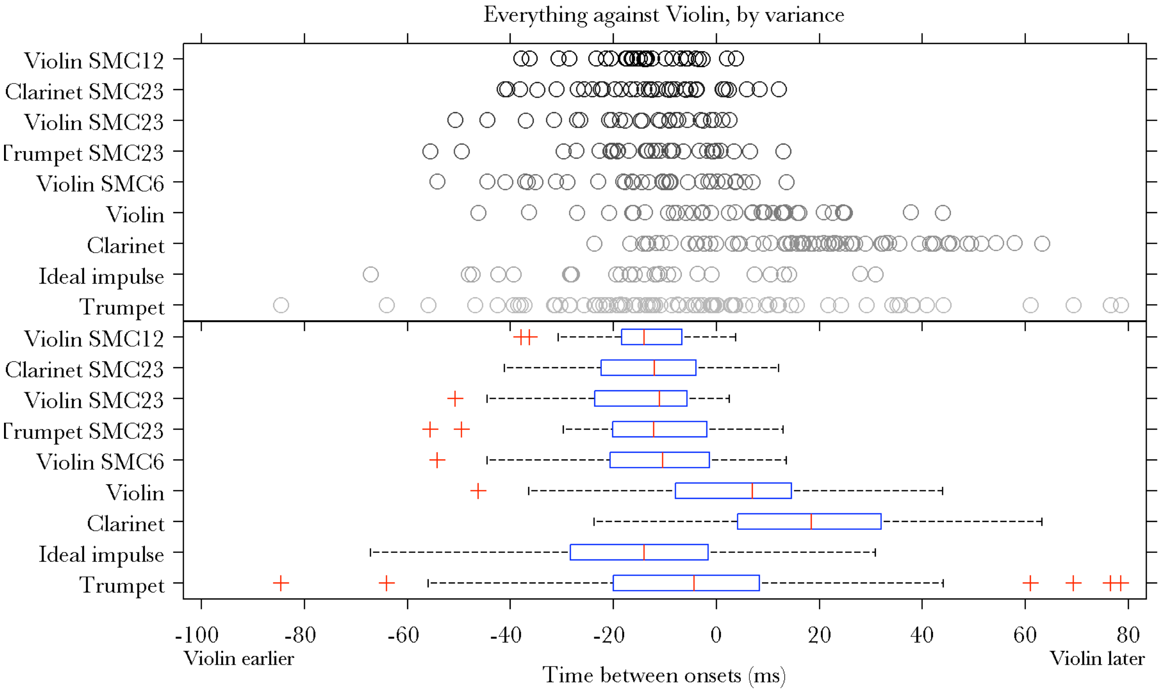


Figure 61: Results of all trials involving Violin, in order of variance

4.4.2 Why Do Subjects Align Shorter Sounds Before Longer Sounds?

Since a click is relatively percussive and short in duration, we would expect its PAT to be close to its physical onset, while we would expect a sound with a more gradual attack to have a relatively later PAT. The naïve assumption of the synchrony method for measuring PAT using these two sounds is that subjects will on average start the longer sound earlier than the click, so that their PATs will line up. On average this was indeed the case, but there was often also a secondary effect that added another mode near starting the click at the same time as the londer sound, or even before. Many natural sounds begin with an impulsive attack followed by a sustained or decaying segment. I believe that in some cases subjects perceived the click and the longer sound as fused into a single sound event (with a stronger attack than the longer sound by itself), and adjusted their relative timing to make the composite result fit this model of impulsive attack followed by sustaining portion. Sound examples such as *quart-Clarinet.vs.C_SMC23* and *quart-C_SMC23.vs.Ideal* demonstrate this percept.

Section 4.3.11 (page 104) discusses this same effect in the context of aligning the ideal impulse with the relatively much longer Clarinet SMC23 click.

4.4.3 Effectiveness of Spectrally Matched Clicks

One of the hypotheses that this experiment investigated was that the distributions would have less variance when the reference sound is spectrally more similar to the sound being tested. This experiment compared the Clarinet, Trumpet, Violin, and Snare sounds to a variety of clicks. In each case at least one of the clicks was spectrally matched to the longer sound, so we can compare the standard deviations for all clicks versus to see whether SMCs do indeed make better reference sounds, as shown in the next four tables (Table 16 through Table 19). The first point to note is that in all four cases the Ideal Impulse is the “worst” reference for measuring the PAT, in the sense that comparisons to the Ideal impulse have the highest standard deviation of any reference.

Sound	*Mean*	*Median*	*STD*	*N*
Trumpet SMC12	-9.27	-8.73	8.46	31
Trumpet SMC23	-9.52	-8.16	9.93	32
Clarinet SMC23	-8.28	-7.11	10.36	28
Violin SMC23	-8.01	-6.20	14.47	30
Trumpet SMC6	-10.26	-6.61	16.91	28
Trumpet	-1.88	0.27	19.70	74
Ideal impulse	-10.39	-8.88	20.62	34

Table 16: Summary statistics for click sounds compared to the trumpet.

Mean, median, and standard deviation are in ms. See also Figure 54: Results of all trials involving Trumpet, in order of variance (page 114).

For the Trumpet (Table 16) the results support the hypothesis: the two clicks with the lowest standard deviation are the 12 and 23 ms clicks that are spectrally matched to the trumpet. The other two 23 ms clicks had higher standard deviations even though the durations are identical. Note that the 6ms SMC had higher variance than any of the 23ms clicks; this suggests that the spectral match at this shorter duration was not good enough to keep Trumpet SMC6 in the same auditory stream as Trumpet.[122]

Sound	***Mean***	***Median***	***STD***	***N***
Violin SMC12	-14.63	-14.10	10.23	29
Clarinet SMC23	-13.16	-12.03	13.66	36
Violin SMC23	-15.36	-11.08	13.74	28
Trumpet SMC23	-13.11	-12.15	14.73	30
Violin SMC6	-13.61	-10.40	16.04	36
Violin	3.36	6.96	19.30	72
Ideal impulse	-14.24	-14.01	23.16	27

Table 17: Summary statistics for click sounds compared to the Violin. Mean, median, and standard deviation are in ms. See also Figure 61: Results of all trials involving Violin, in order of variance (page 117).

For the Violin (Table 17), again the lowest-variance reference sound is an SMC made from the violin, though in second place the Clarinet SMC23 barely beat the Violin SMC23. Again the 6-ms SMC seems not to be sufficiently spectrally matched to be a good reference.

Sound	***Mean***	***Median***	***STD***	***N***
Clarinet SMC23	-37.47	-40.18	11.13	38
Trumpet SMC23	-34.27	-33.88	11.25	34
Clarinet SMC12	-40.95	-42.36	12.52	33
Clarinet SMC6	-31.76	-33.22	13.37	33
Violin SMC23	-28.31	-28.82	13.78	27
Clarinet	3.21	6.49	15.76	90
Ideal impulse	-33.36	-35.94	18.27	33

Table 18: Summary statistics for click sounds compared to the Clarinet. Mean, median, and standard deviation are in ms. See also Figure 44: Results of all trials involving Clarinet, in order of variance (page 109).

Sound	***Mean***	***Median***	***STD***	***N***
Snare	1.07	0.26	7.21	80
Snare SMC3	-17.53	-15.19	10.94	36
Ideal impulse	-13.89	-14.05	12.82	52

Table 19: Summary statistics for click sounds compared to the Snare. Mean, median, and standard deviation are in ms. See also Figure 49: Results of all trials involving Snare (page 111).

[122] Section 3.7.3 (page 68) suggests some more sophisticated techniques for synthesizing SMCs; perhaps one of these would be able to make a 6ms SMC that was a better spectral match to the Trumpet.

Unlike the synthetic orchestral tones, the Snare drum's smallest standard deviation comes from aligning a second copy of the same sound. I believe this is due to a combination of the greater percussiveness of the Snare sound and possible comb filtering effects. Of the two clicks that subjects aligned with the Snare, the Snare SMC3 had the lower standard deviation, but this is not surprising, since the ideal impulse is the reference with the highest standard deviation in all four of these cases. Again I conclude that the short SMC may not allow the SMC synthesis method to produce a sufficiently close spectral match.

I conclude that Spectrally Matched Clicks are indeed superior reference sounds for measuring PAT, though with the SMC synthesis methods used in this study (see Section 3.7.1 on page 65) it appears that the duration of the SMC needs to be greater than 6ms.

4.4.4 Intrinsic PAT-pdf Means and Variances of all Sound Events

	Algorithm Name	*log likelihood 17*	*log likelihood 11*
1	Shortest-variance path from Ideal Impulse	-5837.4	-5139.4
2	Shortest-variance path from best likelihood reference	-5824.9	-5125.9
3	Greedy next smallest intrinsic variance starting from ideal impulse	-5862.6	-5183.1
4	Greedy next smallest intrinsic variance starting from best likelihood reference	-5882.1	-5217.2
5	Batch estimation starting from Ideal impulse	-6014.1	-5283.5
6	Batch estimation starting from Snare SMC3 (best likelihood)	-6008.0	-5277.4
7	Variances from Trials Against Self & Maximum Likelihood Means	-5720.6	-5054.3
8	Least Squares Means & Variances from Trials Against Self	-5723.6	-5058.1

Table 20: The eight methods used to estimate all sound's intrinsic PAT means and variances by assuming normality. See Section 3.5.5 (page 56) for descriptions of the methods. The last two columns are the overall log likelihood of each model given all the observed experimental data. "17" refers to the first set of estimates, for all 17 single-event sounds, as shown in Table 21 and Table 22, while "11" refers to the second set of estimates, taking into account only the 11 sounds that were each compared against two other sounds, as shown in Table 23 and Table 24. The overall likelihoods are much higher for the second model simply because it considers many fewer trials. Algorithms 4 and 6 have the free parameter of which sound to take as the starting reference point for the grid search, chosen to produce the highest overall likelihood. Algorithm 4 chose Clarinet SMC23 and Clarinet SMC6 respectively for the "17" and "11" runs, while Algorithm 6 chose Snare SMC3 in both cases.

This section applies the theoretical results of Section 3.5.5 (page 56) to the observed data from this experiment. Assuming that every sound's intrinsic PAT-pdf is Gaussian, what are the intrinsic mean[123] and variance for each sound? The first three algorithms derive everything from a known starting point, so I ran each of them twice: once using the ideal impulse as the known starting

[123] Every algorithm uniformly shifts all of its estimated intrinsic means so that the minimum will be zero. This is justified only because in each case some of the sounds are very impulsive and "should" have their PATs near their physical onsets.

point, and a second time trying each of the sounds as a starting point and selecting the result with the highest likelihood. Table 20 names the eight algorithms. Table 21 and Table 22 show the intrinsic means and variances, respectively, resulting from each algorithm when applied to all of the single-event sounds in the experiment.

sound	***alg. 1***	***alg. 2***	***alg. 3***	***alg. 4***	***alg. 5***	***alg. 6***	***alg. 7***	***alg. 8***	***mean***
Clarinet SMC12	0	0	0	0	0	0	0	0	0
Trumpet SMC6	1.503	1.503	1.503	1.503	7.721	7.721	4.325	7.742	4.1903
Violin SMC12	2.606	2.606	2.021	2.021	7.206	7.206	5.615	7.397	4.5846
Maurit. SMC12	0.525	0.525	3.572	3.572	6.505	6.505	9.017	6.899	4.6399
Violin SMC6	0.6251	0.6251	3.672	3.672	6.605	6.605	9.921	8.058	4.9729
Clarinet SMC6	1.94	1.94	1.94	1.94	7.92	7.92	9.025	7.274	4.9872
Trumpet SMC23	2.235	2.235	2.235	2.235	7.951	7.951	8.014	7.975	5.104
Trumpet SMC12	2.485	2.485	2.485	2.485	8.703	8.703	7.082	8.723	5.3937
Clarinet SMC23	3.483	3.483	3.483	3.483	6.443	6.443	9.69	7.563	5.509
Ideal impulse	1.611	1.611	4.657	4.657	7.59	7.59	10.43	7.984	5.7663
Snare SMC3	1.758	1.758	4.805	4.805	7.738	7.738	10.57	7.078	5.781
Brazil SMC23	2.189	2.189	5.236	5.236	8.168	8.168	11.01	8.563	6.3444
Violin SMC23	1.871	1.871	4.918	4.918	9.781	9.781	11.39	9.701	6.7791
Trumpet	11.76	11.76	11.76	11.76	17.98	17.98	17.97	18	14.8688
Violin	17.23	17.23	16.65	16.65	21.83	21.83	22.63	22.02	19.5094
Snare	19.28	19.28	22.33	22.33	21.48	25.26	25.86	22.99	22.3534
Clarinet	40.95	40.95	40.95	40.95	40.95	40.95	42.04	40.95	41.0907

Table 21: Intrinsic means for all single-event sounds as estimated by the algorithms listed in Table 20 given all the results of the experiment. The table is sorted by the last column, which gives the mean result across all 8 algorithms.

sound	***alg. 1***	***alg. 2***	***alg. 3***	***alg. 4***	***alg. 5***	***alg. 6***	***alg. 7***	***alg. 8***	***mean***
Clarinet SMC12	33.03	33.03	43.19	36.83	0.01	0.01	9.785	9.785	20.71
Trumpet SMC6	214.2	214.2	188.7	182.3	0.01	0.01	27.51	27.51	106.8
Violin SMC12	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Maurit. SMC12	10.46	10.7	7.009	0.649	10.4	10.41	0.5396	0.540	6.34
Violin SMC6	8.546	8.546	11.78	5.419	15.17	15.18	0.6893	0.690	8.25
Clarinet SMC6	10.95	10.95	0.7838	7.144	18.48	18.49	3.425	3.425	9.20
Tpt. SMC23	27.01	27.01	1.462	0.01	0.01	0.01	0.01	0.01	6.941
Tpt. SMC12	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Clarinet SMC23	0.01	0.01	10.17	3.814	68.58	68.59	15.79	15.79	22.85
Ideal impulse	0.2515	0.01	3.7	10.06	0.3083	0.3	0.2515	0.2515	1.8915
Snare SMC3	4.698	4.939	1.25	0.01	4.641	4.649	0.487	0.487	2.65
Brazil SMC23	3.564	3.805	0.1156	0.01	3.507	3.515	0.5309	0.5309	1.5
Violin SMC23	0.01	0.01	0.01	3.137	47.38	47.39	7.456	7.456	14.11
Trumpet	71.56	71.56	97.11	103.5	424.9	424.9	194.1	194.1	197.71
Violin	104.7	104.7	176.3	182.7	535.9	535.9	186.3	186.3	251.61
Snare	115.1	26.03	118.5	119.7	164	115.1	26.03	26.03	88.81
Clarinet	123.8	123.8	113.6	120	333.4	333.4	124.2	124.2	174.5

Table 22: Intrinsic variances for all single-event sounds as estimated by the algorithms listed in Table 20 given all the results of the experiment. The last column gives the average estimated intrinsic variance across all 8 algorithms. Rows are in the same order as in Table 21.

Given the dissimilarities among the eight algorithms it is no surprise that the results diverge in many cases. However, there are also points of agreement among the eight algorithms, such as every algorithm's estimated mean of the Clarinet's PAT-pdf being around 41-42 ms after its physical onset. Why was the Clarinet SMC12 the sound with the earliest PAT-pdf in every case? Referring to Figure 41 (page 107), we see that Clarinet SMC was only compared against the Clarinet, and not to any other sounds, so all eight algorithms have no choice but to derive the Clarinet SMC's intrinsic mean from that of the Clarinet.

With that in mind, I re-ran all 8 algorithms on only the 11 sounds that were compared against at least two other sounds (thereby excluding Clarinet SMC12 as well as Brazil SMC23, Mauritania SMC12, Trumpet SMC6, Trumpet SMC12, and Violin SMC12). Table 23 and Table 24 show the results. Note that the ordering of the sounds by average estimated mean is the same (that is, the order of rows in Table 23 is the same as in Table 21 for the sounds that appear in both tables).

sound	***alg. 1***	***alg. 2***	***alg. 3***	***alg. 4***	***alg. 5***	***alg. 6***	***alg. 7***	***alg. 8***	***mean***
Violin SMC6	0.00	0.00	1.73	1.73	0.16	0.16	2.18	0.98	0.87
Clarinet SMC6	1.31	1.31	0.00	0.00	1.48	1.48	1.39	0.20	0.90
Trumpet SMC23	1.61	1.61	0.30	0.30	1.51	1.51	0.00	0.90	0.97
Clarinet SMC23	2.86	2.86	1.54	1.54	0.00	0.00	2.05	0.48	1.42
Ideal impulse	0.99	0.99	2.72	2.72	1.15	1.15	2.98	0.91	1.70
Snare SMC3	1.13	1.13	2.87	2.87	1.29	1.29	3.12	0.00	1.71
Violin SMC23	1.25	1.25	2.98	2.98	3.34	3.34	3.58	2.62	2.67
Trumpet	11.13	11.13	9.82	9.82	11.53	11.53	10.66	10.92	10.82
Violin	16.61	16.61	14.71	14.71	15.39	15.39	14.91	14.94	15.41
Snare	18.66	18.66	20.39	20.39	15.04	18.82	18.41	15.92	18.29
Clarinet	40.33	40.33	39.01	39.01	34.51	34.51	34.43	33.88	37.00

Table 23: Intrinsic means for all sounds that were compared to at least two other sounds, estimated by the algorithms listed in Table 20.

The data is sorted by the last column, the average estimate from all 8 algorithms.

sound	***alg. 1***	***alg. 2***	***alg. 3***	***alg. 4***	***alg. 5***	***alg. 6***	***alg. 7***	***alg. 8***	***mean***
Violin SMC6	8.55	8.55	10.78	5.62	15.18	15.18	0.69	0.69	8.15
Clarinet SMC6	10.95	10.95	1.78	6.95	18.49	18.49	3.43	3.43	9.31
Trumpet SMC23	0.01	0.01	0.46	0.01	0.01	0.01	0.01	0.01	0.07
Clarinet SMC23	0.01	0.01	9.17	4.01	68.59	68.59	15.79	15.79	22.75
Ideal impulse	0.25	0.01	4.70	9.86	0.30	0.30	0.25	0.25	1.99
Snare SMC3	4.70	4.94	0.25	0.01	4.65	4.65	0.49	0.49	2.52
Violin SMC23	0.01	0.01	0.01	2.94	47.39	47.4	7.46	7.46	14.08
Trumpet	98.56	98.6	98.11	103.3	424.9	424.9	194.1	194.1	204.6
Violin	188.8	188.8	177.3	182.5	535. 9	535.9	186.3	186.3	272.7
Snare	115.1	26.03	119.5	119.7	164.0	115.1	26.0	26.03	88.93
Clarinet	123.7	123.7	114.6	119.8	333.4	333.4	124.2	124.2	174.6

Table 24: Intrinsic variances for all sounds that were compared to at least two other sounds, estimated by the algorithms listed in Table 20

4.5 Discussion

As expected subjects certainly were not perfectly consistent in their results for this experiment. The distributions of results for the different pairs of sounds have noticeably different shapes, the most obvious factor being the large differences in standard deviation. Many, but not all, of the results appear to be Gaussian.

The experiment supported all of the factors motivating Spectrally Matched Click Synthesis. The ideal impulse consistently had the highest variance of any reference sound aligned against any non-click sound. Although it is perfectly localized in time, it seems especially difficult to align natural sounds consistently with the ideal impulse, as predicted by the completely broad spectrum of the ideal impulse in light of the auditory streaming factors discussed in Section 3.6 (page 61). However, a Spectrally Matched Click was the most consistent reference for each of the natural instrumental tones.

Also as predicted, aligning the ideal impulse against itself was the lowest-variance task in the entire experiment. As figure Figure 46 (page 110) shows, the variance increased by a small amount when aligning the ideal impulse with very short clicks or clicks derived from percussive sounds, then increased by more when aligning the ideal impulse with longer SMCs derived from the instrumental tones, and then became extremely large when aligning the ideal impulse against the instrumental tones themselves. Table 13 (page 101) shows this property of the ideal impulse in another way: although subjects were extremely consistent aligning two copies of the ideal impulse (standard deviation 0.71 ms, as described in Section 4.3.11 [page 104]), the ideal impulse generally brought a lot of "extra" variance when aligned with another sound.

For modeling purposes, I had hoped that the mean of the ideal impulse would be close to zero, i.e., that all other sounds' PATs would on average be later than that of the ideal impulse. As Section 4.3.11 (page 104) explains, this was not the case; three other short click sounds on average were aligned with the ideal impulse as if their PATs were earlier than that of the ideal impulse, so that in Table 21 the estimated mean of the ideal impulse is 1.6 to 10.4 ms after its physical onset. By excluding the sounds that were not compared against at least two other sounds, all estimates for the mean of the ideal impulse's PAT-pdf stayed below 3ms.

The trials from tasks aligning multiple copies of a single-event sound to a looped recording of metric music should definitely be analyzed in the future.

Chapter 5 Motivations, Implications, and Future Work

5.1 Predictive Modeling of PAT

It would be useful to have a predictive model of PAT that could take in any arbitrary input sound event (whose PAT has not been measured experimentally), compute some deterministic function of the acoustic signal, and output an estimate of the sound's PAT. More generally, a PAT model could take in arbitrary input sound consisting of arbitrarily many sound events, and combine the onset-detection-like task of identifying each individual perceived event along with estimating the PAT for each event.

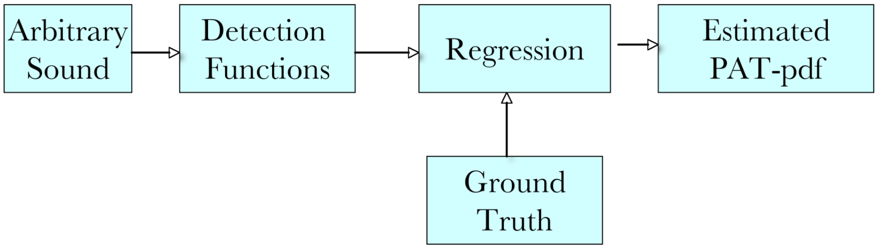


Figure 62: System dataflow diagram for predictive modeling of PAT-pdf with regression.

Existing models for predicting PAT (or P-Center) all treat PAT as a single moment in time (Collins 2006; Villing, Ward, and Timoney 2007). The most important distinction among these models is between those that use information only from the local portion at the beginning of a sound event versus those that consider the entire sound event to compute its PAT. All models compute one or more features from the acoustic signal such as a rectified signal (Vos and Rasch 1981), slope of the energy envelope (Gordon 1987), energy in one or more spectral bands (Marcus 1981; Scott 1998), or modulation of energy in spectral bands (Harsin 1997). All predictive models of which I am aware take in one entire isolated sound event as input, so to use them to find the PAT of multiple sound events in a continuous audio signal would require first segmenting the signal manually and/or with an automatic onset detector.

The novel approach here, continuing with the theme of a continuous probability density function representation for PAT, is to formulate the prediction of PAT as a regression problem, as shown in Figure 62. The inputs to the predictive model are various functions of time computed from the acoustic signal, in particular, a large collection of the *detection functions* (described in Section 5.1.1) known in the onset detection literature. The output of each model is another (sampled) continuous function of time whose value at each moment is proportional to the probability of a listener assigning a sound's PAT to that moment. With formulation we can try a wide variety of regression techniques to attempt to learn a relationship between the detection functions and subjects' PAT-pdf results in a supervised machine learning context.

Another approach bypasses the theoretical (Section 3.5) and practical (Section 4.4.4) difficulties of decomposing relative PAT measurements for pairs from pairs of sounds into intrinsic PATs for each individual sound. Instead of trying to predict the absolute PAT of a single sound, we instead directly try to predict the relative PAT for a pair of sounds.

5.1.1 Detection functions

For each input sound I compute a large number of *detection functions*, a term from the computational onset detection (see Section 2.3.3 on page 17 for a review) and beat induction literature that means a function derived from the raw audio and whose value tends to show peaks at moments of onsets. They are generally low-level descriptors computed from raw audio and sampled at a much lower rate (for example, at a rate determined by Short-Term Fourier Transform (STFT) hop size[124]). For example, one of the most trivial detection functions is signal energy in each 1024-sample audio frame.

I have implemented many of the detection functions listed in Gouyon's comprehensive review (Gouyon 2005): low frequency energy, spectral centroid, spectral flatness, spectrum energy, spectrum energy (normalized and on a db scale), spectrum geometric mean, the four variants of "high frequency content," spectrum maximum magnitude frequency, spectrum mean, spectrum rolloff, spectrum slope, and spectrum spread. I also implemented two of the detection functions from the QMUL group used in the Aubio library: aubio_complex and aubio_phase (Brossier 2006) and compute the thirteen mel-frequency cepstral coefficients (MFCC) using the *Auditory Toolbox* (Slaney 1993-1994). This gives a total of 43 detection functions.

In addition, also following the lead of (Gouyon 2005), for each "raw" detection function I also compute the half-wave rectification of the first-order difference.[125] This produces a second set of 43 detection functions, for a total of 86. Gouyon also suggests using the half-wave rectification of the first-order difference of a mulaw-compressed[126] version of each detection function, but the

[124] http://ccrma.stanford.edu/~jos/sasp/Practical_Computation_STFT.html

[125] "Half-wave rectification" means setting every negative value to zero. "First-order difference" means computing the difference between each successive pair of values; it is an approximation to the first derivative for sampled signals.

[126] Gouyon's formula for mulaw compression (in Matlab notation) is

slight difference between the mulaw and non-mulaw versions of the half-wave rectification of the first-order difference has been a source of trouble in the regression context. Sometimes the resulting vectors are very nearly equal, requiring that one of the two be discarded before attempting regression. In other instances, the regression often results in a huge weight for one function, and then a nearly equal weight of opposite magnitude for the mulaw version of that function, which tends to generalize poorly. Therefore I'm currently eliminating the mulaw versions from further processing.

As most of these functions are based on the STFT[127], an outer "main loop" procedure handles all details of the STFT, passing either an FFT frame or the magnitudes from an FFT frame as appropriate to each detection function.[128] Therefore each detection function is only a few lines of Matlab code. Before further processing I normalize the output of each detection function to the range [0,1] for each sound.[129]

All my sounds have a sampling rate Fs=44100. For now I'm using a frame size of M=1024 samples (23.2 ms), Hamming windows, size $Nfft$=4096 FFTs (that is, zero padding by a factor of four), and a hop size (a.k.a. frame rate) of R=20 samples (0.45 ms). The results of changing some of these settings (particularly frame size) could be considered additional detection functions.

I use the convention that the time of an STFT frame is the time of the *center* of the window, as in the SDIF standard (Wright et al. 1999). The *Auditory Toolbox* instead uses the convention that the time of one sample of the MFCC output is the time of the *beginning* of the STFT window that produced that output, so I insert $Nfft/2$ zeros at the beginning of each input sound before passing

```
        y = log(1+mu*x) ./ log(1+mu);
```

This formula returns an imaginary value when the input is negative, which sometimes happens with, for example, spectral slope. Therefore, rather than using always mu=100, when necessary I set mu to guarantee that 1+mu*x is at least 0.0001.

[127] http://ccrma.stanford.edu/~jos/sasp/Short_Time_Fourier_Transform.html

[128] The detection function procedures are organized in subdirectories according to what input they expect: functions in df-fft take the current complex spectrum as input (and the past 2 complex spectra as optional second and third arguments), functions in df-fft-mag take the magnitude spectrum as the first argument and the frequency sampling interval (in Hertz) as the optional second argument. A makefiile automatically generates the Matlab program all_df_names.m, which produces cell arrays of the names of all the detection function procedures in each input type category. The outer STFT loop then iterates through this cell array using FEVAL to invoke each function on each STFT frame.

[129] The exception is cases in which the detection function output is completely constant, in which case of course no linear remapping can make the function occupy the entire range from 0 to 1. (This sometimes occurs, for example, with Spectrum Maximum Magnitude Frequency, which for many of the tones is the frequency of the FFT bin of the fundamental frequency in every frame.) Of course a completely constant detection function isn't going to contribute anything in a regression context (where there's already a weight for a constant term), so any such detection function will soon be discarded anyway.

it to the MFCC function. Also, because I want to sample the detection function at times centered all throughout the duration of the sound, I add ($Nfft/2$)+R zeros at the end as well.

The following sections describe specific detection functions. In each case X stands for the current STFT frame and X_i stands for the complex value of the i^{th} frequency bin[130] of X, with $N=Nfft/2$ the number of bins with nonnegative frequencies and $0 \leq i \leq N\text{-}1$.

$F=Fs/Nfft=44100Hz/4096\approx 10.76\ Hertz$ is the "frequency sampling interval" or "frequency step," in other words, the distance in Hertz between the frequencies of successive STFT bins: iF is the center frequency (in Hertz) if the i^{th} STFT bin.

5.1.1.1 Zero-crossing rate (ZCR)

The ZCR is the number of waveform time-domain zero-crossings (i.e. sign changes), divided by the number of samples minus one.[131] It will generally be higher when signals are noisier, brighter, or higher in pitch.

5.1.1.2 Spectrum mean

The mean magnitude of the spectrum:

$$mean(X) = \tfrac{1}{N} \sum_{i=0}^{N-1} \left| X_i \right|$$

5.1.1.3 Spectrum Spread

The variance of the magnitudes in the spectrum:

$$spread(X) = \tfrac{1}{N} \sum_{i=0}^{N-1} \left(\left| X_i \right| - mean(X) \right)^2$$

5.1.1.4 Spectrum Geometric Mean

The geometric mean is the N^{th} root of the product of N elements; here we take the geometric mean of the amplitudes:

$$geometricMean(X) = \left(\prod_{i=0}^{N-1} \left| X_i \right| \right)^{1/N}$$

[130] See http://ccrma.stanford.edu/~jos/mdft/Spectral_Bin_Numbers.html for why these are called "bins."

[131] A sequence of length n can have a maximum of n-1 sign changes, so dividing by n-1 keeps the result in the range [0,1].

5.1.1.5 Spectral Flatness

One measure of the flatness of a spectrum is the ratio between the geometric mean and the mean:

$$SpectrumFlatness(X) = GeometricMean(X) \,/\, Mean(X)$$

Note that if the STFT frame is completely silent (e.g., if it comes in a frame after the end of a short click) then this formula causes a division by zero.

5.1.1.6 Spectrum Slope

The slope of the straight line that is the best fit to the magnitude spectrum. I used Matlab's "polyfit" to fit the straight line:

```
function s = spectrum_slope(Xm, freqstep)
freqs = freqstep * [0:length(Xm)-1];
p = polyfit(freqs, Xm, 1);
s = p(1);
```

5.1.1.7 Spectrum Energy

The energy is the sum of the squares of the time-domain samples, which is equal to the sum of the squares of the STFT magnitudes:[132]

$$energy(X) = \sum_{i=0}^{N-1} |X_i|^2$$

We can also normalize to find the energy per sample

$$NormalizedEnergy(X) = \tfrac{1}{N} \sum_{i=0}^{N-1} |X_i|^2$$

and also convert to the decibel scale:[133]

$$NormalizedEnergyDB(X) = 20 * \log_{10} \tfrac{1}{N} \sum_{i=0}^{N-1} |X_i|^2$$

5.1.1.8 Low-Frequency Energy

Low-Frequency Energy is the proportion of the energy below 100 Hz. I approximate this by taking the proportion of energy in STFT bins whose bin frequency is ≤ 100 Hz, in other words, I

[132] http://ccrma.stanford.edu/~jos/mdft/Rayleigh_Energy_Theorem_Parseval_s.html

[133] http://ccrma.stanford.edu/~jos/mdft/Decibels.html

don't worry about dividing up the energy in the STFT bin whose range of frequencies spans 100 Hz.[134]

Note that, as with Spectral Flatness, the value of this detection function becomes undefined (because of a division by zero) when there is no energy at all in the frame.

5.1.1.9 **Chroma Energy**

We can partition the spectrum energy into the twelve pitch classes of equal temperament. So, for example, the Chroma Energy for the pitch class F# is the sum of the squared amplitudes of all STFT bins whose (center) frequency is closer to one of the octaves of F# than to any other equal-tempered note. As with Low-Frequency Energy, each bin counts all-or-nothing towards exactly one chroma.

5.1.1.10 **High Frequency Content Family**

The detection function "High Frequency Content" ("HFC"), working on an STFT magnitude spectrum, has been defined by different researchers to be the sum of amplitude or energy weighted by frequency or frequency squared. Therefore I propose the following naming convention[135] for the four possible variants:

- HFC: sum of amplitude weighted by frequency
- HFCC: sum of energy (amplitude squared) weighted by frequency
- HFFC: sum of amplitude weighted by frequency squared
- HFFCC: sum of energy weighted by frequency squared

$$HFC(X) = \sum_{i=0}^{N-1} iF\left|X_i\right|$$

$$HFFC(X) = \sum_{i=0}^{N-1} (iF)^2\left|X_i\right|$$

$$HFCC(X) = \sum_{i=0}^{N-1} iF\left|X_i\right|^2$$

$$HFFCC(X) = \sum_{i=0}^{N-1} (iF)^2\left|X_i\right|^2$$

[134] A better approximation would consider that each STFT bin represents the signal's energy over a range of frequencies (http://ccrma.stanford.edu/~jos/mdft/Spectral_Bin_Numbers.html). So in the case of *fs=44100* and *Nfft*=4096, *freqstep* is 10.766 Hz, so the ninth STFT bin, with center frequency 9*10.766 =96.8994, actually extends from 91.5161 Hz to 102.2827 Hz, so only (100-91.161)/ 10.766 = 0.821 times the energy in the ninth STFT bin would be counted as below 100 Hz. A yet more sophisticated approximation would consider the filter response of the window function and the DFT (http://ccrma.stanford.edu/~jos/mdft/Frequencies_Cracks.html).

[135] This naming convention is obviously inspired by the algebraic notation for multiplication: since *FF* means "*F* times *F*", the versions that use frequency have "F" in their names while those that use frequency squared have "FF" in their names.

5.1.1.11 Spectral Centroid

The spectral centroid is the "center of gravity" of the magnitude frequency spectrum, and is an acoustic correlate of the perceptual "brightness" of a signal (Wessel 1979):

$$centroid(X) = \frac{\sum_{i=0}^{N-1} iX_i}{\sum_{i=0}^{N-1} X_i}$$

Note that this definition (following Gouyon) is in units of bin number rather than frequency; multiply by *F* to get Hertz.

5.1.1.12 Spectrum Maximum Magnitude Frequency

This is the frequency of the STFT bin which has the highest frequency. For harmonic tones this is often but not always the fundamental frequency.

$$\forall i, \left| X_{SpectrumMaximumMagnitudeFrequency(X)/F} \right| \geq \left| X_i \right|$$

5.1.1.13 Spectrum Rolloff

Spectrum Rolloff is the frequency below which 85% of the signal energy remains. As with Low-Frequency Energy and Chroma Energy we don't worry about dividing the energy in each bin, so *SpectrumRolloff(X)* will always be an integer multiple of *F*.

$$\sum_{i=0}^{(SpectrumRolloff(X)/F)-2} |X_i|^2 < 0.85 Energy(X)$$

$$\sum_{i=0}^{(SpectrumRolloff(X)/F)-1} |X_i|^2 \geq 0.85 Energy(X)$$

5.1.1.14 Two Phase/Complex Based Detection Functions from *Aubio*

The basic idea is that for a tonal signal in the steady state, the unwrapped phase in each frequency bin should advance by approximately $2\pi fR$ during each frame of the STFT. So given $\phi_n(i-1)$, the phase in the n^{th} frequency bin in STFT frame *i*-1, we can estimate the phase of that bin in STFT frame *i:*

$$\hat{\phi}_n(i) = \mathrm{mod}(\phi_n(i-1) + 2\pi f_n R, 2\pi)$$

We compare the actual phase in the i^{th} STFT frame to this estimate, and treat the amount of error as the detection function. Higher values of the error indicate a greater likelihood of an attack, rearticulation, etc.

For details on this family of detection functions see (Bello et al. 2005; Bello et al. 2004; Bello and Sandler 2003; Dixon 2006; Duxbury et al. 2003a, 2003b; Duxbury, Sandler, and Davies 2002); my implementation is a Matlab translation of these two detection functions from Paul Brossier's C language *Aubio*[136] library (Brossier 2006).

5.1.2 Setup for Intrinsic PAT-pdf Regression

We're going to be given any arbitrary sound s in the form of a sampled audio signal $s(t)$, and we will output a continuous estimate of the relative probability $p(t)$ of PAT being perceived in each moment. The scaling is immaterial, but for consistency let's say that $-1 \leq s(t) \leq 1$ and that $p(t)$ is in units of probability per millisecond of PAT occurring.

We will learn a function h (for "hypothesis") that generates the estimate $\hat{p}(t)$ given $s(t)$:

$$\hat{p} = h(s)$$

Our training data will consist of the estimated PAT-pdf curves for a set of sounds whose PAT-pdf we have measured experimentally. Specifically the "inputs" are the detection functions computed for each of these sounds, and the "targets" are the measured/estimated PAT-pdf curves.

If h is causal then $\hat{p}(t_0) = h(s(t)), t \leq t_0$, in other words, h can only "see" the portion of the input before the current moment.

If h has bounded look-ahead then $\hat{p}(t_0) = h(s(t)), t \leq t_0 + k$, in other words, h can only "see" the input from the past and up to k seconds into the future.

If $\hat{p}(t_0)$ is a linear combination of a finite number of $h(s(t)), t \leq t_0$ then h is a (causal) FIR filter.

Rather than directly trying to estimate $p(t)$ directly from the sampled waveform $s(t)$, we will instead compute all N_{df} of our detection functions and use them as the inputs to h:

$$h(s) = h(df_1(s), df_2(s), df_3(s), \ldots df_{Ndf}(s))$$

Now if h is causal, than it can only "see" the values of the detection functions up to the present.

[136] http://aubio.org

There is good reason to assume that PAT depends on features of the input signal within, say 100 ms of each moment, taking into account the recent past and what is about to occur in the future.[137]

$$\hat{p}(t_0) = h(df_1(t), df_2(t), \ldots df_{Ndf}(t)),\ t_0 - memory \le t \le t_0 + lookahead$$

We can begin with the drastically simplifying assumption that h will look only at the current value for each detection function, i.e.,

$$\hat{p}(t_0) = h(df_1(t_0), df_2(t_0), \ldots df_{Ndf}(t_0))$$

With this assumption, the goal of our regression model is to find the vector of weights Θ such that

$$A\Theta \approx B$$

where A is the matrix of all the inputs and B is the (column) vector of targets. Each column of A is a detection function. We must make sure to sample the detection functions and the ground truth intrinsic PAT-pdf shapes with the same sampling interval T and aligned with the same time zero. We will "stack" all of the sounds vertically in both A and B, so that the rows of A and B are all of the times for sound 1, then all of the times for sound 2, etc.:

$$A = \begin{pmatrix} df_1(s_1)(0) & \cdots & df_{Ndf}(s_1)(0) \\ df_1(s_1)(T) & \cdots & df_{Ndf}(s_1)(T) \\ df_1(s_1)(2T) & \cdots & df_{Ndf}(s_1)(2T) \\ \vdots & \cdots & \vdots \\ df_1(s_1)(end) & \cdots & df_{Ndf}(s_1)(end) \\ df_1(s_2)(0) & \cdots & df_{Ndf}(s_2)(0) \\ \vdots & \cdots & \vdots \\ df_1(s_2)(end) & \cdots & df_{Ndf}(s_2)(end) \\ \vdots & \cdots & \vdots \\ df_1(s_{num\_sounds})(end) & \cdots & df_{Ndf}(s_{num\_sounds})(end) \end{pmatrix}, B = \begin{pmatrix} I_1(0) \\ I_1(T) \\ I_1(2T) \\ \vdots \\ I_1(end) \\ I_2(0) \\ \vdots \\ I_2(end) \\ \vdots \\ I_{num\_sounds}(end) \end{pmatrix}$$

As in Section 3.4.3 (page 42), I stands for "intrinsic PAT-pdf," and here I'm treating each intrinsic PAT-pdf and each detection function as a sampled function of time and indexing it with a time in seconds. The notation $df_i(s_j)$ means the ith detection function computed on the jth sound. (Also

[137] Many models of P-Center, for example, look at the audio signal for an entire syllable before estimating the P-center.

I'm using the Octave/Matlab convention that "end" is a special index meaning "the final sample of the given signal.")

Without our simplifying assumption, there are at least three options for considering the influence of detection functions from nearby times:

1. The brute force solution: consider every possible time shift between *memory* and *lookahead* to be its own valid subset of regression inputs. So, for example, if R is 0.45 ms, *memory* is 200 ms and *lookahead* is 100 ms, then consider all (100+200)/0.45= 667 possible time shifts, for a total of $667 \times 86 = 57362$ input signals to the regression. Obviously this is an unwieldy number of dimensions, so going this route will require the use of feature selection or other techniques for dimensionality reduction.

2. Time shift each detection function individually. Rather than considering *all* possible time shifts of each detection function as above, look for the single best time shift for each one. One approach would be to cross-correlate each detection function with the ground truth PAT-pdf estimates for our training set, choose the lag that maximizes the absolute value of the cross-correlation, and then always time-shift that detection function by the resulting lag.[138]

3. The system identification solution: consider each detection function separately, and use any of a huge variety of FIR system identification techniques[139] to produce an FIR filter (possibly constrained to be low order) whose output best approximates the intrinsic PAT-pdfs when the input is the given detection function. Now we have produced the "best" weighted linear combination of time shifts for each detection function, so there is only one input to the full regression step for each detection function.

5.1.3 Setup for Pairwise Relative PAT-pdf Regression

The previous section describes a setup for predicting (estimated) intrinsic PAT-pdf for each individual sound from the acoustic properties of that sound. This section suggests another setup in which we directly estimate the pdf of the difference in PAT between a pair of sounds, thereby the theoretical (Section 3.5) and practical (Section 4.4.4) difficulties of recovering each sound's own PAT from measurements of relative PAT for pairs of sounds.

[138] In an early phase of this research, before I had carefully thought through the issues discussed in Section 3.4.4 (page 43), when I naively thought that the distribution of subjects' delay times could be interpreted directly as the PAT-pdf of the test sound, I used this method with some success to predict the shapes of some of the figures in (Gordon 1987) directly from the audio. I have not yet tried this method on estimates of intrinsic PAT-pdf.

[139] http://ccrma.stanford.edu/~jos/mdft/FIR_System_Identification.html

Now the target functions are the distributions of the time delay between physical onsets to make a pair of sounds be perceived as synchronous; these are the distributions that can actually be directly estimated from listening experiments. (In other words, these are the shapes of the distributions for the random variable D as described in Section 3.4.4 (page 43).)

We're going to be given two arbitrary sounds s_i and s_j in the form of sampled audio signals, and we will output an estimate of the relative probability of s_i and s_j being perceived as synchronous as a function of the delay time between their physical onsets.

The inputs to the regression algorithm are the cross-correlation of each detection function as computed on s_i and s_j. Let $x_{i,j,n}(l)$ be the cross-correlation between $df_n(s_i)$ and $df_n(s_j)$ as a functin of lag time l, where l ranges from l_{min} to l_{max} in the usual way depending on the lengths of $s1$ and $s2$.

Again the goal of our regression model is to find the vector of weights Θ such that

$$A\Theta = B$$

where A is the matrix of all the inputs and B is the (column) vector of targets. Again each column of A corresponds to a single detection function, and we vertically "stack" the signals from our training data. The difference is that now each group of rows is a pair of sounds (that were compared to each other in the experiment), rather than an individual sound. For sake of illustration, here's how A and B would look in a situation where we had experimental PAT alignment results for all six unordered pairs of the three sounds s_1, s_2, and s_3 (1-1, 1-2, 1-3, 2-2, 2-3, and 3-3):

$$A = \begin{pmatrix} x_{1,1,1}(l_{\min}) & \cdots & x_{1,1,n}(l_{\min}) \\ \vdots & \cdots & \vdots \\ x_{1,1,1}(l_{\max}) & \cdots & x_{1,1,n}(l_{\max}) \\ x_{1,2,1}(l_{\min}) & \cdots & x_{1,2,n}(l_{\min}) \\ \vdots & \cdots & \vdots \\ x_{1,2,1}(l_{\max}) & \cdots & x_{1,2,n}(l_{\max}) \\ x_{1,3,1}(l_{\min}) & \cdots & x_{1,3,n}(l_{\min}) \\ \vdots & \cdots & \vdots \\ x_{1,3,1}(l_{\max}) & \cdots & x_{1,3,n}(l_{\max}) \\ x_{2,2,1}(l_{\min}) & \cdots & x_{2,2,n}(l_{\min}) \\ \vdots & \cdots & \vdots \\ x_{2,2,1}(l_{\max}) & \cdots & x_{2,2,n}(l_{\max}) \\ x_{2,3,1}(l_{\min}) & \cdots & x_{2,3,n}(l_{\min}) \\ \vdots & \cdots & \vdots \\ x_{2,3,1}(l_{\max}) & \cdots & x_{2,3,n}(l_{\max}) \\ x_{3,3,1}(l_{\min}) & \cdots & x_{3,3,n}(l_{\min}) \\ \vdots & \cdots & \vdots \\ x_{3,3,1}(l_{\max}) & \cdots & x_{3,3,n}(l_{\max}) \end{pmatrix}, B = \begin{pmatrix} D_{1,1}(l_{\min}) \\ \vdots \\ D_{1,1}(l_{\max}) \\ D_{1,2}(l_{\min}) \\ \vdots \\ D_{1,2}(l_{\max}) \\ D_{1,3}(l_{\min}) \\ \vdots \\ D_{1,3}(l_{\max}) \\ D_{2,2}(l_{\min}) \\ \vdots \\ D_{2,2}(l_{\max}) \\ D_{2,3}(l_{\min}) \\ \vdots \\ D_{2,3}(l_{\max}) \\ D_{3,3}(l_{\min}) \\ \vdots \\ D_{3,3}(l_{\max}) \end{pmatrix}$$

Note that the set of possible lag times that is the domain of the function $x_{i,j,n}$ is also the domain of the function $D_{i,j}$.

By construction this formulation has the property that time-shifting one of the sounds (e.g., by inserting some amount of silence at the beginning) will automatically shift every x by the proper amount.

One weakness of this formulation is that it doesn't provide a mechanism to model the pairwise alignment penalty terms described in Section 3.4.4.2 (page 48).

5.1.4 Orthogonality Checking

Early attempts at regression failed because my matrix A (whose columns are individual detection functions) was rank deficient. In an attempt to gain some insight into the relationship among various detection functions on these sounds, I ran the following tests.

First I looked for completely constant functions of time and rejected them outright as obviously containing no information.[140] Then I used the vector cosine to compute the collinearity[141] between each pair of detection functions, rejecting one of the pair whenever the collinearity was below 0.001. Finally, I projected[142] each remaining detection function onto the space spanned by all of the other detection functions; when the mean squared difference between the function and its projection is low it means that a linear combination of all the other detection functions can closely approximate the given function, so this detection function is redundant in a regression context.

After these two passes (which in early tests tended to eliminate about 10% of the detection functions) I was always able to proceed with the regression.

5.1.4.1 Most and Least Orthogonal Pairs of Detection Functions

I computed all 86 detection functions for the 17 single-event sounds used in my listening experiment. (In effect, I concatenated all 17 sounds in the time domain, with enough zero-padding between them to ensure that every STFT frame will include energy from only one sound, then computed all detection functions on the 17-sound sequence.)

Low-Frequency Energy, Spectral Flatness, and their halfwave rectified first-order differences all had to be removed because some input frames (for the short clicks) are completely silent. Of the remaining detection functions, all had a collinearity of at least 0.025π with every other detection function, as shown in Table 25 and Table 26, so the vector cosine test did not discard any detection functions.

On the other hand, the projection test found that two of the detection functions were almost linear combinations of the others. One was Spectrum Energy, which by definition is exactly equal to the sum of the twelve Chroma Energy functions described in Section 5.1.1.9. The other is HFCC (the "sum of frequency weighted by energy" flavor of "high frequency content"). Examining the weights chosen by the projection algorithm reveals that (for these 17 sounds), HFCC is almost equal to 0.62 times Spectrum Energy plus 0.5241 times HFFCC, plus small

[140] An example of this is Spectrum Max Magnitude Frequency, which for many of Grey's tones is always the fundamental frequency.

[141] Actually, I computed the collinearity, which ignores the sign of the direction of each vector, so that it goes from 0 (meaning the two vectors are linearly proportional to each other, i.e., parallel) to $\pi/2$ (meaning the two vectors are completely orthogonal in the multidimensional space): http://ccrma.stanford.edu/~jos/mdft/Vector_Cosine_I.html

[142] http://ccrma.stanford.edu/~jos/mdft/Projection_I.html

weights (absolute value never more than 0.06 and usually under 0.01) times all the other detection functions.

C	*Detection Function 1*	*Detection Function 2*
0.5	MFCC 13	halfwave(diff(chroma-G))
0.49998	MFCC 13	halfwave(diff(zcr))
0.49995	MFCC 7	halfwave(diff(spectrum-hffcc))
0.49995	chroma-A	halfwave(diff(MFCC 9))
0.49995	halfwave(diff(spectrum-energy-norm-db))	halfwave(diff(aubio-complex))
0.49994	halfwave(diff(chroma-Bb))	halfwave(diff(spectrum-max-mag-freq))
0.49992	halfwave(diff(spectrum-max-mag-freq))	halfwave(diff(MFCC 11))
0.49989	MFCC 9	halfwave(diff(zcr))
0.49988	MFCC 12	halfwave(diff(chroma-D))
0.49986	MFCC 4	halfwave(diff(MFCC 12))
0.49985	halfwave(diff(chroma-G))	halfwave(diff(spectrum-max-mag-freq))
0.49984	MFCC 7	halfwave(diff(MFCC 2))
0.49984	MFCC 5	halfwave(diff(MFCC 9))
0.49983	halfwave(diff(chroma-E))	halfwave(diff(MFCC 12))
0.49982	spectrum-geometric-mean	MFCC 6
0.49982	halfwave(diff(chroma-E))	halfwave(diff(spectrum-max-mag-freq))
0.4998	halfwave(diff(spectral-centroid))	halfwave(diff(MFCC 6))
0.49978	MFCC 12	halfwave(diff(spectrum-max-mag-freq))
0.49978	halfwave(diff(spectrum-max-mag-freq))	halfwave(diff(aubio-complex))
0.49977	halfwave(diff(spectrum-max-mag-freq))	halfwave(diff(aubio-phase))

Table 25: The twenty most orthogonal pairs of detection functions for the 17 single-event sounds used in the listening experiment.

"C" is the collinearity between the two vectors, divided by pi.

C	*Detection Function 1*	*Detection Function 2*
0.025459	spectrum-energy	spectrum-spread
0.033428	halfwave(diff(spectrum-energy))	halfwave(diff(spectrum-spread))
0.039343	chroma-D	spectrum-energy
0.03992	spectrum-hfcc	spectrum-hffcc
0.03998	chroma-C	chroma-Db
0.04507	chroma-D	chroma-Eb
0.046348	chroma-D	spectrum-spread
0.047311	chroma-Eb	spectrum-energy
0.047533	spectrum-energy	spectrum-hfcc
0.050125	chroma-Db	spectrum-energy
0.054868	chroma-Db	spectrum-hfcc
0.055695	chroma-D	chroma-E
0.056543	chroma-Eb	spectrum-spread
0.057911	spectrum-hfcc	spectrum-spread
0.058094	halfwave(diff(spectrum-energy))	halfwave(diff(spectrum-hfcc))
0.05839	chroma-Db	spectrum-spread
0.06266	chroma-Bb	chroma-G
0.063057	chroma-Eb	spectrum-hfcc
0.063814	chroma-E	chroma-F
0.066182	chroma-E	chroma-Eb

Table 26: The twenty least orthogonal pairs of detection functions for the 17 single-event sounds used in the listening experiment.

5.1.5 Regression and Machine Learning Future Work

It is time-consuming to obtain subjective ground truth PAT-pdf data from listening tests, so the amount of training data is quite small by the standards of machine learning. Therefore models with too many degrees of freedom (e.g., linear combinations of many dozens of detection functions or thousands of time-shifted detection functions) will tend to overfit. I believe that to get good results with predictive modeling of PAT (without an impractically large amount of training data) it will be necessary to perform some kind of feature selection to choose a small subset of the detection functions.

The integral of any PAT-pdf will be 1 since it's a probability density function.[143] Since the variances of intrinsic PAT-pdfs vary greatly, so then do the maximum values: low-variance PAT-pdfs will be very narrow and tall, while high-variance PAT-pdfs will be wide and short. This presents a challenge to the regression techniques, since none of the detection functions will have such a wide difference in magnitude between impulsive and less impulsive sounds. Perhaps some form of scaling or preprocessing could alleviate this problem.

In addition to the straightforward linear regression that I have tried so far, more sophisticated forms of regression might perform better, including linear regression restricted to positive weights, logistic and other forms of nonlinear regression, and kernel ridge regression.

5.2 Nonparametric Modeling of PAT-pdf

All five of the intrinsic PAT-pdf estimation algorithms described in Section 3.5.5 (page 56) assume that intrinsic PAT-pdf is normal, but Section 0 (page 93) indicates that many of the intrinsic PAT-pdfs must not be normal. Getting from PAT measurements to shapes of intrinsic PAT-pdf curves in a completely nonparametric way (in other words, without making assumptions about the PAT-pdfs fitting particular statistical distributions) brings up all the same issues of allocating observed variance among the two sounds' intrinsic PAT-pdf variances and the pairwise alignment penalties, plus additional signal processing issues to find the exact shapes of the distributions.

5.2.1 Nonparametric Model

Given a set of samples of the random variable *D*, we can use *nonparametric kernel density estimation* to estimate the shape of *D*'s pdf without making any assumptions about *D* fitting a particular statistical distribution (Martinez and Martinez 2002, 280-285). This gives us an estimate of *D*'s

[143] Actually, this is only true for single-event sounds that will be perceived as one discrete event with probability 1.

pdf in the form of a sampled table, as illustrated in the top plot of Figure 16 (page 46). Let's use the notation *pdf(x)* to mean "the (sampled) shape of the estimated probability density function of the random variable *x*."

Ignoring the penalty terms for the moment, $D_{T,R}=I_T\text{-}I_R$, so we know that $pdf(D_{T,R})$ is the cross-correlation of the two intrinsic PAT-pdfs $pdf(I_T)$ and $pdf(I_R)$:

$$pdf(D_{T,R}) = pdf(T)\bigstar pdf(R)$$

(This follows from the assumption that the random variables T and R are independent.)

We can use signal-processing methods to estimate the underlying intrinsic PAT-pdfs. The basic problem is the following: given $pdf(D_{T,R})$ in the form of a sampled signal, find signals $pdf(T)$ and $pdf(R)$ such that $pdf(D_{T,R}) \approx pdf(T)\bigstar pdf(R)$.

The penalty terms complicate this so that the problem becomes the following: given an estimate of $pdf(D_{T,R})$ in the form of a sampled signal, find signals $pdf(T)$, $pdf(R)$, and $pdf(Penalty_{T,R})$ such that $pdf(D_{T,R}) \approx (pdf(T)\bigstar pdf(R)) * pdf(Penalty_{T,R})$. We can assume that $Penalty_{T,R}$ is zero-mean Gaussian noise.

5.2.2 Deconvolution of x*x with Spectral Square Root

As in Section 3.5.3 (page 50), we can look at trials aligning two copies of the same sound *S* to try to find the shape of the intrinsic PAT-pdf for that sound.

If we assume that $pdf(I_S)$ is symmetric in time, then we can set the zero point of our time axis to be the center of $pdf(I_S)$ so that the left/right flip that differentiates cross-correlation from convolution does nothing, in which case

$$pdf(D_{S,S}) = pdf(S)\bigstar pdf(S) = pdf(S)*pdf(S)$$

Since $x*x$ (convolved with itself) is equivalent to XX (with X meaning the spectrum of x), then given as input a signal y assumed to be $x*x$, we first find its spectrum Y with a discrete Fourier transform (Smith 2007b). Now we want to find $X=\sqrt{Y}$. What makes this complicated is that Y is complex, so every element (in other words, each FFT bin) has two square roots. Therefore there are $2^{\mathrm{Nfft}/2}$ possible spectra X such that $XX=Y$. The trick is that for each bin we choose the square root that results in the least difference in phase between this bin and the previous bin, so that the end result will have a maximally “smooth” spectrum.[144] Taking an IFFT of this “smooth” spectrum produces *x*.

5.2.3 Deconvolution of x*noise in the Frequency Domain

Since the spectrum of any Gaussian (e.g., a Gaussian noise penalty term) is another Gaussian[145], we can “sharpen” a signal to remove any given amount of variance by dividing its spectrum by the appropriate Gaussian spectrum.

5.2.4 Spectral Decorrelation of x★x?

Since *x*★*x* (the cross-correlation of *x* with *x*) is equivalent to $X\overline{X}$, how can we find *x* given *x*★*x*? First of all, note that for any *X*, $X\overline{X}$ will be real, so *x*★*x* must be symmetric, which makes sense as discussed above.

Unfortunately, there are too many degrees of freedom. We’re given *y*, a symmetric distribution for $D_{S,S}$. We’ll call *y*’s spectrum *Y*; *Y* must be real since *y* must be symmetric. Now we must find *X* such that $X\overline{X} = Y$. Let each X_i be $a + bj$.

$$Y_i = X_i\overline{X}_i = (a+bj)(a-bj) = a^2 + abj - abj - b^2 j^2 = a^2 + b^2$$

Each Y_i is real, so there are infinitely many choices for *a* and *b* such that $Y_i = a^2 + b^2$.

Of the infinitely many *x* such that *x*★*x* matches a given *y*, one approach is to choose *x* with minimum phase, in which case this is the well-known problem of *spectral factorization* (Sayed and Kailath 2001).

5.2.5 Generalization of Normal Parameter Estimation Algorithms to Nonparametric Case

Some of the steps of our algorithms for estimating Gaussian model parameters have equivalents in the nonparametric case, as Table 27 shows.

[144] Thanks to Julius Smith for coming up with this solution.

[145] See http://ccrma.stanford.edu/~jos/sasp/Gaussian_Window_Transform.html

Gaussian Model	***Nonparametric Model***
Estimating $\hat{\mu}_S$	Time-shifting *pdf(S)* so its centroid is $\hat{\mu}_S$
$\hat{\sigma}^2_{unknown} = \sigma^2_{D_{unknown,known}} - \hat{\sigma}^2_{known}$	Decorrelation to solve *pdf(D*$_{unknown.known}$*) =pdf(known))*★*pdf(unknown)* for *pdf(unknown)*
$var(I_S) \triangleq var(D_{S,S})\ /\ 2$	Spectral Decorrelation of x★x

Table 27: Correspondance between steps in estimating parameters of a Gaussian model of PAT-pdf and steps in estimating PAT-pdf nonparametrically.

5.3 Applications

5.3.1 Scheduling with PAT-pdf

The main practical use of models of PAT and P-Center is to synthesize music or speech with a desired rhythm by concatenating individual sound events. By representing PAT as PAT-pdf we gain two additional benefits over accounting for PAT as a single instant. First, we get a sense of the rhythmic tolerance inherent in each sound: a click might sound rhythmically different if moved by a few milliseconds, while the rhythmic contribution of another sound might sound exactly the same with the same shift.

More subtly, use of PAT-pdf allows us to adjust the rhythmic feel of a sequence of sounds. Sound example *cycle-noPATcorrection* plays the sequence of sounds listed in Table 28 with the *physical onset* of the sounds spaced exactly 400 ms apart. The result sounds somewhat incorrect rhythmically, because the PATs are not equally spaced. Sound example *cycle-PATmean* is the same sequence of sounds, but scheduled so that the means of the sounds' PAT-pdfs (according to the results shown in Table 23 of the "Variances from Trials Against Self & Maximum Likelihood Means" algorithm described in Section 3.5.5.4) are spaced every 400 ms. The difference between these two sound examples indicates that PAT matters at all. The added expressive power of the PAT-pdf formulation comes from considering also the standard deviations of each sound's PAT-pdf (namely, the square root of the variance shown in the "alg. 7" column of Table 24). Sound example *cycle-PAT+1STD* is just like the previous two examples except that it schedules the sounds so that the times of the mean plus one standard deviation of the PAT-pdf will be isochronous; sound example *cycle-PAT+1STD* is the same but making the times of the mean *minus* one standard deviation of the PAT-pdf be isochronous. To my ears, although the 400 ms period gives the same perceptual tempo in all four examples, *cycle-PAT-1STD* has a "laid-back," rhythmically "easy" feel, while *cycle-PAT+1STD* has a more "pushing," "on top of the beat" feel.

Ideal impulse
Ideal impulse
Ideal impulse
Ideal impulse
Clarinet
Ideal impulse
Clarinet SMC23
Ideal impulse
Clarinet SMC6
Ideal impulse
Snare
Ideal impulse
Snare SMC3
Ideal impulse
Ideal impulse
Ideal impulse
Trumpet
Ideal impulse
Trumpet SMC23
Ideal impulse
Violin
Ideal impulse
Violin SMC23
Ideal impulse
Violin SMC6
Ideal impulse

Table 28: Sequence of sounds in the "cycle" sound examples.

Sound examples *samba-noPAT, samba-PATmeans, samba-PAT-1STD*, and *samba-PAT+1STD* have the same relationship to each other as the previous four sound examples, but this time following the "score" shown in Figure 63. Again the *noPAT* version sounds wrong, the *PATmeans* version sounds correct, and the plus and minus one standard deviation versions have "forward leaning" or "backward leaning" rhythmic feels.

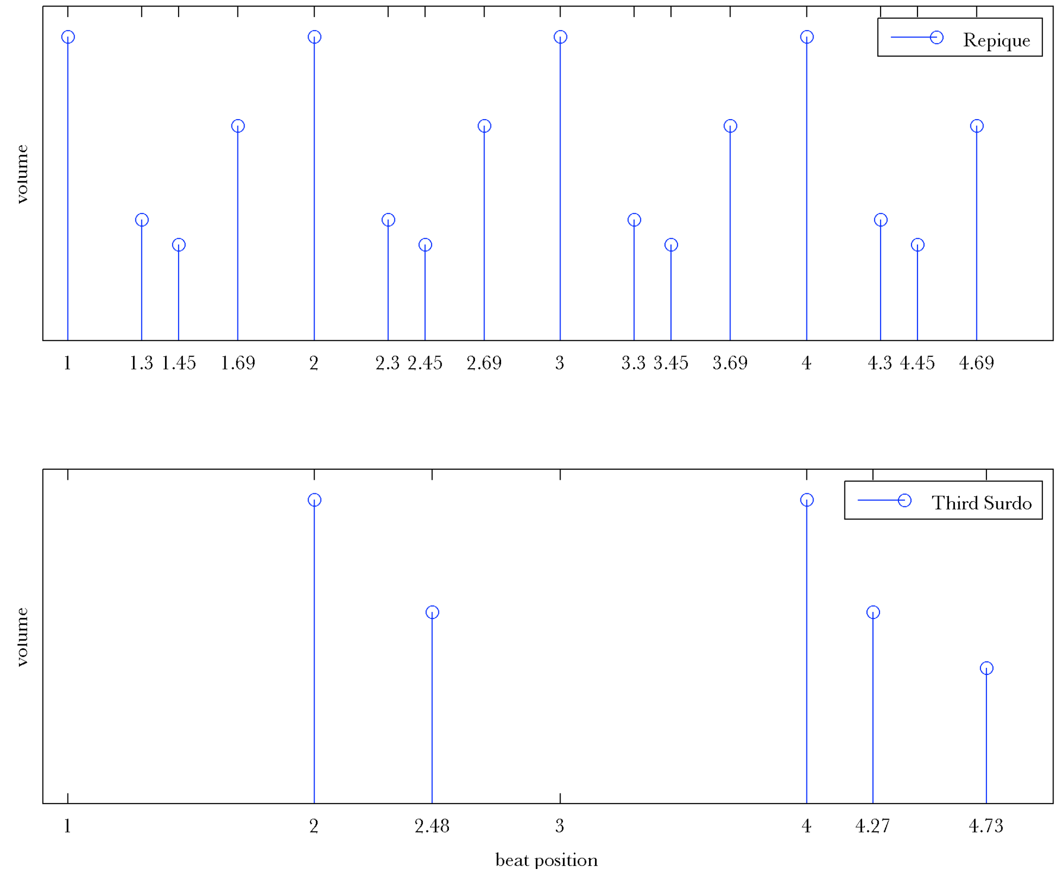


Figure 63: Typed-in "expressive" timing and event amplitudes for two of the parts of one version of the Samba Batucada rhythm.

More work is required to create useful and flexible tools for scheduling rhythmic sequences according to different points in each sound's estimated PAT-pdf curves.

5.3.2 Taking Advantage of Pair-Specific Alignment Difficulty

Throughout section 3.5 the "penalty term for each pair of sounds" representing the difficulty of perceiving the rhythmic alignment of sounds in separate audio streams was a source of complexity and difficulty. However, I believe it is possible to take advantage of this aspect of perception in compositionally interesting ways. Sounds A and B might be in one auditory stream with sounds C and D in another. A rhythm between sounds A and B will be perceived more exactly, possibly allowing for a larger number of distinct categories, or a clearer ability to hear microtiming subtlety. If sound B gradually morphed to sound C, then it would lose not only its timbral and registral similarity to sound A, but also perhaps its "rhythmic affinity" to sound A; details of timing might be less apparent, and the rhythm might give the same perceptual impression even if the timing of the rhythm played on C adjusted slightly so as to have a different relationship to the rhythm played on D.

5.3.3 Design of New Interfaces for Musical Expression

In the NIME field, there is much discussion of latency and jitter between the time of musicians' physical motion and the time of the corresponding sonic effect in the audio output of the "interface" (Chafe and Gurevich 2004; Lago and Kon 2004; Wessel and Wright 2002; Wright 2002; Wright, Cassidy, and Zbyszynski 2004). Musicians can easily learn and adjust to latency as shown in Figure 9 (page 24), but jitter replaces the possibility of expressive microtiming with uncontrolled random temporal results. The engineering question is "how much latency and jitter can a device have without compromising the possibility of expressive timing?"; an understanding of PAT-pdf could provide a way to quantify the way in which the answer to this question depends on the specific sounds that the device will make.

On a more abstract level, I believe that expert musicians control not only the temporal placement of the PAT of each sound event, but also in some case perhaps the entire shape of the PAT-pdf. For percussive instruments this may be a bit far-fetched, but for the singing voice and for continuously bowed or blown instruments, skilled performers can certainly produce events that are rhythmically distinct to varying degrees, and possibly even specifically shaped PAT-pdfs such as plateau, bimodal, etc. Perhaps a truly rhythmically expressive musical interface is one that allows the performer not just to control PAT with imperceptibly low jitter, but also allows the performer to control these shapes.

5.3.4 Use of PAT and PAT-pdf in Computer-Assisted Rhythmic Analysis

Analysis of the timing of recorded music generally considers the times of automatically detected note onsets. But as Section 2.5 (page 23) asserts, musicians control the timing of events' PATs rather than onsets. Incorporating models of PAT into this kind of analysis will allow exploration of the *perceived* rhythm of a given musical recording.

Appendix A Software for Administering the Listening Experiment

Prior work with online listening experiments (Cox 2007; Disley 2006; Disley, Howard, and Hunt 2006; Honing and Ladinig 2008) fits a paradigm in which experimenters create a small set of fixed sound files which are downloaded or streamed to subjects' computers, and then the subjects listen to these sounds and enter some form of multiple-choice response via simple graphical user interfaces. This level of interactivity would be enough to answer yes/no questions such as "do these two sounds seem like they're rhythmically together?"[146] However, I chose to use the standard method of adjustment for measuring PAT (see Section 3.3), which required giving the user complete control of the relative timing between two fixed sounds. It would not be practical to pre-record every possible temporal alignment of each pair of sounds, nor to synthesize them on the fly from a central server and stream each option back to the user in real time; the only efficient network architecture is for the subject to download the individual sounds and then have software running locally on each subject's machine render the sounds in a user-adjustable temporal relationship.

Therefore I had to build custom software for administering this software in an environment with the following features:

- High quality audio output
- Fully general playback of sound samples with sample-accurate control of timing
- Ability to design a graphical user interface to control sound synthesis interactively
- Ability to send results back to me via the Internet
- Cross-platform
- Ability for subjects to run the software without having to pay for it

[146] In retrospect, it might have been better to design the experiment around that kind of question. In addition to being easier to implement software to administer such a test, that kind of task and its associated simpler interface would have been easier for subjects to understand. On the other hand, having a more challenging and potentially more interesting task could be considered a form of the "high-hurdle technique"(Reips 2002, 7) for biasing data collection towards more dedicated and serious volunteer subjects.

I chose Max/MSP (Zicarelli 1998).

This appendix describes the software in some detail with two overlapping goals. The first is to provide a more complete picture of what exactly happened as subjects took this experiment. The second is based on the idea that this software and some of its components are generally useful beyond their use in this one experiment. The software itself is freely available online under the GNU Public License; this appendix serves as documentation of the software. At the time of publication of this dissertation, at least two other researchers have already used elements of this software for their own related experiments. I have also used the signal-based sample-accurate tap time input (described in section A.2.2) for other research involving the exact microtiming of the notes of Afro-Cuban *clave* patterns.

A.1 Principles in the Design of the Listening Experiment Software

This section lists the principles that guided the design of the software for this experiment and how I applied them.

A.1.1 Do not hurt the subjects

Compared to many other forms of research with human subjects, listening experiments are relatively safe. However, I was concerned about two main risks to subjects' health and well-being. To avoid *hearing loss* from excessive sound loudness (Rossing, Moore, and Wheeler 2002, 717-722) there is a volume calibration step in which the volume begins at zero with instructions to the subject to adjust it slowly upward until reaching a comfortable listening level. To avoid repetitive strain injury, vision problems, and related computer-use ailments from excessive time spent on the computer taking the experiment, there is a feature that enforces a 60-second break every 15 minutes, during which time audio pauses and the user sees the message "Please stretch and/or look far away for the next 53 seconds," with the number of seconds counting down until the end of the break, at which point the message becomes "I hope you enjoyed or are still enjoying your break" and the subject is once again able to turn on audio.[147] Figure 64 shows the implementation of the scheduling of breaks and Figure 65 shows how the breaks themselves were implemented.

[147] One subject commented on this feature by email: "The forced pause happened at exactly the right time!! I totally needed a break, but I wasn't aware of it."

Another feature addressing both of these concerns is that the user can pause the experiment at any time for as long as desired.

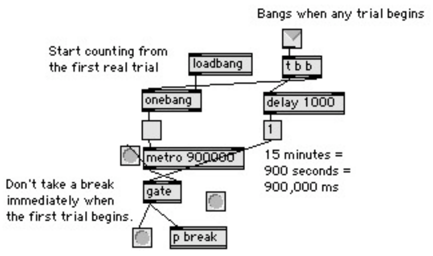


Figure 64: Sub-patch for scheduling breaks

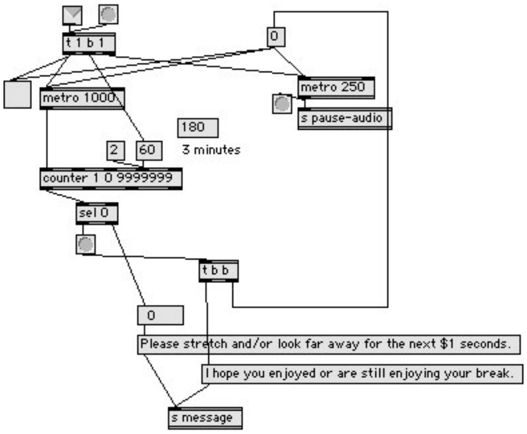
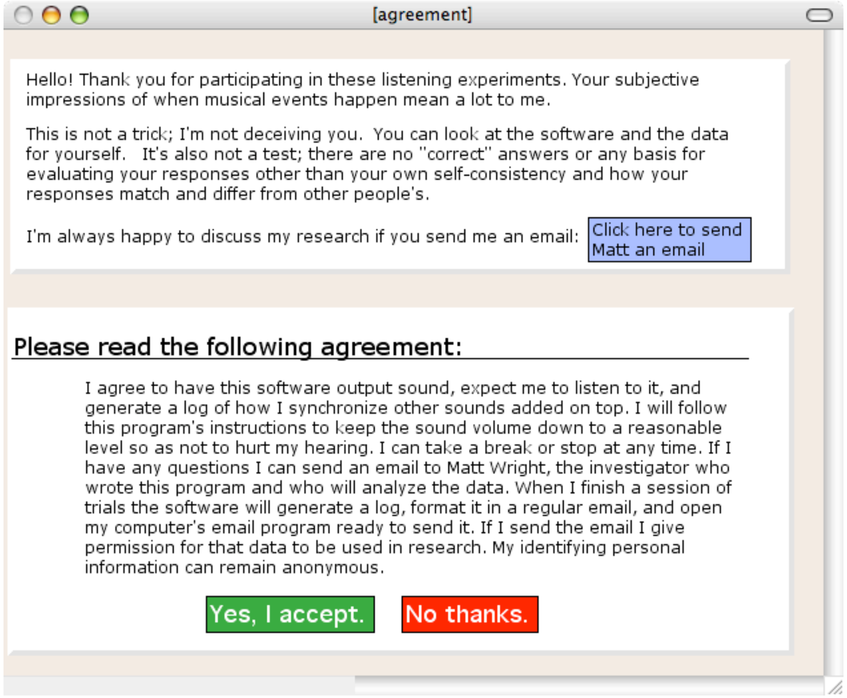

t 1 b 1
0
metro 1000
metro 250
s pause-audio
180
2
60
3 minutes
counter 1 0 9999999
sel 0
t b b
0
Please stretch and/or look far away for the next $1 seconds.
I hope you enjoyed or are still enjoying your break.
s message

Figure 65: Sub-patch for enforcing breaks. A "bang" received in the inlet causes a break to begin.

A.1.2 Get subjects' agreement to participate



[agreement]

Hello! Thank you for participating in these listening experiments. Your subjective impressions of when musical events happen mean a lot to me.

This is not a trick; I'm not deceiving you. You can look at the software and the data for yourself. It's also not a test; there are no "correct" answers or any basis for evaluating your responses other than your own self-consistency and how your responses match and differ from other people's.

I'm always happy to discuss my research if you send me an email: Click here to send Matt an email

Please read the following agreement:

I agree to have this software output sound, expect me to listen to it, and generate a log of how I synchronize other sounds added on top. I will follow this program's instructions to keep the sound volume down to a reasonable level so as not to hurt my hearing. I can take a break or stop at any time. If I have any questions I can send an email to Matt Wright, the investigator who wrote this program and who will analyze the data. When I finish a session of trials the software will generate a log, format it in a regular email, and open my computer's email program ready to send it. If I send the email I give permission for that data to be used in research. My identifying personal information can remain anonymous.

Figure 66 Opening "agreement" screen seen when the software begins.

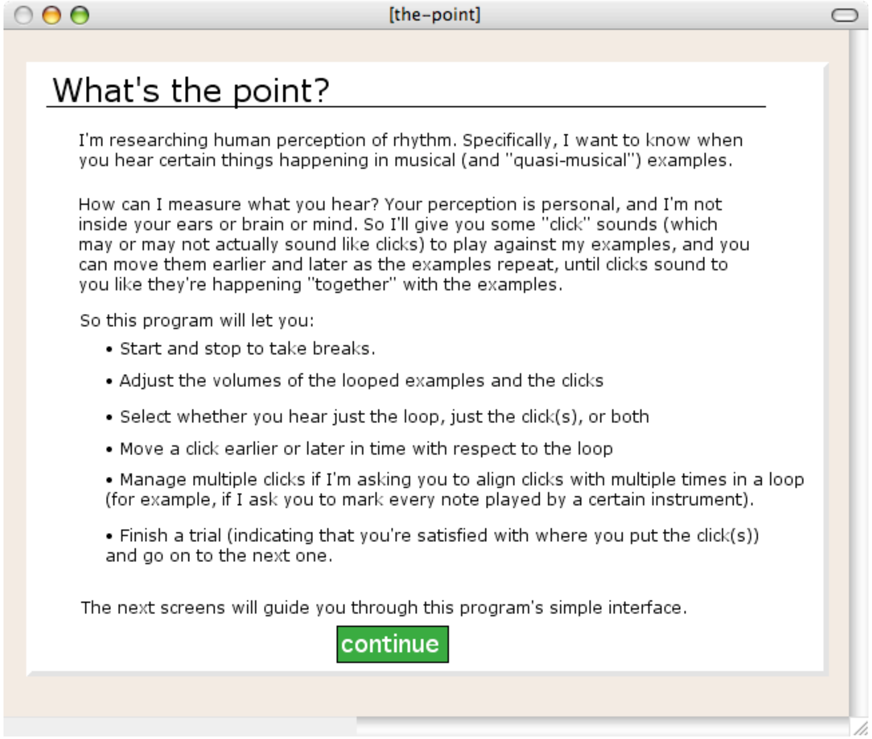
The process of downloading and installing this software was a major deterrent for any would-be subject who did not actually want to be a part of the experiment. However, the ethical principle of *informed consent* requires that experimenters disclose relevant information about the experiment to subjects, that experimenters make their best attempt to ensure that subjects understand that information, and that subjects voluntarily agree to participate in the research (Ryan et al. 1979). Therefore the software begins with the screens shown in Figure 66 and Figure 67. Clicking "No thanks" in the agreement[148] screen (Figure 66) quits the program. Clicking "Click here to send Matt an email" opens the subject's default email software[149] with a blank email message addressed to matt@ccrma.stanford.edu with the subject "Let's discuss your research."

Several would-be subjects (all musicians) whom I knew personally raised the unanticipated concern that this research would a "test" of their own temporal acuity, that somehow their own musicianship would be placed under scrutiny, hence the language "it's also not a test; there are no "correct" answers or any basis for evaluating your responses other than your own self-consistency and how your responses match and differ from other people's."

Subjects were allowed to stop the experiment at any time, and I encouraged them to send me however many trials they completed even if not the full 75 that I requested. (Figure 18 on page 79 shows the number of trials that each subject completed.) Subjects also had to manually email the results back to me, providing one final opportunity to withdraw participation in the experiment.

[148] Many experiments using human subjects begin with a step in which the subject consents to participate. For this experiment, Stanford's Institutional Review Board ruled that I was in fact exempt from this requirement because of the extremely low level of risk to the subject and the fact that the technical difficulties of downloading and running the experiment would deter any "potentially vulnerable subjects" such as those with impaired decision making. Therefore I had to be careful not to use the word "consent" anywhere in my software, because I technically was not getting consent.

[149] This works by opening the url "`mailto:matt@ccrma.stanford.edu?subject=Let's discuss your research.`" in the user's default web browser, by sending the "launchbrowser" message to "max."



What's the point?

I'm researching human perception of rhythm. Specifically, I want to know when you hear certain things happening in musical (and "quasi-musical") examples.

How can I measure what you hear? Your perception is personal, and I'm not inside your ears or brain or mind. So I'll give you some "click" sounds (which may or may not actually sound like clicks) to play against my examples, and you can move them earlier and later as the examples repeat, until clicks sound to you like they're happening "together" with the examples.

So this program will let you:

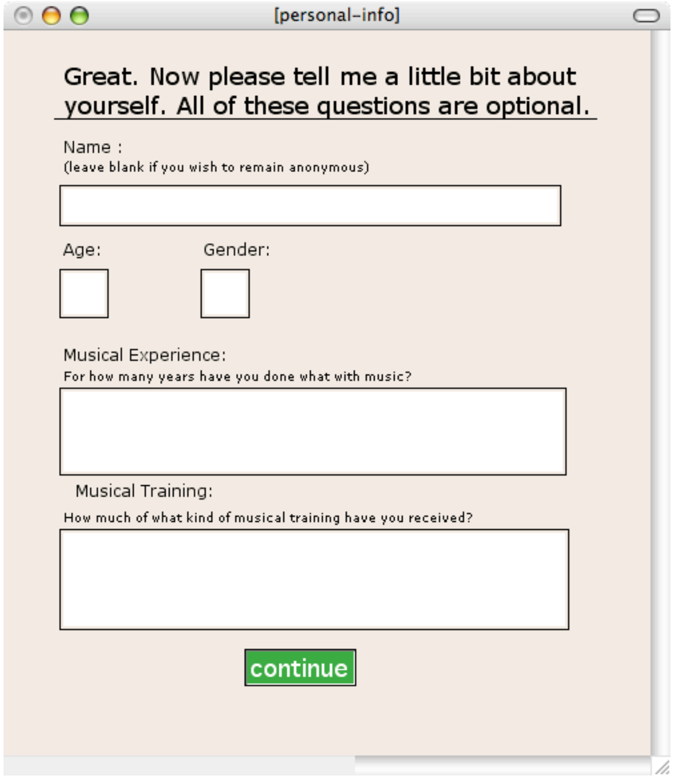
- Start and stop to take breaks.
- Adjust the volumes of the looped examples and the clicks
- Select whether you hear just the loop, just the click(s), or both
- Move a click earlier or later in time with respect to the loop
- Manage multiple clicks if I'm asking you to align clicks with multiple times in a loop (for example, if I ask you to mark every note played by a certain instrument).
- Finish a trial (indicating that you're satisfied with where you put the click(s)) and go on to the next one.

The next screens will guide you through this program's simple interface.

Figure 67: Second screen seen by subjects, giving the context of the research and an overview of the program's functions.

A.1.3 Allow anonymous participation

The software asks each subject for some personal information (as shown in Figure 68), emphasizing that the questions are optional.



Great. Now please tell me a little bit about yourself. All of these questions are optional.

Name :
(leave blank if you wish to remain anonymous)

Age:

Gender:

Musical Experience:
For how many years have you done what with music?

Musical Training:
How much of what kind of musical training have you received?

Figure 68: Screen for users to enter (optional) personal information.

The section "Sending Back Data from Completed Trials" below discusses technical aspects of enabling anonymity. I note that in spite of these allowances none of my subjects chose to participate anonymously; each provided his or her full name and email address. (Many, however, declined to state age and/or gender.)

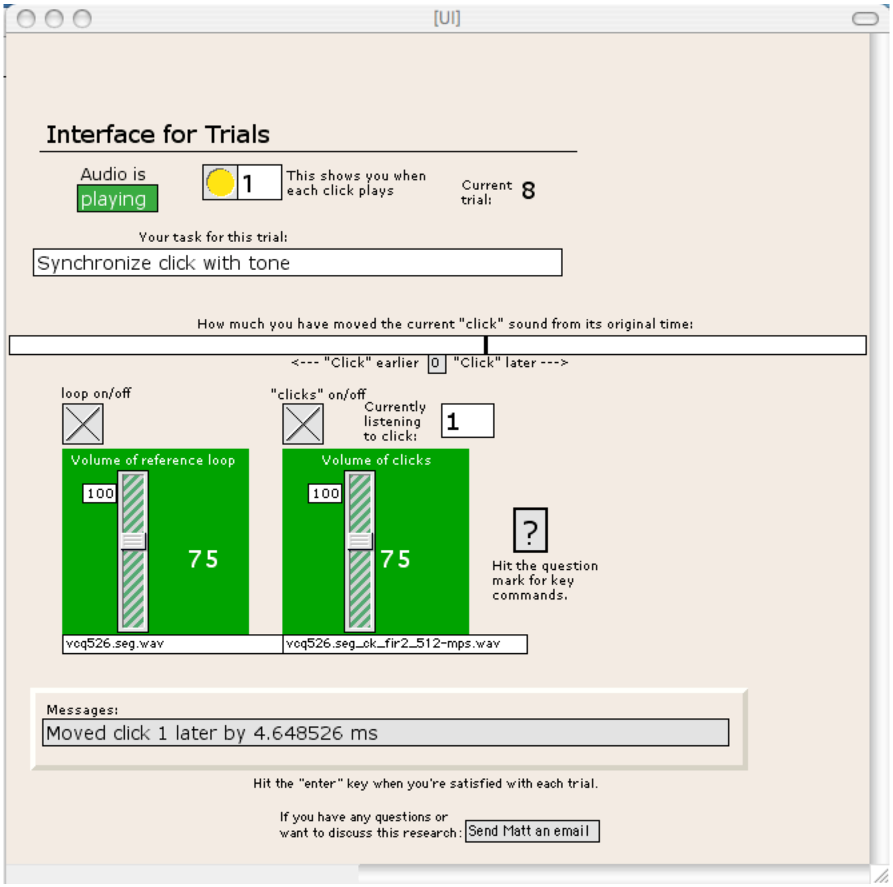
A.1.4 Make it fun to participate

"Make the experiment challenging and fun to do… Music lovers tend to like listening experiments and are usually very motivated, resulting in large numbers of responses" (Honing and Ladinig 2008, 5). Some subjects commented that they were fascinated with being able to hear the effect of extremely small temporal adjustments. One subject, in fact, was completely engrossed by this, and took over an hour per trial listening to hundreds of possible responses!

I attempted to make the experiment more enjoyable by making the text for the subjects to read be humorous and light-hearted yet succinct. I also alternated blocks of more musically interesting examples (based on looped metric examples from real music) with the dryer, more "scientific" examples (aligning isolated pairs of sounds).

A.1.5 Make it obvious what the task is

Figure 69 shows the main user interface for this software. The large text field labeled “Your task for this trial” contained a string such as “Synchronize two synthetic tones,” “Synchronize two clicks,” etc. (Table 7 on page 77 lists all such tasks.) The initial training segment of the software (described in the next section) gave a more detailed description of the task: “The goal of the experiment is to move the click so that it sounds like it's lined up exactly with the tone, i.e., to "synchronize" the click with the tone. You could pretend to be a conductor trying to get two musicians to play at the same time.” The analogy to conducting was inspired by Gordon’s instructions that subjects “pretend to be a conductor, trying to get the two "players" to perform exactly together on the beat” (Gordon 1987, 91).



Interface for Trials

Audio is playing

1 This shows you when each click plays

Current trial: 8

Your task for this trial:
Synchronize click with tone

How much you have moved the current "click" sound from its original time:
<--- "Click" earlier 0 "Click" later --->

Currently listening to click: 1

vcq526.seg.wav

vcq526.seg_ck_fir2_512-mps.wav

Hit the question mark for key commands.

Messages:
Moved click 1 later by 4.648526 ms

Hit the "enter" key when you're satisfied with each trial.

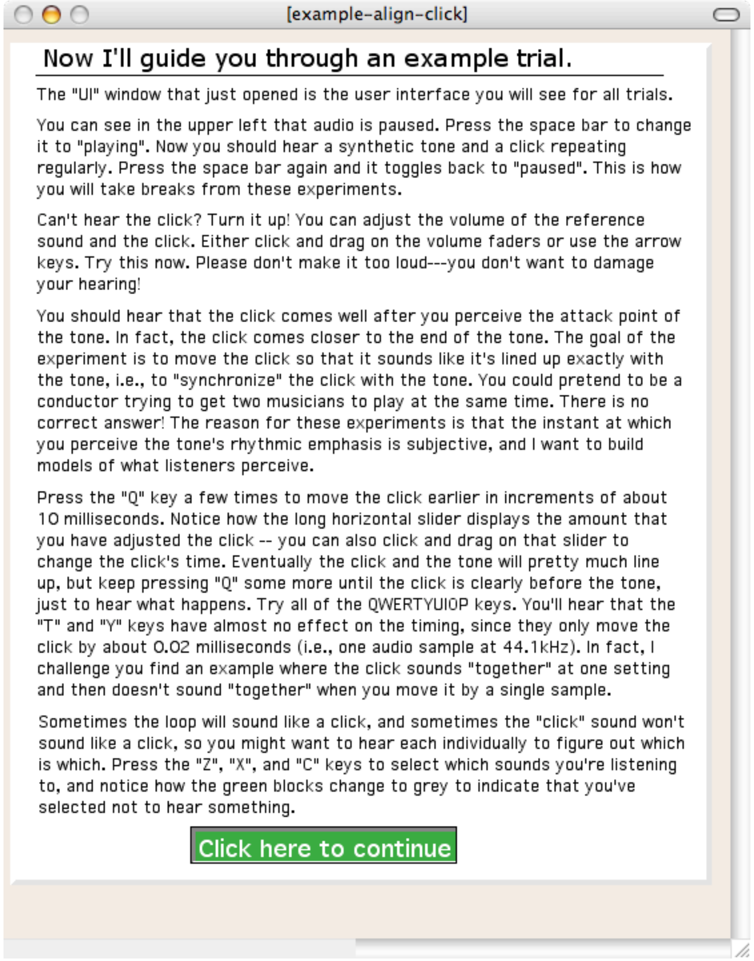
If you have any questions or want to discuss this research: Send Matt an email

Figure 69: Main user interface for the experiment

A.1.6 Train subjects how to do the task

Because this experiment is administered entirely by computer, the training takes place in the form of two “example” trials. Each of these trials put the software into the same state as for a real trial and guided the user step-by-step through the software features needed to perform the experiment.

The first example used the *Clarinet* and *Ideal impulse* sounds, starting from an initial alignment in which the impulse occurs 136 ms (6000 samples) after the physical onset of the clarinet. The clarinet sound has over 400 times as much energy as the single-sample impulse, so it's much louder than the impulse (so much so, in fact, that for some subjects the clarinet completely masked the impulse when presented at equal volumes); this motivates the teaching of the user interface for controlling volumes individually.



Now I'll guide you through an example trial.

The "UI" window that just opened is the user interface you will see for all trials.

You can see in the upper left that audio is paused. Press the space bar to change it to "playing". Now you should hear a synthetic tone and a click repeating regularly. Press the space bar again and it toggles back to "paused". This is how you will take breaks from these experiments.

Can't hear the click? Turn it up! You can adjust the volume of the reference sound and the click. Either click and drag on the volume faders or use the arrow keys. Try this now. Please don't make it too loud---you don't want to damage your hearing!

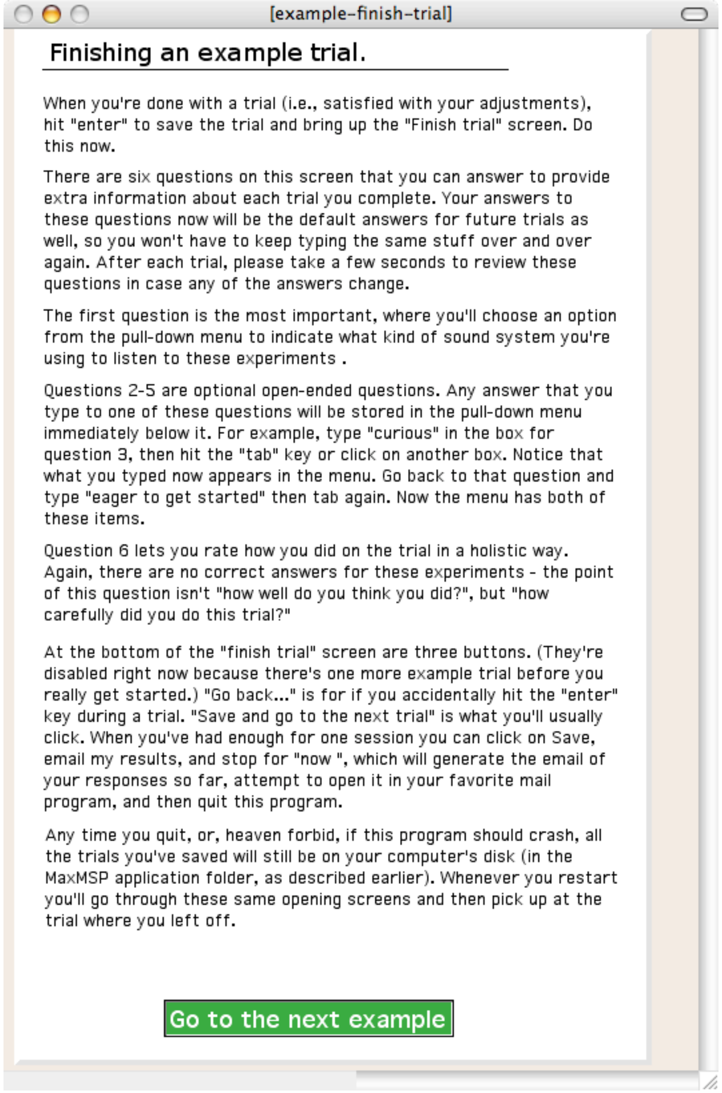
You should hear that the click comes well after you perceive the attack point of the tone. In fact, the click comes closer to the end of the tone. The goal of the experiment is to move the click so that it sounds like it's lined up exactly with the tone, i.e., to "synchronize" the click with the tone. You could pretend to be a conductor trying to get two musicians to play at the same time. There is no correct answer! The reason for these experiments is that the instant at which you perceive the tone's rhythmic emphasis is subjective, and I want to build models of what listeners perceive.

Press the "Q" key a few times to move the click earlier in increments of about 10 milliseconds. Notice how the long horizontal slider displays the amount that you have adjusted the click -- you can also click and drag on that slider to change the click's time. Eventually the click and the tone will pretty much line up, but keep pressing "Q" some more until the click is clearly before the tone, just to hear what happens. Try all of the QWERTYUIOP keys. You'll hear that the "T" and "Y" keys have almost no effect on the timing, since they only move the click by about 0.02 milliseconds (i.e., one audio sample at 44.1kHz). In fact, I challenge you find an example where the click sounds "together" at one setting and then doesn't sound "together" when you move it by a single sample.

Sometimes the loop will sound like a click, and sometimes the "click" sound won't sound like a click, so you might want to hear each individually to figure out which is which. Press the "Z", "X", and "C" keys to select which sounds you're listening to, and notice how the green blocks change to grey to indicate that you've selected not to hear something.

Figure 70: Instructions for the first example trial, demonstrating how to synchronize two isolated sounds.

After the subject clicks "Click here to continue," the window shown in Figure 71 appears.



Finishing an example trial.

When you're done with a trial (i.e., satisfied with your adjustments), hit "enter" to save the trial and bring up the "Finish trial" screen. Do this now.

There are six questions on this screen that you can answer to provide extra information about each trial you complete. Your answers to these questions now will be the default answers for future trials as well, so you won't have to keep typing the same stuff over and over again. After each trial, please take a few seconds to review these questions in case any of the answers change.

The first question is the most important, where you'll choose an option from the pull-down menu to indicate what kind of sound system you're using to listen to these experiments .

Questions 2-5 are optional open-ended questions. Any answer that you type to one of these questions will be stored in the pull-down menu immediately below it. For example, type "curious" in the box for question 3, then hit the "tab" key or click on another box. Notice that what you typed now appears in the menu. Go back to that question and type "eager to get started" then tab again. Now the menu has both of these items.

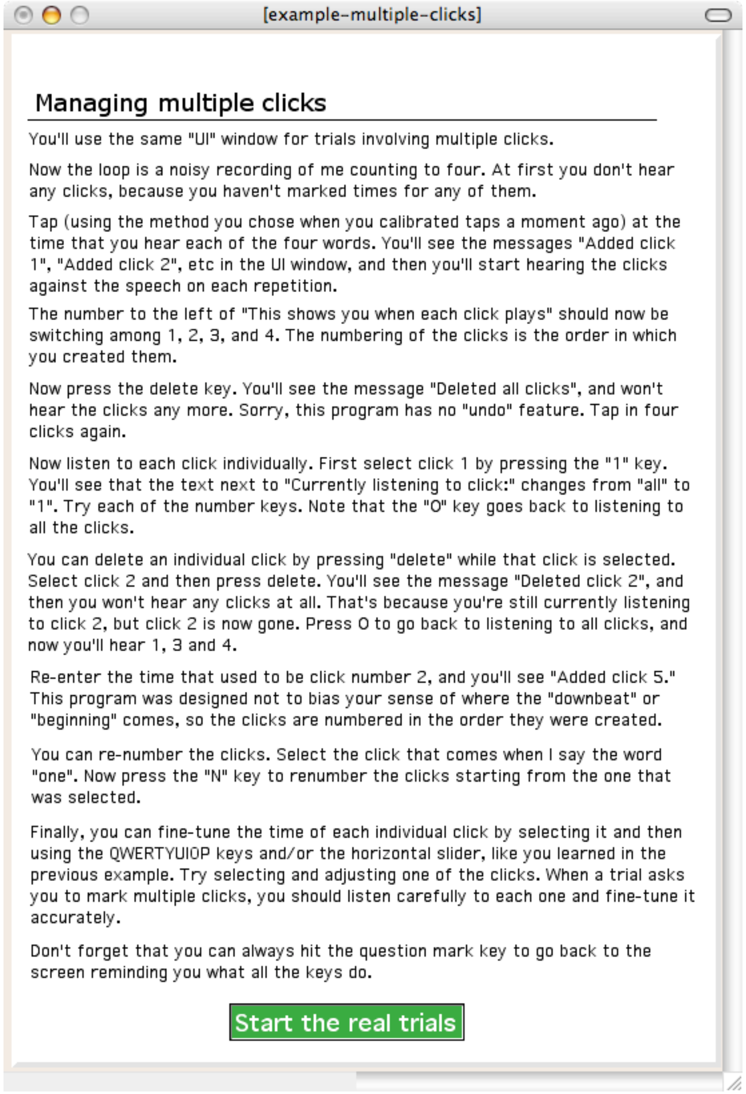
Question 6 lets you rate how you did on the trial in a holistic way. Again, there are no correct answers for these experiments - the point of this question isn't "how well do you think you did?", but "how carefully did you do this trial?"

At the bottom of the "finish trial" screen are three buttons. (They're disabled right now because there's one more example trial before you really get started.) "Go back..." is for if you accidentally hit the "enter" key during a trial. "Save and go to the next trial" is what you'll usually click. When you've had enough for one session you can click on Save, email my results, and stop for "now ", which will generate the email of your responses so far, attempt to open it in your favorite mail program, and then quit this program.

Any time you quit, or, heaven forbid, if this program should crash, all the trials you've saved will still be on your computer's disk (in the MaxMSP application folder, as described earlier). Whenever you restart you'll go through these same opening screens and then pick up at the trial where you left off.

Figure 71: Instructions for finishing a trial.

After the subject clicks "Go to the next example," the screen shown in Figure 72 appears, explaining how to enter multiple clicks, select and move each individually, and renumber them if necessary.



Managing multiple clicks

You'll use the same "UI" window for trials involving multiple clicks.

Now the loop is a noisy recording of me counting to four. At first you don't hear any clicks, because you haven't marked times for any of them.

Tap (using the method you chose when you calibrated taps a moment ago) at the time that you hear each of the four words. You'll see the messages "Added click 1", "Added click 2", etc in the UI window, and then you'll start hearing the clicks against the speech on each repetition.

The number to the left of "This shows you when each click plays" should now be switching among 1, 2, 3, and 4. The numbering of the clicks is the order in which you created them.

Now press the delete key. You'll see the message "Deleted all clicks", and won't hear the clicks any more. Sorry, this program has no "undo" feature. Tap in four clicks again.

Now listen to each click individually. First select click 1 by pressing the "1" key. You'll see that the text next to "Currently listening to click:" changes from "all" to "1". Try each of the number keys. Note that the "0" key goes back to listening to all the clicks.

You can delete an individual click by pressing "delete" while that click is selected. Select click 2 and then press delete. You'll see the message "Deleted click 2", and then you won't hear any clicks at all. That's because you're still currently listening to click 2, but click 2 is now gone. Press 0 to go back to listening to all clicks, and now you'll hear 1, 3 and 4.

Re-enter the time that used to be click number 2, and you'll see "Added click 5." This program was designed not to bias your sense of where the "downbeat" or "beginning" comes, so the clicks are numbered in the order they were created.

You can re-number the clicks. Select the click that comes when I say the word "one". Now press the "N" key to renumber the clicks starting from the one that was selected.

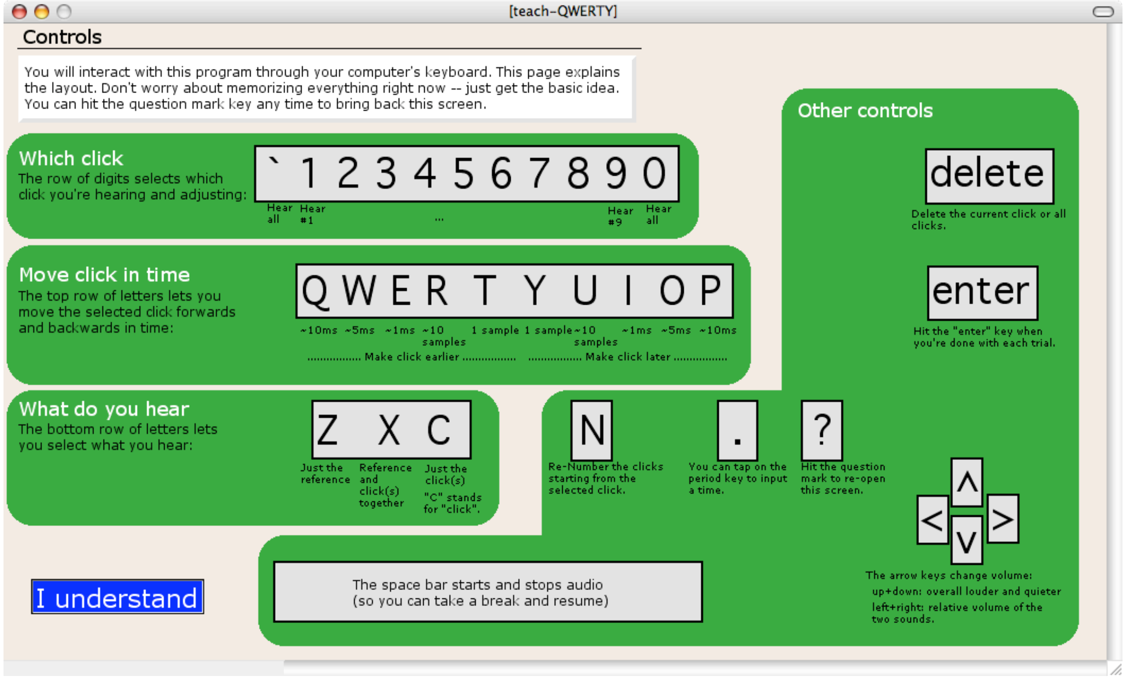
Finally, you can fine-tune the time of each individual click by selecting it and then using the QWERTYUIOP keys and/or the horizontal slider, like you learned in the previous example. Try selecting and adjusting one of the clicks. When a trial asks you to mark multiple clicks, you should listen carefully to each one and fine-tune it accurately.

Don't forget that you can always hit the question mark key to go back to the screen reminding you what all the keys do.

Figure 72: Instructions for the second example trial, explaining how to perform trials with multiple clicks.

All in all the interface uses 34 keys to select clicks individually or all together, to move the selected click in time, to hear only the loop, only the click, or both, to pause and resume the experiment, to adjust the relative and overall volume, to enter new clicks, and to finish each trial. Although these were laid out spatially in a relatively logical fashion[150] it would have been easy to forget details of the interface, so at any time the user pressed the question mark key, the window shown in Figure 73 would appear.

[150] At least the physical layout made sense on the American-style keyboard I used to design the user interface. Thanks to Laurent Daudet for reminding me that this layout becomes scrambled on various international keyboards.



Controls

You will interact with this program through your computer's keyboard. This page explains the layout. Don't worry about memorizing everything right now -- just get the basic idea. You can hit the question mark key any time to bring back this screen.

Which click
The row of digits selects which click you're hearing and adjusting:

Move click in time
The top row of letters lets you move the selected click forwards and backwards in time:

.... Make click earlier .... Make click later ....

What do you hear
The bottom row of letters lets you select what you hear:

Just the reference
Reference and click(s) together
Just the click(s) "C" stands for "click".

Re-Number the clicks starting from the selected click.

You can tap on the period key to input a time.

Hit the question mark to re-open this screen.

Other controls

Delete the current click or all clicks.

Hit the "enter" key when you're done with each trial.

The arrow keys change volume:
up+down: overall louder and quieter
left+right: relative volume of the two sounds.

The space bar starts and stops audio
(so you can take a break and resume)

Figure 73: Screen documenting the keyboard-based user interface for the software.

A.1.7 Test only what should be tested

I wanted to measure only subjects' auditory perception of sound, so there was no visual feedback such as a waveform display or numeric readout of the amount of time between the onsets of the two sounds. Figure 69 shows a slider labeled "How much you have moved the current "click" sound from its original time," but the "zero" point of this slider is the (randomly assigned) initial temporal relationship between the two sounds, not, for example, a delay of zero between the acoustic beginnings of the two sounds.

For trials in which the subject enters click times by tapping, I did not want to consider latency or jitter in the measurement of subjects' taps, or variance from subjects' motor noise, or the tendency to anticipate when tapping. (See Section 3.3.1 on page 35.) Therefore tap times entered in this way were then verified sonically by playing the reference sound at the measured time of each tap, with a subsequent phase for the subject to fine-tune the timing until it sounded correct.

A.1.8 Arrange trials in blocks

The software presented blocks of trials with the same task before moving to each new task, as shown in Table 7 (p. 77). A global counter kept track of the current trial number, which was the input to the trivial "num-trials-per-block" subpatcher shown in Figure 74.

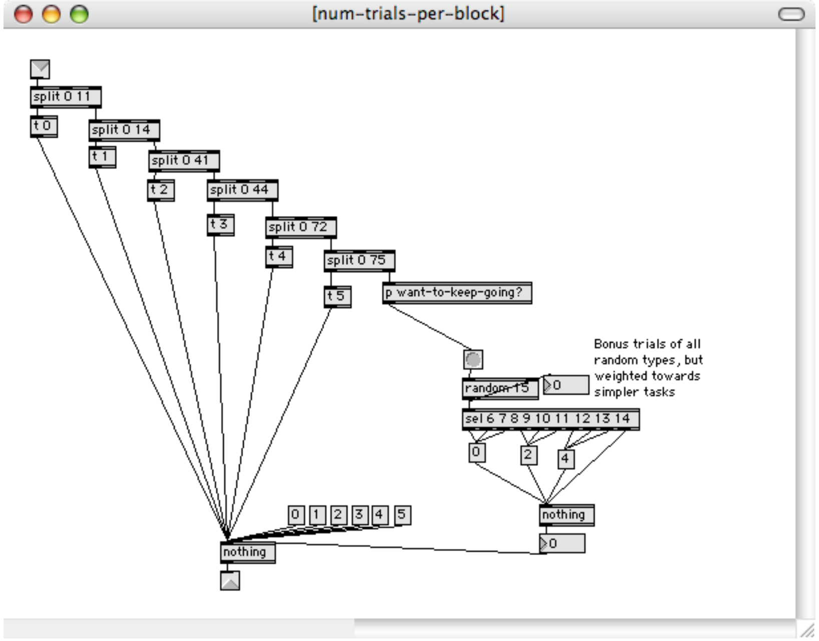
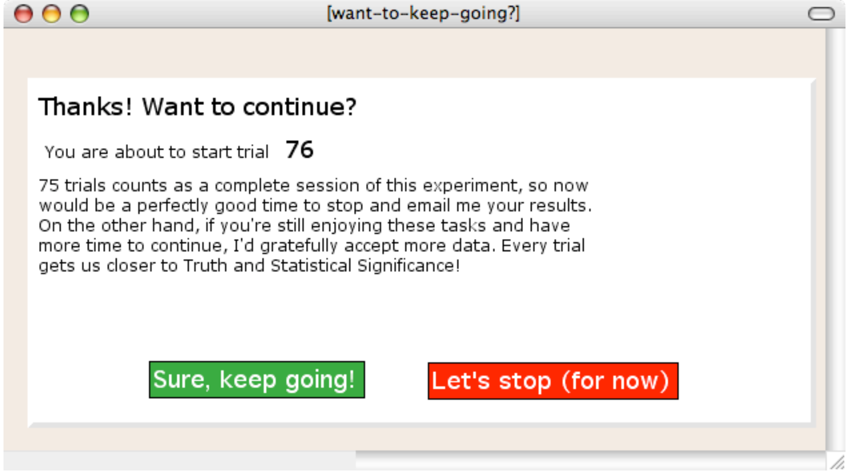


Figure 74: Subpatcher to determine the type of task for each trial, implementing the arrangement of trials into blocks and the option to continue after completing 75 trials.

After the end of the 75th trial (as well as the 85th, 95th, etc.) the software invokes the "want to keep going?" patcher, shown in Figure 75.



Thanks! Want to continue?

You are about to start trial 76

75 trials counts as a complete session of this experiment, so now would be a perfectly good time to stop and email me your results. On the other hand, if you're still enjoying these tasks and have more time to continue, I'd gratefully accept more data. Every trial gets us closer to Truth and Statistical Significance!

Figure 75: Screen informing the user that 75 trials were completed and offering the choice to stop or continue with extra trials.

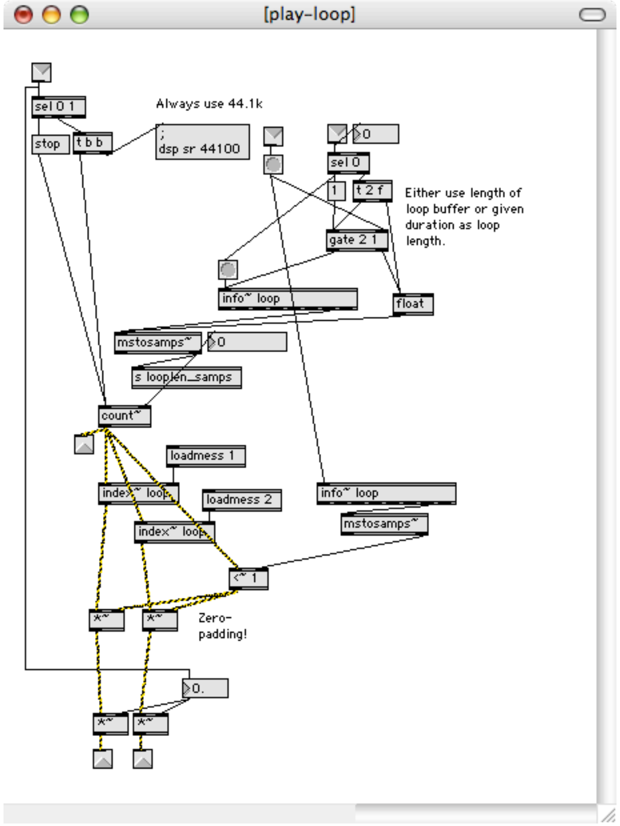
A.1.9 Randomize initial conditions

The software randomly selected reference and test sounds for each trial from the options for the appropriate block. Most importantly, for trials in which there is only a single sound to adjust, the initial time relationship between the test and reference sounds was random. For the first block of trials, which had two full-length instrumental tones, the initial offset was chosen uniformly from the range ± 340 ms. For blocks in which one of the sounds was a short click, the initial offset was chosen uniformly from the range ± 113 ms.

A.2 Challenges in the Implementation of Listening Experiment Software

This section details some of the technical issues I faced in building the software for these listening tests.

A.2.1 Sample-Accurate Playback Scheduling



Always use 44.1k

Either use length of loop buffer or given duration as loop length.

Zero-padding!

Figure 76: Subpatcher for playing the looped "reference" sound and outputting a synchronization signal.

This software provides the subject with control of the relative timing of sounds down to a single digital audio sample (approximately 0.02 ms). Achieving this kind of temporal accuracy in

Max/MSP requires bypassing the normal event-domain scheduling features and controlling timing completely in the audio domain.[151]

The “master clock” for the software comes from the “play loop” subpatcher (shown in Figure 76). The “count~” object outputs a signal consisting of consecutive integers from zero up to the specified maximum (in this case, the period, in samples, of the repeating sound). This signal goes out the left outlet to control the scheduling of the test sounds (described below) and also goes to a pair of index~ objects (for the two channels of the potentially stereo reference sound). Note that MSP’s “index~” object outputs the last sample of the audio buffer for any input value greater than the duration of the buffer (in other words, it outputs a DC offset after it finishes playing a sound), hence the construction with the “<~” object and the two multipliers to force the output signal to zero after exceeding the duration of the buffer.

The main data structure for the scheduling of click sounds is the collection “click-times”, which simply lists the time (in samples) that each click sound should begin with respect to the repeating reference loop. Here’s an example collection that defines the times for four clicks:

```
1, 12485;
2, 41861;
3, 68677;
4, 96901;
```

The software allows the user to enter up to nine clicks (because it uses the keys 1-9 to select each click individually). For trials involving only a single test sound the multiple-click features are disabled.

A poly~ object containing nine instances of the “clicker voice” patch shown in Figure 77 synthesizes the click sounds. This patch is important enough to warrant describing in detail:

1. The “r click-times-changed” outputs whenever the contents of the “click-times” collection change.

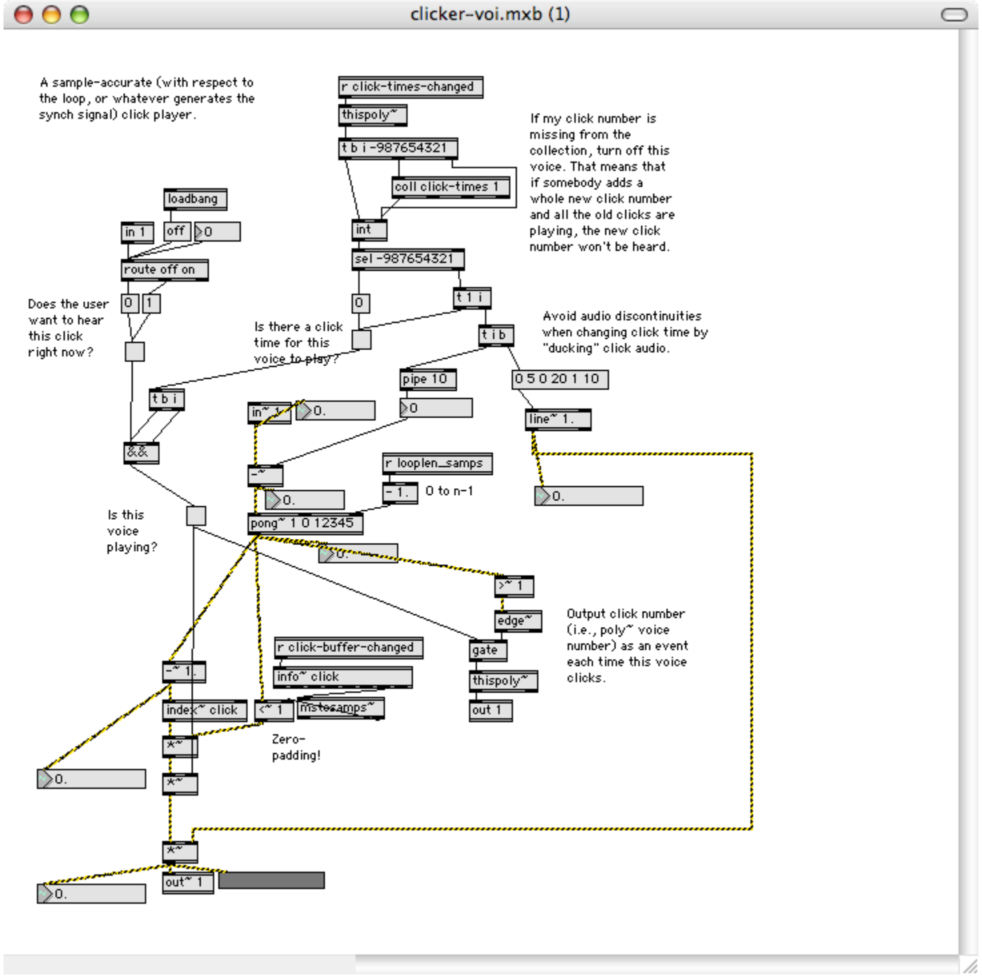
2. The “thispoly~” object outputs the voice number (1-9) of the “clicker voice” patch; each voice is responsible for producing only one click.

3. The construction with “t b i -987654321”, the collection and the “int” provides a default value of -987654321 in case this voice’s click number does not appear in the “click-times”

[151] See (Puckette 1991) for a discussion of the distinction between the event scheduler and the signal scheduler. The modern version of Max/MSP inherits this architecture.

collection. When the “sel” object receives this default value it disables the voice so no click is produced.

4. Otherwise, the time that this voice should play its click has changed. In order to avoid audio discontinuities when the subject changes the time of a click while it’s playing, there’s a ducking mechanism (implemented with the line~ object) that brings the gain of this click down to zero over 5 ms, holds it there for 20 ms (during which time the new click time passes through the “pipe 10” and to the “-~” object, changing the time the click will play), and then brings the gain back up to full volume over another 10ms.
5. The signal input to this poly~ object (and hence all 9 instances of “clicker voice”) is the synchronization signal produced by the count~ object in Figure 76. This signal comes in through the “in~ 1” object.
6. Subtracting the offset for this voice’s click from the global synchronization signal produces a new signal whose value is zero at precisely the sample when this voice’s click should play. The pong~ object wraps this signal to be between zero and the loop length of the reference sound.
7. The index~ object actually plays the click sound.
8. The construction with the info~, <~, and *~ objects is a workaround for the problem mentioned above about index~ outputting a DC offset after finishing the buffer.
9. The control inlet to this patch (which comes through the “in 1” object) determines whether this voice should sound (based on the user’s selection of whether to listen to all clicks or just a single click).
10. The control outlet of this patch (which goes out via the “out 1” object) outputs the current click number whenever this voice plays.



A sample-accurate (with respect to the loop, or whatever generates the synch signal) click player.

If my click number is missing from the collection, turn off this voice. That means that if somebody adds a whole new click number and all the old clicks are playing, the new click number won't be heard.

Does the user want to hear this click right now?

Is there a click time for this voice to play?

Avoid audio discontinuities when changing click time by "ducking" click audio.

Is this voice playing?

Output click number (i.e., poly~ voice number) as an event each time this voice clicks.

Zero-padding!

Figure 77: Abstraction (to be used inside poly~) for sample-accurate playback of click sounds.

A.2.2 Sample-Accurate Tap Time Input

The software supports two methods of entering tap times for multiple-click trials: via the QWERTY keyboard or via the audio input. The advantage of using the QWERTY keyboard is that it requires no configuration or calibration and the implementation is extremely simple (see Figure 78); the disadvantage is that the latency between physical key-press and the moment the software registers the key may be both long and variable (Wright, Cassidy, and Zbyszynski 2004). The audio input has the opposite properties: audio input latency is constant and potentially very small, but it is more difficult to configure.

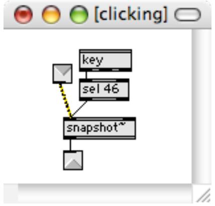
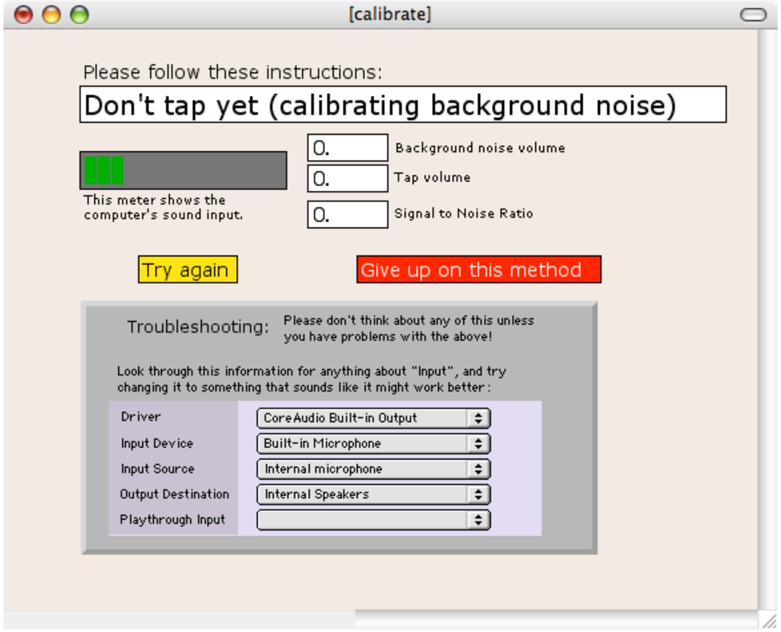


Figure 78: Sub-patch for finding times when the subject taps on the period key.

The inlet of this patch is connected to the synchronization signal produced by the count~ object in Figure 76


[calibrate]

Please follow these instructions:

Don't tap yet (calibrating background noise)

0.	Background noise volume
0.	Tap volume
0.	Signal to Noise Ratio

This meter shows the computer's sound input.

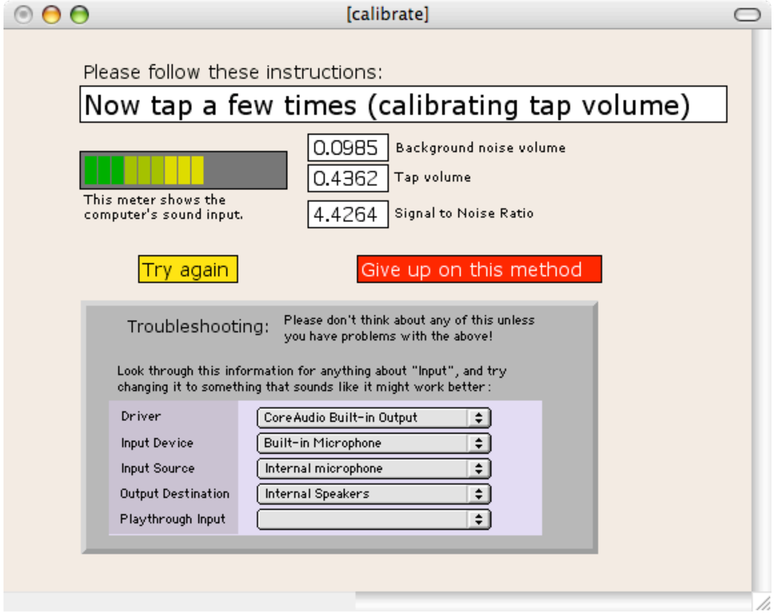
Troubleshooting: Please don't think about any of this unless you have problems with the above!

Look through this information for anything about "Input", and try changing it to something that sounds like it might work better:

Driver	CoreAudio Built-in Output
Input Device	Built-in Microphone
Input Source	Internal microphone
Output Destination	Internal Speakers
Playthrough Input	

Figure 79: Step one of audio tap input calibration: find the level of the background noise.

A calibration phase in the opening sequence of screens offers the subject a choice of methods for inputting tap times. If the user chooses to use an audio input then a calibration process attempts to set an amplitude threshold to distinguish subjects' taps from the background noise. (The software suggests that the subject insert a male-to-male audio cable into the computer's sound input and tap directly on the cable, inspired by (Dixon and Goebl 2002). Otherwise the subject taps on the computer's built-in microphone, and sees the message "this method might not work very well if you're listening through a laptop's built-in speakers.") First the subject is instructed not to tap for two seconds, during which time the software finds the maximum instantaneous amplitude of the background noise, as shown in Figure 79. Then the subject is instructed to tap, as shown in Figure 80. If the maximum amplitude of the tapping is less than twice the maximum amplitude of the background noise, the “instructions” become “Sorry, I couldn't tell the difference. Try again?” Otherwise, the arithmetic mean of the noise level and the tapping level becomes the amplitude threshold at which the software registers the subject’s audio taps. This window also contains some settings that subjects can adjust in case Max/MSP does not register audio input at all.



[calibrate]

Please follow these instructions:

Now tap a few times (calibrating tap volume)

This meter shows the computer's sound input.

0.0985	Background noise volume
0.4362	Tap volume
4.4264	Signal to Noise Ratio

Troubleshooting: Please don't think about any of this unless you have problems with the above!

Look through this information for anything about "Input", and try changing it to something that sounds like it might work better:

Driver	CoreAudio Built-in Output
Input Device	Built-in Microphone
Input Source	Internal microphone
Output Destination	Internal Speakers
Playthrough Input	

Figure 80: Step two of audio tap input calibration: find the level of tap inputs and hence the signal-to-noise ratio.

Once the software has determined an amplitude threshold that discriminates the subject’s taps from the background noise, the process of detecting audio tap input is conceptually simple: Whenever the audio input first crosses the threshold, sample and output the current value of the synchronization signal, then wait 100 milliseconds before allowing the next tap to occur.[152] Figure 81 shows the sub-patch implementing this idea; it is somewhat tricky because of the desire to sample the synchronization signal with sample accuracy, i.e., in the signal domain, but needing to output a Max event to the rest of the program. The “>~” object implements the check against the amplitude threshold and outputs a signal each of whose samples is 0 or 1 depending on whether or not the corresponding audio input is above the threshold. This goes into a minmax~

[152] A simple, obvious, and incorrect implementation of this idea would be a >~ connected to an edge~ connected to a snapshot~. The problem is that by the time the bang from edge~ were transmitted, the synchronization signal input to the snapshot~ object would probably have advanced past the correct value. (Also, the snapshot~ object is not sample accurate; it returns the first entry of the signal vector.)

object, whose second output is a signal giving the maximum value seen on the input since it was reset. In other words, the output of minmax~ is zero for as long as the audio input is below the threshold, then changes to one at the instant the audio input crosses the threshold and stays there until reset. This single transition from zero to one is used as the control signal for a sample-and-hold ("sah~") object, which samples the synchronization signal at the proper instant. Everything described so far takes place in the signal domain on a per-sample basis. Now we must cross into Max's event domain, incurring a scheduling delay on the order of milliseconds. The "edge~" object outputs a bang when the tap occurs, and this bang comes out after the aforementioned scheduling delay. There's a chance that the bang from the edge~ object may actually occur before the sah~ object starts outputting the proper new value, so the bang passes through an additional 10 milliseconds' delay. The bang finally reaches the snapshot~ object well after the tap occurred, but the sah~ object continues to output the proper value of the sync signal, so the proper value is output. After 100 milliseconds the minmax~ object is reset and the process may begin again.

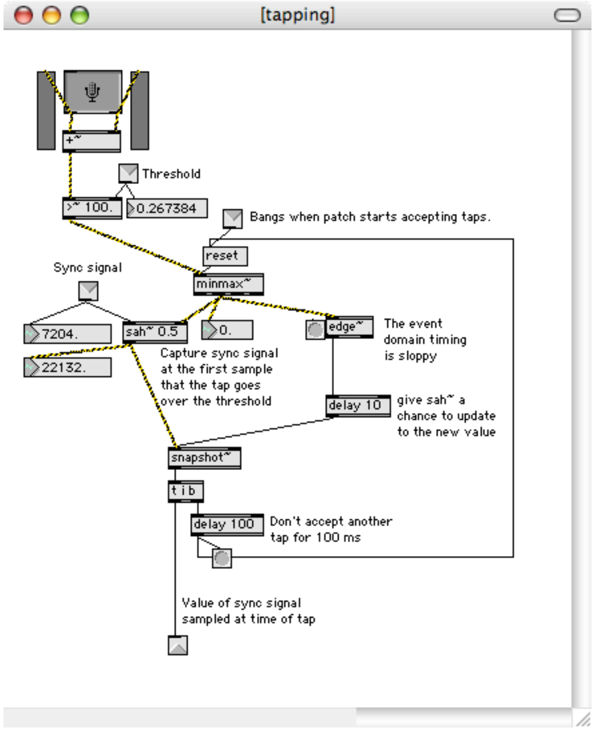
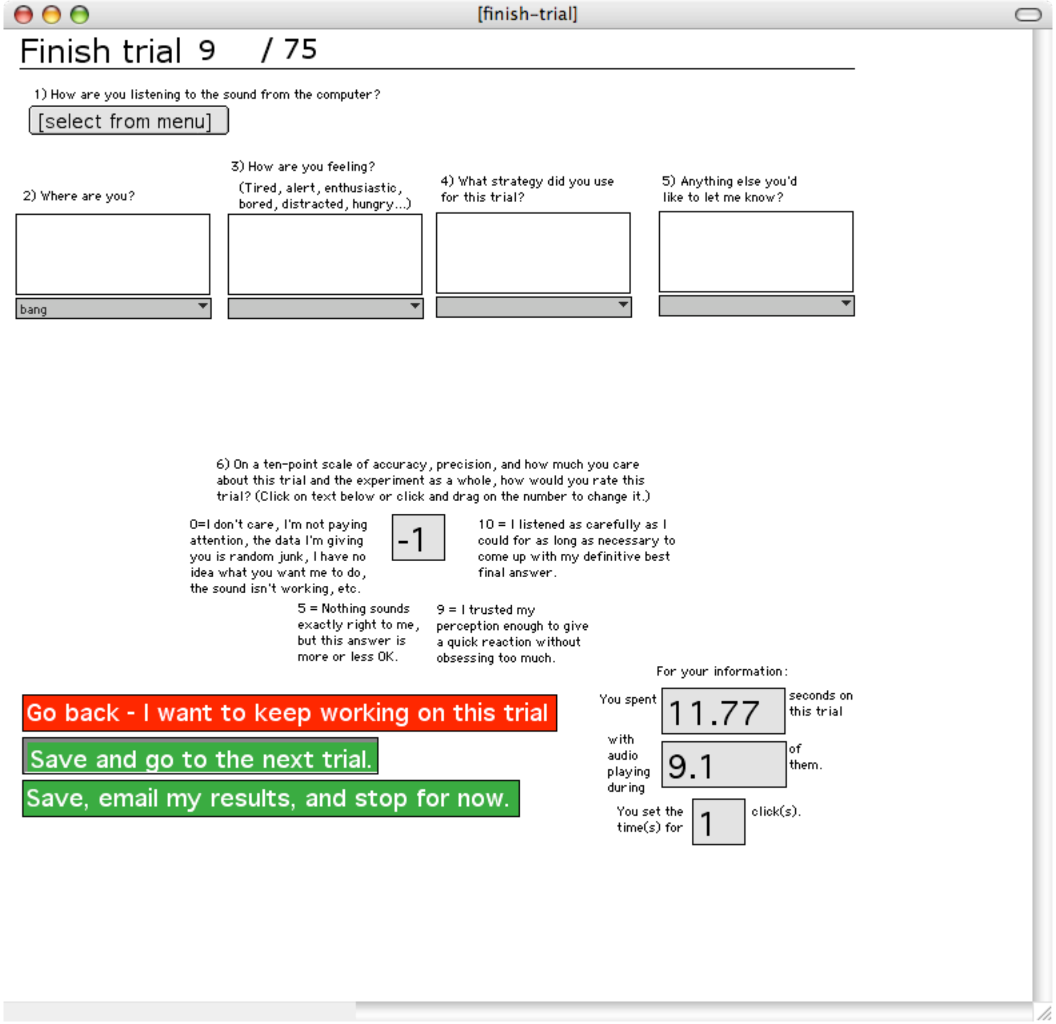


Figure 81: Sub-patch for finding times when the subject taps on the audio input. The left inlet of this patch ("Sync signal") is connected to the synchronization signal produced by the count~ object in Figure 76. The middle input ("Threshold") comes from the threshold set by the patch shown in Figure 80. The patch outputs a Max event (an integer) containing the value of the sync signal sampled at the instant the audio input crosses the threshold. See the text for a more detailed explanation.

A.2.3 Logging the Subject's Actions in Each Trial


[finish-trial]

Finish trial 9 / 75

1) How are you listening to the sound from the computer?

[select from menu]

2) Where are you?

bang

3) How are you feeling?
(Tired, alert, enthusiastic, bored, distracted, hungry...)

4) What strategy did you use for this trial?

5) Anything else you'd like to let me know?

6) On a ten-point scale of accuracy, precision, and how much you care about this trial and the experiment as a whole, how would you rate this trial? (Click on text below or click and drag on the number to change it.)

0=I don't care, I'm not paying attention, the data I'm giving you is random junk, I have no idea what you want me to do, the sound isn't working, etc.

-1

10 = I listened as carefully as I could for as long as necessary to come up with my definitive best final answer.

5 = Nothing sounds exactly right to me, but this answer is more or less OK.

9 = I trusted my perception enough to give a quick reaction without obsessing too much.

Go back - I want to keep working on this trial

Save and go to the next trial.

Save, email my results, and stop for now.

For your information:

You spent 11.77 seconds on this trial

with audio playing during 9.1 of them.

You set the time(s) for 1 click(s).

Figure 82: "Finish trial" screen, for collecting information at the completion of each trial.

The most important result from each trial is simply the final time for each click sound relative to the loop. To provide data for further analysis in the future, and to allow me to investigate and discard certain kinds of bogus trials, I also collected a time-stamped log of the subject's actions during each trial:

- Any adjustment of the time of a click (whether via the keyboard, the on-screen slider, or tapping). The log records which click was moved, which key was pressed, the amount of time (in samples) by which the click was moved, the resulting new time for that click, and the time (in milliseconds since the beginning of the trial) that the adjustment occurred.
- Switching between listening to just the loop, just the click(s), or both.
- The time (in milliseconds since the beginning of the trial) that user completed the trial.

The trial log also contains the date and time that the trial began (according to the clock on the subject's computer), the version number of the software used to run the experiment, the trial number, the names of the two sound files used for the trial, the period of the loop, the randomly chosen initial physical offset, the string displayed as the task for this trial (as shown in Figure 69), as well as the personal information gathered from the screen shown in Figure 68. Finally, at the end of each trial the software brought up the "finish trial" screen (shown in Figure 82); all of these responses also appeared in the trial log.

```
/time 2007 12 9 9 8 39
/looplen_samps 26459
/click-offset 1 -881
/software/version 1.4.1
/subject/gender M
/subject/age 36
/subject/name Matt Wright
/listening/thru good speakers
/rating 8
/trial-number 4
/select-click 1
/allow-click-selection 0
/loop-length 600
/task Synchronize two synthetic tones
/click vcq526.seg.wav
/loop tpq642.seg.wav
/init-offset -1524
/slider -681 1 2342
/slider -553 1 2376
/slider -1390 1 3365.81103515625
/adjustment 1 1 5 439 -2475 5112.81103515625
/adjustment 2 1 5 460 -2015 5716.81103515625
/adjustment 3 1 5 429 -1586 6324.81103515625
/adjustment 4 1 5 429 -1157 6932.81103515625
/adjustment 5 1 4 227 -930 8228.810546875
/adjustment 6 1 3 49 -881 9252.810546875
/completed 14600.1396484375
```

Figure 83: Example trial log. This was trial number four for a 36-year-old male named Matt Wright, and it took place at 9:08:39 am on December 9, 2007 using version 1.4.1 of the software, with sound coming through "good speakers." The task was "synchronize two synthetic tones." The tones were tpq642.seg.wav (the trumpet, which was the fixed "loop") and vcq526.seg.wav (the violin, which was the moveable "click"). The trumpet repeated every 600 ms (26459 samples). Initially the violin began 1524 samples (about 35 ms) before the trumpet. After 2.3 seconds the subject used the slider to move the violin earlier by an additional 681 samples (about 15 ms) and then 553 samples (about 12.5 ms) with respect to the trumpet (these numbers are in addition to the initial 1524 samples of offset). About a second later the subject used the slider to move the violin earlier. Around five seconds into the trial the subject used the QWERTY keyboard to move the violin later, first in four large increments of about 10 ms each, then in a smaller increment of about 5 ms, then a yet smaller increment of about 1 ms. The subject finished the trial in a total of 14.6 seconds, ending up with a final offset amount of -881 samples, i.e., the violin preceding the trumpet by about 20ms.

Figure 83 shows an example log for a trial I performed; note the format inspired by Open Sound Control (Wright and Freed 1997) in which each line is an individual message with a symbolic address followed by arguments.

A.2.4 Sending Back Data from Completed Trials

The most challenging implementation issue for this program was the method for sending data from completed trials back to me. At first I wanted to use Open Sound Control (Wright 2005; Wright and Freed 1997). This would have required setting up a server that would constantly listen for Open Sound Control messages sent by subjects, and write them to files. CCRMA's system administration team was appropriately concerned about the possibility of exploiting this kind of server via a denial-of-service attack, by which a malicious party could waste unlimited disk space by sending bogus "trial data" messages, so we abandoned this idea.

Max has the ability to open URLs using the user's default web browser. For "mailto" URLs (Hoffman, Masinter, and Zawinski 1998) this opens the user's default mail program (which might be a standalone mail client application or the "compose message" screen of a web-based email client). This method works perfectly for short messages (such as the "Let's discuss your research" email created by the screen shown in Figure 66), but unfortunately both Max and various operating systems impose limits on the size of email messages that can be opened with this mechanism, and these limits are too small for the trial logs I needed to send.

So I had to use a more roundabout solution. I wrote a Javascript program, running within Max's "js" object, to take in each line of the trial log and store it in an internal data structure. When the user was ready to send the data back to me, this Javascript program wrote an HTML file (Berners-Lee 1995) containing an enormous "mailto" link around the text "Click here to email results to Matt." Even this method was not completely reliable, because of a Windows limit to the amount of text that can be in the body of a mailto link, so the html file also contained a form with a text area containing the trial log and with a "mailto" action invoked when the subject hits the "submit" button.[153] Figure 84 shows the html file that allows the user to email me the trial log shown in Figure 83.

```
<HTML><BODY><H1><A HREF="mailto:matt@ccrma.stanford.edu?subject=listening test
data&body=/time%202007%2012%209%209%208%2039%20%0A
/looplen_samps%2026459%20%0A
/click-offset%201%20-881%20%0A
/software/version%201.4.1%20%0A
/subject/gender%20M%20%0A
```

[153] Thanks to Andrew Schmeder for suggesting this method and helping me implement it.

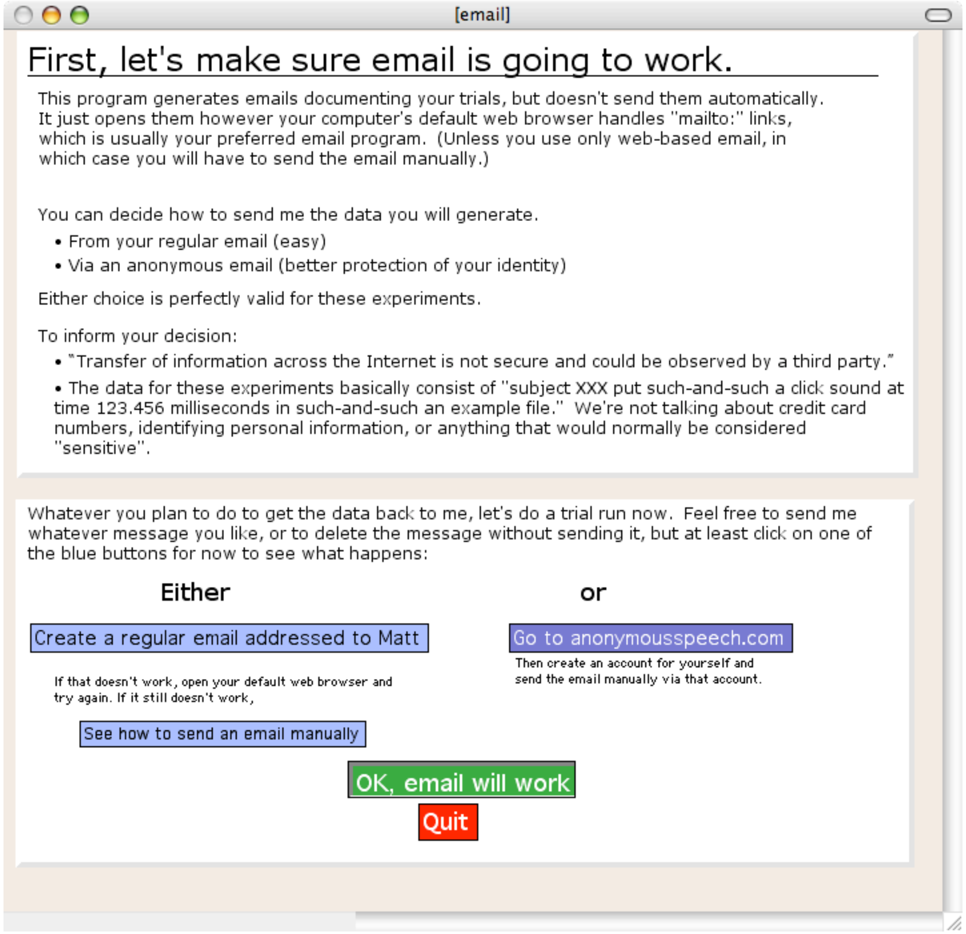
```
/subject/age%2036%20%0A
/subject/name%20Matt%20Wright%20%0A
/listening/thru%20good%20speakers%0A
/rating%208%20%0A
/trial-number%204%20%0A
/select-click%201%20%0A
/allow-click-selection%200%20%0A
/loop-length%20600%20%0A
/task%20Synchronize%20two%20synthetic%20tones%20%0A
/click%20vcq526.seg.wav%20%0A
/loop%20tpq642.seg.wav%20%0A
/init-offset%20-1524%20%0A
/slider%20-681%201%202342%20%0A
/slider%20-553%201%202376%20%0A
/slider%20-1390%201%203365.81103515625%20%0A
/adjustment%201%201%205%20439%20-2475%205112.81103515625%20%0A
/adjustment%202%201%205%20460%20-2015%205716.81103515625%20%0A
/adjustment%203%201%205%20429%20-1586%206324.81103515625%20%0A
/adjustment%204%201%205%20429%20-1157%206932.81103515625%20%0A
/adjustment%205%201%204%20227%20-930%208228.810546875%20%0A
/adjustment%206%201%203%2049%20-881%209252.810546875%20%0A
/completed%2014600.1396484375%20%0A
--------------------------------------------------%20%0A"> Click here to email results to Matt </a></H1><hr>
If that doesn't work (which is likely on Windows), try this instead:
<form action="mailto:matt@ccrma.stanford.edu?subject=listening test data" method="post"
<textarea name="
/time 2007 12 9 9 8 39
/looplen_samps 26459
/click-offset 1 -881
/software/version 1.4.1
/subject/gender M
/subject/age 36
/subject/name Matt Wright
/listening/thru good speakers
/rating 8
/trial-number 4
/select-click 1
/allow-click-selection 0
/loop-length 600
/task Synchronize two synthetic tones
/click vcq526.seg.wav
/loop tpq642.seg.wav
/init-offset -1524
/slider -681 1 2342
/slider -553 1 2376
/slider -1390 1 3365.81103515625
/adjustment 1 1 5 439 -2475 5112.81103515625
/adjustment 2 1 5 460 -2015 5716.81103515625
/adjustment 3 1 5 429 -1586 6324.81103515625
/adjustment 4 1 5 429 -1157 6932.81103515625
/adjustment 5 1 4 227 -930 8228.810546875
/adjustment 6 1 3 49 -881 9252.810546875
/completed 14600.1396484375
-------------------------------------------------- ">
</textarea> <input type="submit"></form>
<hr>
<p>If neither of the above work, your last resort is to open your
usual mail program manually and send this HTML file as an attachment.
This file is named sendemail-2.html
It's located in the same folder as the application for running this experiment.
If you're reading this in a web browser, it should be telling you
where on your computer's disk this file is located.</p>
```

```
<p>Use whichever of these three methods works best for you.</p>
</BODY></HTML>
```

Figure 84: Example html file used for emailing a trial log

As a backup, the software also wrote individual text files containing the log for each completed trial, so that users could manually attach the logs to an email to me in case the other methods failed. The software included a link to the anonymous email service anonymousspeech.com, where I had created an account that subjects could use if they wished to submit data anonymously. (No subjects used this method to submit their results.)

Because of all the complication and potential problems with the sending of this email, the third screen of the software (shown in Figure 85) tests the full process of creating an email inside an html file. Clicking on "See how to send an email manually" brings up a screen (shown in Figure 86) of instructions for doing so.



First, let's make sure email is going to work.

This program generates emails documenting your trials, but doesn't send them automatically. It just opens them however your computer's default web browser handles "mailto:" links, which is usually your preferred email program. (Unless you use only web-based email, in which case you will have to send the email manually.)

You can decide how to send me the data you will generate.

- From your regular email (easy)
- Via an anonymous email (better protection of your identity)

Either choice is perfectly valid for these experiments.

To inform your decision:

- "Transfer of information across the Internet is not secure and could be observed by a third party."
- The data for these experiments basically consist of "subject XXX put such-and-such a click sound at time 123.456 milliseconds in such-and-such an example file." We're not talking about credit card numbers, identifying personal information, or anything that would normally be considered "sensitive".

Whatever you plan to do to get the data back to me, let's do a trial run now. Feel free to send me whatever message you like, or to delete the message without sending it, but at least click on one of the blue buttons for now to see what happens:

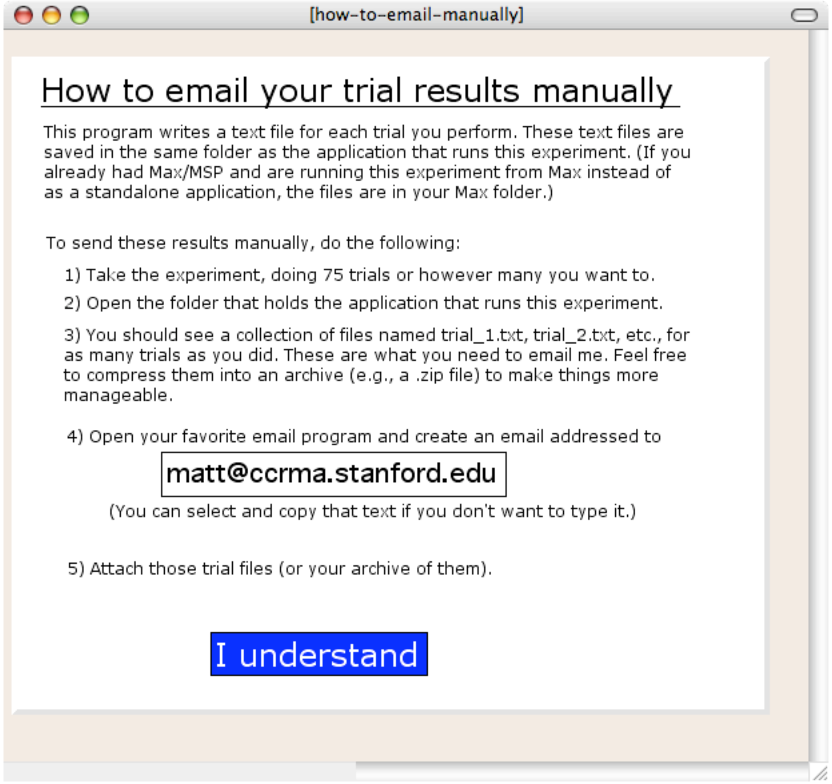
Either

or

If that doesn't work, open your default web browser and try again. If it still doesn't work,

Then create an account for yourself and send the email manually via that account.

Figure 85: Email configuration test screen.



How to email your trial results manually

This program writes a text file for each trial you perform. These text files are saved in the same folder as the application that runs this experiment. (If you already had Max/MSP and are running this experiment from Max instead of as a standalone application, the files are in your Max folder.)

To send these results manually, do the following:

1) Take the experiment, doing 75 trials or however many you want to.

2) Open the folder that holds the application that runs this experiment.

3) You should see a collection of files named trial_1.txt, trial_2.txt, etc., for as many trials as you did. These are what you need to email me. Feel free to compress them into an archive (e.g., a .zip file) to make things more manageable.

4) Open your favorite email program and create an email addressed to

matt@ccrma.stanford.edu

(You can select and copy that text if you don't want to type it.)

5) Attach those trial files (or your archive of them).

Figure 86: Screen of instructions for manually emailing trial results.

An unforeseen complication was the liberties that various email transmission software took in modifying the text in the bodies of these emails. Some trial emails arrived with extra line breaks, or different text encodings or other corruptions. In retrospect it would probably have been better to send trial results as email attachments, or with a transmission medium other than email. At least the consistent use of the email subject line “listening test data” made it easy to gather all such messages and separate them from spam and other messages.

Appendix B Maximum Likelihood Solution for Intrinsic PAT-pdfs?

Let's try to use a maximum likelihood method to solve for all of the parameters of a model of Gaussian intrinsic PAT-pdfs plus a zero-mean Gaussian penalty term for each pair of sounds. What we see is samples of a random variable $D_{T,R}$ (as explained in Section 3.4.4 on page 43). We assume that it has three components (the random variables I_T *and* I_R for the intrinsic PATs of the test and reference sounds, and the penalty term $\varepsilon_{T,R}$ expressing the alignment difficulty penalty for aligning sounds T and R):

$$D_{T,R} = I_T - I_R + \varepsilon_{T,R}$$

We assume all three terms are independent Gaussian random variables and that the alignment penalty $\varepsilon_{T,R}$ has mean zero:

$$I_T \sim \mathcal{N}(\mu_T, \sigma_T^2)$$
$$I_R \sim \mathcal{N}(\mu_R, \sigma_R^2)$$
$$\varepsilon_{T,R} \sim \mathcal{N}(0, \sigma_{T,R}^2)$$

Since the sum of Gaussian random variables is itself a Gaussian random variable, we know

$$D_{T,R} \sim \mathcal{N}(\mu_{D_{T,R}} = \mu_T - \mu_R, \sigma_{D_{T,R}}^2 = \sigma_T^2 + \sigma_R^2 + \sigma_{T,R}^2)$$

The probability density function for $D_{T,R}$ is therefore

$$\frac{1}{\sqrt{\sigma_T^2 + \sigma_R^2 + \sigma_{T,R}^2}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\left(x - \left(\mu_T - \mu_R\right)\right)^2}{2\left(\sigma_T^2 + \sigma_R^2 + \sigma_{T,R}^2\right)} \right).$$

The likelihood of each observation $d_{T,R}(k)$ is

$$\frac{1}{\sqrt{\sigma_T^2 + \sigma_R^2 + \sigma_{T,R}^2}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\left(d_{T,R}(k) - \mu_T + \mu_R\right)^2}{2\left(\sigma_T^2 + \sigma_R^2 + \sigma_{T,R}^2\right)} \right)$$

We take the log to find the log-likelihood of each observed value $d_{T,R}(k)$:

$$\log\left(\left(\sigma_T^2+\sigma_R^2+\sigma_{T,R}^2\right)^{-\frac{1}{2}}\right)+\log(\frac{1}{\sqrt{2\pi}})-\frac{\left(d_{T,R}(k)-\mu_T+\mu_R\right)^2}{2\left(\sigma_T^2+\sigma_R^2+\sigma_{T,R}^2\right)}$$

$$=-\frac{1}{2}\log\left(\sigma_T^2+\sigma_R^2+\sigma_{T,R}^2\right)+\log(\frac{1}{\sqrt{2\pi}})-\frac{\left(d_{T,R}(k)-\mu_T+\mu_R\right)^2}{2\left(\sigma_T^2+\sigma_R^2+\sigma_{T,R}^2\right)}$$

There are $\mathcal{N}_{i,j}$ trials aligning test sound *i* with reference sound *j*, with no particular relationship between $\mathcal{N}_{i,j}$ and $\mathcal{N}_{j,i}$. We pay attention to trials comparing a sound to itself, so $\mathcal{N}_{i,i} \geq 0$.

The total log likelihood of all trials comparing test sound *i* with reference sound *j* is

$$\log(l_{i,j})=\sum_{k=1}^{\mathcal{N}_{i,j}}\left(-\frac{1}{2}\log\left(\sigma_i^2+\sigma_j^2+\sigma_{i,j}^2\right)+\log(\frac{1}{\sqrt{2\pi}})-\frac{\left(d_{i,j}(k)-\mu_i+\mu_j\right)^2}{2\left(\sigma_i^2+\sigma_j^2+\sigma_{i,j}^2\right)}\right)$$

$$=\mathcal{N}_{i,j}\log(\frac{1}{\sqrt{2\pi}})-\mathcal{N}_{i,j}\frac{1}{2}\log\left(\sigma_i^2+\sigma_j^2+\sigma_{i,j}^2\right)-\frac{1}{2\left(\sigma_i^2+\sigma_j^2+\sigma_{i,j}^2\right)}\sum_{k=1}^{\mathcal{N}_{i,j}}\left(d_{i,j}(k)-\mu_i+\mu_j\right)^2$$

The total log likelihood of all of the results of an entire experiment involving *n* sounds is

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\mathcal{N}_{i,j}\log(\frac{1}{\sqrt{2\pi}})-\mathcal{N}_{i,j}\frac{1}{2}\log\left(\sigma_i^2+\sigma_j^2+\sigma_{i,j}^2\right)-\frac{1}{2\left(\sigma_i^2+\sigma_j^2+\sigma_{i,j}^2\right)}\sum_{k=1}^{\mathcal{N}_{i,j}}\left(d_{i,j}(k)-\mu_i+\mu_j\right)^2\right)$$

We want to estimate the $2n$ intrinsic means and variances $\sigma_1^2,\sigma_2^2,...,\sigma_n^2,\mu_1,\mu_2,...,\mu_n$ and the $n(n-1)/2$ alignment penalty variances $\sigma_{2,1}^2,\sigma_{3,1}^2,\sigma_{3,2,}^2,...,\sigma_{n,1,}^2...,\sigma_{n,n-1}^2$. (There would be n^2 alignment penalty variances except that we assume $\sigma_{i,i}^2=0$ and $\sigma_{i,j}^2=\sigma_{j,i}^2$.) There are therefore $n(n+3)/2$ parameters to our model, which we'll indicate collectively with the column vector θ. We will consider the negative of the total log likelihood to be a cost function to be minimized:

$$\mathcal{J}(\theta)=-\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\mathcal{N}_{i,j}\log(\frac{1}{\sqrt{2\pi}})-\mathcal{N}_{i,j}\frac{1}{2}\log\left(\sigma_i^2+\sigma_j^2+\sigma_{i,j}^2\right)-\frac{1}{2\left(\sigma_i^2+\sigma_j^2+\sigma_{i,j}^2\right)}\sum_{k=1}^{\mathcal{N}_{i,j}}\left(d_{i,j}(k)-\mu_i+\mu_j\right)^2\right)$$

Since $\mathcal{N}_{i,j}\log(\frac{1}{\sqrt{2\pi}})$ is a constant, we can equivalently minimize

$$\mathcal{J}(\boldsymbol{\theta}) = -\sum_{i=1}^{n}\sum_{j=1}^{n}\left(-\mathcal{N}_{i,j}\frac{1}{2}\log\left(\sigma_i^2+\sigma_j^2+\sigma_{i,j}^2\right) - \frac{1}{2\left(\sigma_i^2+\sigma_j^2+\sigma_{i,j}^2\right)}\sum_{k=1}^{\mathcal{N}_{i,j}}\left(d_{i,j}(k)-\mu_i+\mu_j\right)^2 \right)$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n}\left(\mathcal{N}_{i,j}\frac{1}{2}\log\left(\sigma_i^2+\sigma_j^2+\sigma_{i,j}^2\right) + \frac{1}{2\left(\sigma_i^2+\sigma_j^2+\sigma_{i,j}^2\right)}\sum_{k=1}^{\mathcal{N}_{i,j}}\left(d_{i,j}(k)-\mu_i+\mu_j\right)^2 \right)$$

We'll take the partial derivative of $\mathcal{J}$ with respect to each parameter of our model, setting each to zero.

B.1 Solving for the Means

First we consider μ_S, the intrinsic mean of each sound *S*:

$$0 = \frac{\partial}{\partial\mu_s}\mathcal{J}(\theta)$$

$$= \frac{\partial}{\partial\mu_s}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\mathcal{N}_{i,j}\frac{1}{2}\log\left(\sigma_i^2+\sigma_j^2+\sigma_{i,j}^2\right) + \frac{1}{2\left(\sigma_i^2+\sigma_j^2+\sigma_{i,j}^2\right)}\sum_{k=1}^{\mathcal{N}_{i,j}}\left(d_{i,j}(k)-\mu_i+\mu_j\right)^2 \right)$$

$$= \frac{\partial}{\partial\mu_s}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{1}{2\left(\sigma_i^2+\sigma_j^2+\sigma_{i,j}^2\right)}\sum_{k=1}^{\mathcal{N}_{i,j}}\left(d_{i,j}(k)-\mu_i+\mu_j\right)^2 \right)$$

$$= \frac{\partial}{\partial\mu_s}\sum_{j=1}^{n}\left(\frac{1}{2\left(\sigma_s^2+\sigma_j^2+\sigma_{s,j}^2\right)}\sum_{k=1}^{\mathcal{N}_{s,j}}\left(d_{s,j}(k)-\mu_s+\mu_j\right)^2 \right) + \frac{\partial}{\partial\mu_s}\sum_{i=1}^{n}\left(\frac{1}{2\left(\sigma_i^2+\sigma_s^2+\sigma_{i,s}^2\right)}\sum_{k=1}^{\mathcal{N}_{i,s}}\left(d_{i,s}(k)-\mu_i+\mu_s\right)^2 \right)$$

$$= \frac{\partial}{\partial\mu_s}\sum_{i=1}^{n}\left(\frac{1}{2\left(\sigma_s^2+\sigma_i^2+\sigma_{s,i}^2\right)}\sum_{k=1}^{\mathcal{N}s,i}\left(d_{s,i}(k)-\mu_s+\mu_i\right)^2 \right) + \frac{\partial}{\partial\mu_s}\sum_{i=1}^{n}\left(\frac{1}{2\left(\sigma_s^2+\sigma_i^2+\sigma_{i,s}^2\right)}\sum_{k=1}^{\mathcal{N}_{i,s}}\left(-d_{i,s}(k)+\mu_i-\mu_s\right)^2 \right)$$

$$= \tfrac{1}{2}\frac{\partial}{\partial\mu_s}\sum_{i=1}^{n}\left(\left(\frac{1}{\sigma_s^2+\sigma_i^2+\sigma_{s,i}^2}\sum_{k=1}^{\mathcal{N}s,i}\left(d_{s,i}(k)-\mu_s+\mu_i\right)^2 \right) + \left(\frac{1}{\sigma_s^2+\sigma_i^2+\sigma_{i,s}^2}\sum_{k=1}^{\mathcal{N}_{i,s}}\left(-d_{i,s}(k)+\mu_i-\mu_s\right)^2 \right)\right)$$

$$= \tfrac{1}{2}\sum_{i=1}^{n}\left(\left(\frac{1}{\sigma_s^2+\sigma_i^2+\sigma_{s,i}^2}\sum_{k=1}^{\mathcal{N}s,i}\frac{\partial}{\partial\mu_s}\left(d_{s,i}(k)-\mu_s+\mu_i\right)^2 \right) + \left(\frac{1}{\sigma_s^2+\sigma_i^2+\sigma_{i,s}^2}\sum_{k=1}^{\mathcal{N}_{i,s}}\frac{\partial}{\partial\mu_s}\left(-d_{i,s}(k)+\mu_i-\mu_s\right)^2 \right)\right)$$

$$= \tfrac{1}{2}\sum_{i=1}^{n}\left(\left(\frac{1}{\sigma_s^2+\sigma_i^2+\sigma_{s,i}^2}\sum_{k=1}^{\mathcal{N}s,i}2\left(d_{s,i}(k)-\mu_s+\mu_i\right) \right) + \left(\frac{1}{\sigma_s^2+\sigma_i^2+\sigma_{i,s}^2}\sum_{k=1}^{\mathcal{N}_{i,s}}2\left(-d_{i,s}(k)+\mu_i-\mu_s\right) \right)\right)$$

$$= \sum_{i=1}^{n}\left(\left(\frac{1}{\sigma_s^2+\sigma_i^2+\sigma_{s,i}^2}\left(-\mathcal{N}_{s,i}\mu_s+\mathcal{N}_{s,i}\mu_i+\sum_{k=1}^{\mathcal{N}s,i}d_{s,i}(k)\right) \right) + \left(\frac{1}{\sigma_s^2+\sigma_i^2+\sigma_{i,s}^2}\left(\mathcal{N}_{i,s}\mu_i-\mathcal{N}_{i,s}\mu_s+\sum_{k=1}^{\mathcal{N}_{i,s}}-d_{i,s}(k)\right) \right)\right)$$

To get from the third line to the fourth we used the fact that of all terms of the expansion of the double summation, only those for *i*=*s* or *j*=*s* will contain μ_S. At this point we need to assume $\sigma^2_{i,j} = \sigma^2_{j,i}$:

$$0 = \sum_{i=1}^{n} \frac{1}{\sigma_s^2 + \sigma_i^2 + \sigma_{i,s}^2} \left(-\left(\mathcal{N}_{s,i} + \mathcal{N}_{i,s}\right)\mu_s + \left(\mathcal{N}_{s,i} + \mathcal{N}_{i,s}\right)\mu_i + \sum_{k=1}^{\mathcal{N}s,i} d_{s,i}(k) - \sum_{k=1}^{\mathcal{N}_{i,s}} d_{i,s}(k) \right)$$

This is a system of *n* nonlinear equations in 3*n* unknowns.

B.1.1 Solving for the Means if Variances are Known Constants

By treating the variances as constants we can set up *n* linear equations in the *n* unknown means:

$$\begin{aligned}
0 &= \sum_{i=1}^{n} \frac{1}{\sigma_s^2 + \sigma_i^2 + \sigma_{i,s}^2} \left(-\left(\mathcal{N}_{s,i} + \mathcal{N}_{i,s}\right)\mu_s + \left(\mathcal{N}_{s,i} + \mathcal{N}_{i,s}\right)\mu_i + \sum_{k=1}^{\mathcal{N}s,i} d_{s,i}(k) - \sum_{k=1}^{\mathcal{N}_{i,s}} d_{i,s}(k) \right) \\
&= -\mu_s \sum_{i=1}^{n} \frac{\mathcal{N}_{s,i} + \mathcal{N}_{i,s}}{\sigma_s^2 + \sigma_i^2 + \sigma_{i,s}^2} + \sum_{i=1}^{n} \frac{\mathcal{N}_{s,i} + \mathcal{N}_{i,s}}{\sigma_s^2 + \sigma_i^2 + \sigma_{i,s}^2}\mu_i + \sum_{i=1}^{n} \frac{1}{\sigma_s^2 + \sigma_i^2 + \sigma_{i,s}^2} \left(\sum_{k=1}^{\mathcal{N}s,i} d_{s,i}(k) - \sum_{k=1}^{\mathcal{N}_{i,s}} d_{i,s}(k) \right) \\
&= \frac{\mathcal{N}_{s,1} + \mathcal{N}_{1,s}}{\sigma_s^2 + \sigma_1^2 + \sigma_{1,s}^2}\mu_1 + \frac{\mathcal{N}_{s,2} + \mathcal{N}_{2,s}}{\sigma_s^2 + \sigma_2^2 + \sigma_{2,s}^2}\mu_2 + \ldots + \left(\frac{2\mathcal{N}_{s,s}}{2\sigma_s^2 + \sigma_{s,s}^2} - \sum_{i=1}^{n} \frac{\mathcal{N}_{s,i} + \mathcal{N}_{i,s}}{\sigma_s^2 + \sigma_i^2 + \sigma_{i,s}^2} \right)\mu_s + \ldots \\
&\quad \ldots + \frac{\mathcal{N}_{s,n} + \mathcal{N}_{n,s}}{\sigma_s^2 + \sigma_n^2 + \sigma_{n,s}^2}\mu_n + \sum_{i=1}^{n} \frac{1}{\sigma_s^2 + \sigma_i^2 + \sigma_{i,s}^2} \left(\sum_{k=1}^{\mathcal{N}s,i} d_{s,i}(k) - \sum_{k=1}^{\mathcal{N}_{i,s}} d_{i,s}(k) \right)
\end{aligned}$$

Setting $\frac{\partial}{\partial \mu_s}\mathcal{J}(\boldsymbol{\theta}) = 0$ for all *s* produces the following system of *n* linear equations in the *n* unknown values of μ_S:

$$\begin{pmatrix} \left(\frac{2\mathcal{N}_{1,1}}{2\sigma_1^2+\sigma_{1,1}^2} - \sum_{i=1}^{n} \frac{\mathcal{N}_{1,i}+\mathcal{N}_{i,1}}{\sigma_i^2+\sigma_1^2+\sigma_{i,1}^2} \right) & \frac{\mathcal{N}_{2,1}+\mathcal{N}_{1,2}}{\sigma_2^2+\sigma_1^2+\sigma_{2,1}^2} & \cdots & \frac{\mathcal{N}_{n,1}+\mathcal{N}_{1,n}}{\sigma_n^2+\sigma_1^2+\sigma_{n,1}^2} \\ \frac{\mathcal{N}_{1,2}+\mathcal{N}_{2,1}}{\sigma_1^2+\sigma_2^2+\sigma_{1,2}^2} & \left(\frac{2\mathcal{N}_{2,2}}{2\sigma_2^2+\sigma_{2,2}^2} - \sum_{i=1}^{n} \frac{\mathcal{N}_{i,2}+\mathcal{N}_{2,i}}{\sigma_i^2+\sigma_2^2+\sigma_{i,2}^2} \right) & \cdots & \frac{\mathcal{N}_{n,2}+\mathcal{N}_{2,n}}{\sigma_n^2+\sigma_2^2+\sigma_{n,2}^2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\mathcal{N}_{1,n}+\mathcal{N}_{n,1}}{\sigma_1^2+\sigma_n^2+\sigma_{1,n}^2} & \frac{\mathcal{N}_{2,n}+\mathcal{N}_{n,2}}{\sigma_2^2+\sigma_n^2+\sigma_{2,n}^2} & \cdots & \left(\frac{2\mathcal{N}_{n,n}}{2\sigma_n^2+\sigma_{n,n}^2} - \sum_{i=1}^{n} \frac{\mathcal{N}_{i,n}+\mathcal{N}_{n,i}}{\sigma_i^2+\sigma_n^2+\sigma_{i,n}^2} \right) \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} \frac{\sum_{k=1}^{\mathcal{N}1,i} d_{1,i}(k) - \sum_{k=1}^{\mathcal{N}_{i,1}} d_{i,1}(k)}{\sigma_i^2+\sigma_1^2+\sigma_{i,1}^2} \\ \sum_{i=1}^{n} \frac{\sum_{k=1}^{\mathcal{N}_{2,i}} d_{2,i}(k) - \sum_{k=1}^{\mathcal{N}_{i,2}} d_{i,2}(k)}{\sigma_i^2+\sigma_2^2+\sigma_{i,2}^2} \\ \vdots \\ \sum_{i=1}^{n} \frac{\sum_{k=1}^{\mathcal{N}_{n,i}} d_{n,i}(k) - \sum_{k=1}^{\mathcal{N}_{i,n}} d_{i,n}(k)}{\sigma_i^2+\sigma_n^2+\sigma_{i,n}^2} \end{pmatrix}$$

Since $\sigma_{i,j}^2 = \sigma_{j,i}^2$, the matrix of coefficients is symmetric. By construction this matrix of coefficients is also singular, because we can add any constant to every μ_S and still equally explain the observed experimental results, as described in section 3.5.4.2 (page 53). As before, we can solve this system of equations except for the constant either by computing the pseudo-inverse or by arbitrarily removing one of the columns of the matrix and the corresponding μ_i.

It is also instructive to compare this system of linear equations with the one on page 55; they are essentially identical except for the variances in the denominators of all the fractions. An intuitive explanation for this difference is that this method scales the data from each pair of sounds by the inverse of its variance. For example, if subjects were more consistent overall in their alignment of sounds A and B than in their alignment of sounds A and C, then $\sigma_A^2+\sigma_B^2+\sigma_{A,B}^2 < \sigma_A^2+\sigma_C^2+\sigma_{A,C}^2$ and so the trials comparing A and B will "count more" in the overall solution than the trials comparing A and C. Section 3.6 (page 61) motivates the benefit of this down-weighting by variance.

B.2 Solving for the Intrinsic Variances

Next we consider the intrinsic variance of each sound:

$$
\begin{aligned}
0 &= \frac{\partial}{\partial \sigma_s^2} \mathcal{J}(\boldsymbol{\theta}) \\
&= \frac{\partial}{\partial \sigma_s^2} \sum_{i=1}^{n} \sum_{j=1}^{i} \left(\mathcal{N}_{i,j} \frac{1}{2} \log\left(\sigma_i^2 + \sigma_j^2 + \sigma_{i,j}^2\right) + \frac{1}{2\left(\sigma_i^2 + \sigma_j^2 + \sigma_{i,j}^2\right)} \sum_{k=1}^{\mathcal{N}_{i,j}} \left(d_{i,j}(k) - \mu_i + \mu_j\right)^2 \right) \\
&= \frac{\partial}{\partial \sigma_s^2} \sum_{j=1}^{s} \left(\mathcal{N}_{s,j} \frac{1}{2} \log\left(\sigma_s^2 + \sigma_j^2 + \sigma_{s,j}^2\right) + \frac{1}{2\left(\sigma_s^2 + \sigma_j^2 + \sigma_{s,j}^2\right)} \sum_{k=1}^{\mathcal{N}_{s,j}} \left(d_{s,j}(k) - \mu_s + \mu_j\right)^2 \right) \\
&\quad + \frac{\partial}{\partial \sigma_s^2} \sum_{i=s+1}^{n} \left(\mathcal{N}_{i,s} \frac{1}{2} \log\left(\sigma_i^2 + \sigma_s^2 + \sigma_{i,s}^2\right) + \frac{1}{2\left(\sigma_i^2 + \sigma_s^2 + \sigma_{i,s}^2\right)} \sum_{k=1}^{\mathcal{N}_{i,s}} \left(d_{i,s}(k) - \mu_i + \mu_s\right)^2 \right) \\
&= \frac{\partial}{\partial \sigma_s^2} \sum_{i=1}^{n} \left(\mathcal{N}_{i,s} \frac{1}{2} \log\left(\sigma_i^2 + \sigma_s^2 + \sigma_{i,s}^2\right) + \frac{1}{2\left(\sigma_i^2 + \sigma_s^2 + \sigma_{i,s}^2\right)} \sum_{k=1}^{\mathcal{N}_{i,s}} \left(d_{i,s}(k) - \mu_i + \mu_s\right)^2 \right) \\
&= \sum_{i=1}^{n} \left(\mathcal{N}_{s,i} \frac{\partial}{\partial \sigma_s^2} \log\left(\sigma_s^2 + \sigma_i^2 + \sigma_{s,i}^2\right) + \left(\sum_{k=1}^{\mathcal{N}_{s,i}} \left(d_{s,i}(k) - \mu_s + \mu_i\right)^2 \right) \frac{\partial}{\partial \sigma_s^2} \left(\sigma_s^2 + \sigma_i^2 + \sigma_{s,i}^2\right)^{-1} \right) \\
&= \sum_{i=1}^{n} \left(\mathcal{N}_{s,i} \frac{1}{\sigma_s^2 + \sigma_i^2 + \sigma_{s,i}^2} + \left(\sum_{k=1}^{\mathcal{N}_{s,i}} \left(d_{s,i}(k) - \mu_s + \mu_i\right)^2 \right) \left(-1\right) \left(\sigma_s^2 + \sigma_i^2 + \sigma_{s,i}^2\right)^{-2} \right) \\
&= \sum_{i=1}^{n} \left(\left(\sigma_s^2 + \sigma_i^2 + \sigma_{s,i}^2\right)^{-2} \left(\mathcal{N}_{s,i} \left(\sigma_s^2 + \sigma_i^2 + \sigma_{s,i}^2\right) - \left(\sum_{k=1}^{\mathcal{N}_{s,i}} \left(d_{s,i}(k) - \mu_s + \mu_i\right)^2 \right) \right) \right)
\end{aligned}
$$

Again we have *n* nonlinear equations in *3n* unknowns. (This time the equations are nonlinear in the variances we're trying to optimize, so even if we treat the means as constants the equations are still nonlinear.)

B.3 Solving for the Pairwise Alignment Penalty Variances

Finally we attempt to solve for the variance of the alignment penalty for each pair of sounds:

$$
\begin{aligned}
0 &= \frac{\partial}{\partial \sigma_{s,t}} \mathcal{J}(\boldsymbol{\theta}) \\
&= \frac{\partial}{\partial \sigma_{s,t}} \sum_{i=1}^{n} \sum_{j=1}^{i} \left(\mathcal{N}_{i,j} \frac{1}{2} \log\left(\sigma_i^2 + \sigma_j^2 + \sigma_{i,j}^2\right) + \frac{1}{2\left(\sigma_i^2 + \sigma_j^2 + \sigma_{i,j}^2\right)} \sum_{k=1}^{\mathcal{N}_{i,j}} \left(d_{i,j}(k) - \mu_i - \mu_j\right)^2 \right) \\
&= \frac{\partial}{\partial \sigma_{s,t}} \mathcal{N}_{s,t} \frac{1}{2} \log\left(\sigma_s^2 + \sigma_t^2 + \sigma_{s,t}^2\right) + \frac{1}{2\left(\sigma_s^2 + \sigma_t^2 + \sigma_{s,t}^2\right)} \sum_{k=1}^{\mathcal{N}_{s,t}} \left(d_{s,t}(k) - \mu_s - \mu_t\right)^2, s > t \\
&= \mathcal{N}_{s,t} \frac{1}{2} \frac{\partial}{\partial \sigma_{s,t}} \log\left(\sigma_s^2 + \sigma_t^2 + \sigma_{s,t}^2\right) + \frac{1}{2} \left(\sum_{k=1}^{\mathcal{N}_{s,t}} \left(d_{s,t}(k) - \mu_s - \mu_t\right)^2 \right) \frac{\partial}{\partial \sigma_{s,t}} \left(\sigma_s^2 + \sigma_t^2 + \sigma_{s,t}^2\right)^{-1}, s > t
\end{aligned}
$$

(continued on next page)

$$= \mathcal{N}_{s,t} \frac{1}{\sigma_s^2 + \sigma_t^2 + \sigma_{s,t}^2} + \left(\sum_{k=1}^{\mathcal{N}_{s,t}} \left(d_{s,t}(k) - \mu_s - \mu_t \right)^2 \right) \left(-1 \right) \left(\sigma_s^2 + \sigma_t^2 + \sigma_{s,t}^2 \right)^{-2}, s > t$$

$$= \mathcal{N}_{s,t} \left(\sigma_s^2 + \sigma_t^2 + \sigma_{s,t}^2 \right) - \left(\sum_{k=1}^{\mathcal{N}_{s,t}} \left(d_{s,t}(k) - \mu_s - \mu_t \right)^2 \right), s > t$$

$$= \mathcal{N}_{s,t} \sigma_s^2 + \mathcal{N}_{s,t} \sigma_t^2 + \mathcal{N}_{s,t} \sigma_{s,t}^2 - \left(\sum_{k=1}^{\mathcal{N}_{s,t}} \left(d_{s,t}(k) - \mu_s - \mu_t \right)^2 \right), s > t$$

Once again the resulting equations are nonlinear, but only in the unknown means.

All in all we have $2n + \frac{n^2}{2}$ nonlinear equations in $2n + \frac{n^2}{2}$ unknowns.

B.4 Future Work

It still may be possible to find the θ that minimizes $\mathcal{J}(\theta)$. A nonlinear equation solver might be able to solve the combined system of all n(n-3)/2 equations that result from setting each partial derivative of $\mathcal{J}(\theta)$ to zero. Alternately, based on the work so far, it seems straightforward to find every second derivative of $\mathcal{J}(\theta)$, (in other words, the Hessian matrix) analytically; this could be the basis of a gradient descent optimization algorithm.

Another approach would be an iterative "relaxation" algorithm that starts with initial guesses for all parameters then alternately treats the variances as constants and uses their values to optimize the means (according to Section B.1.1) and then treats the means as constants and uses their values to optimize the variances. This process could continue back and forth until convergence.

Note that every variance always appears only as part of the sum $\sigma_i^2 + \sigma_j^2 + \sigma_{i,j}^2$, so it might be a good idea to define a new variable $x_{i,j}$ equal to this sum, and use the iterative relaxation algorithm to find all μ_i and all $x_{i,j}$. Then the "partitioning" of the variance $x_{i,j}$ among σ_i^2, σ_j^2, and $\sigma_{i,j}^2$ could be a separate step.

Bibliography

Abdallah, Samer and Mark Plumbley. 2003. Unsupervised onset detection: a probabilistic approach using ICA and a hidden markov classifier. In *Proceedings of the Cambridge Music Processing Colloquium* (Cambridge, UK).

Agawu, Kofi. 2006. Structural Analysis or Cultural Analysis? Competing Perspectives on the "Standard Pattern" of West African Rhythm. *Journal of the American Musicological Society* 59, no. 1: 1-46.

Alonso, Miguel, Bertrand David, and Gaël Richard. 2004. Tempo and beat estimation of musical signals. In *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR)* (Barcelona): 158–163.

Anku, Willie. 2000. Circles and Time: A Theory of Structural Organization of Rhythm in African Music. *Music Theory Online, The Online Journal of the Society for Music Theory* 6, no. 1. (http://mto.societymusictheory.org/issues/mto.00.6.1/mto.00.6.1.anku.html)

ANSI. 1973. Psychoacoustical terminology: S3.20. New York: American National Standards Institute.

Arom, Simha. 1989. Time Structure in the Music of Central Africa: Periodicity, Meter, Rhythm and Polyrhythmics. *Leonardo* 22, no. 1: 91-99.

Aschersleben, Gisa. 2002. Temporal Control of Movements in Sensorimotor Synchronization. *Brain and Cognition* 48, no. 1: 66-79.

Atlas, Les. 2003. Modulation Spectral Transforms: Application to Speech Separation and Modification. *Institute of Electronics, Information and Communication Engineers (IEIC) Technical Report* 103, no. 155: 49-54.

Atlas, Les and Shihab A. Shamma. 2003. Joint Acoustic and Modulation Frequency. *EURASIP Journal on Applied Signal Processing* 7: 668-675.

Baily, John. 1988. *Music of Afghanistan: Professional Musicians in the City of Herat*. Cambridge: Cambridge University Press.

Baily, John. 1991. Some Cognitive Aspects of Motor Planning in Musical Performance. *Psychologica Belgica* 31, no. 2: 147-162.

Beek, Peter J., C. (Lieke) E. Peper, and Andreas Daffertshofer. 2000. Timekeepers versus nonlinear oscillators: how the approaches differ. In *Rhythm Perception and Production*, ed. Peter Desain and Luke Windsor:9-33. Lisse: Swets and Zeitlinger.

Bello, Juan Pablo, Laurent Daudet, Samer Abdallah, Chris Duxbury, Mike Davies, and Mark B. Sandler. 2005. A Tutorial on Onset Detection in Music Signals. *IEEE Transactions on Speech and Audio Processing* 13, no. 5: 1035-1047.

Bello, Juan Pablo, Chris Duxbury, Mike Davies, and Mark B. Sandler. 2004. On the use of phase and energy for musical onset detection in the complex domain. *IEEE Signal Processing Letters* 11, no. 6: 553-556.

Bello, Juan Pablo and Mark B. Sandler. 2003. Phase-based note onset detection for music signals. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (Hong Kong): 49-52.

Berners-Lee, Tim. 1995. *RFC1866: Hypertext Markup Language - 2.0*. Internet Engineering Task Force. Accessed. Available from http://www.ietf.org/rfc/rfc1866.txt.

Bilmes, Jeff. 1993. Timing is of the Essence: Perceptual and Computational Techniques for Representing, Learning, and Reproducing Timing in Percussive Rhythm. Master's thesis, Media Lab, Massachusetts Institute of Technology.

Bregman, Albert S. 1990. *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA: MIT Press.

Brossier, Paul M. 2006. Automatic Annotation of Musical Audio for Interactive Applications. Ph.D. dissertation, Centre for Digital Music, Queen Mary, University of London. (http://aubio.piem.org/phdthesis)

Brown, Howard Mayer and Claus Bockmaier. *Tactus*. Accessed 13 December 2007. Available from http://www.grovemusic.com.

Brown, Judith C. 1993. Determination of meter of musical scores by autocorrelation. *Journal of the Acoustical Society of America* 94, no. 4: 1953-1957.

Cemgil, Ali Taylan and Bert Kappen. 2002. Integrating Tempo Tracking and Quantization using Particle Filtering. In *Proceedings of the International Computer Music Conference* (Gothenborg): 419-422.

Chafe, Chris and Michael Gurevich. 2004. Network Time Delay and Ensemble Accuracy: Effects of Latency, Asymmetry. In *Proceedings of the 117th Convention of the Audio Engineering Society* (San Francisco, CA): 1-7 (convention paper 6208).

Chafe, Chris, Bernard Mont-Reynaud, and Loren Rush. 1982. Toward an Intelligent Editor of Digital Audio: Recognition of Musical Constructs. *Computer Music Journal* 6, no. 1: 30-41.

Chernoff, John Miller. 1979. *African Rhythm and African Sensibility: Aesthetics and Social Action in African Musical Idioms*. Chicago: University of Chicago Press.

Clarke, Eric. 1995. Expression in performance: generativity, perception and semiosis. In *The practice of performance: studies in musical interpretation*, ed. John Rink:21-54. Cambridge, UK: Cambridge University Press.

Clarke, Eric F. 1999. Rhythm and Timing in Music. In *The Psychology of Music*, ed. Diana Deutsch:473-500. San Diego: Academic Press.

Clayton, Martin. 2000. *Time in Indian Music: Rhythm, Metre, and Form in North Indian Râg Performance*. Oxford Monographs on Music. Oxford: Oxford University Press.

Clayton, Martin, Rebecca Sager, and Udo Will. 2004. In time with the music: The concept of entrainment and its significance for ethnomusicology. *ESEM CounterPoint* 1: 1-45.

Cleveland, William S. 1993. *Visualizing Data*. Summit, New Jersey: Hobart Press.

Clynes, Manfred. 1983. Expressive Microstructure in Music, Linked to Living Qualities. In *Studies of Music Performance*, ed. J. Sundberg:76-181. Stockholm: Royal Swedish Academy of Music.

Collier, Geoffrey L. and James Lincoln Collier . 2002. A Study of Timing in Two Louis Armstrong Solos. *Music Perception* 19, no. 3: 463-483.

Collins, Nick. 2004a. Beat Induction and Rhythm Analysis for Live Audio Processing: 1st Year Ph.D. Report:26. (http://www.cus.cam.ac.uk/~nc272/papers/pdfs/report1.pdf)

Collins, Nick. 2004b. On Onsets On-the-fly: Real-time Event Segmentation and Categorisation as a Compositional Effect. In *Proceedings of the Sound and Music Computing (SMC04)* (IRCAM, Paris): (online proceedings). (http://recherche.ircam.fr/equipes/repmus/SMC04/scm04actes/P35.pdf)

Collins, Nick. 2005a. A Change Discrimination Onset Detector with Peak Scoring Peak Picker and Time Domain Correction. In *Proceedings of the 1st Annual Music Information Retrieval Evaluation eXchange (MIREX '05)*, ed. J. Stephen Downie (London): (online publication). (http://www.music-ir.org/evaluation/mirex-results/articles/onset/collins.pdf)

Collins, Nick. 2005b. A Comparison of Sound Onset Detection Algorithms with Emphasis on Psychoacoustically Motivated Detection Functions. In *Proceedings of the 118th Convention of the Audio Engineering Society* (Barcelona, Spain): 1-12.

Collins, Nick. 2005c. An Automated Event Analysis System with Compositional Applications. In *Proceedings of the International Computer Music Conference* (Barcelona).

Collins, Nick. 2005d. Using a Pitch Detector for Onset Detection. In *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR)* (University of London): 100-106.

Collins, Nick. 2006. Investigating Computational Models of Perceptual Attack Time. In *Proceedings of the 9th International Conference on Music Perception and Cognition* (Bologna): 923-929.

Cook, Nicholas. 2002. Epistemologies of Music Theory. In *The Cambridge history of Western music theory*, ed. Thomas Christensen:78-105 Cambridge: Cambridge University Press.

Cooper, Grosvenor W. and Leonard B. Meyer. 1960. *The Rhythmic Structure of Music*. Chicago: University of Chicago Press.

Cox, Trevor J. 2007. Bad Vibes: An Investigation into The Worst Sounds in the World. In *Proceedings of the 19th International Congress on Acoustics (ICA)* (Madrid, Spain): PPA-09-003.

Crook, Larry. 2005. *Brazilian Music: Northeastern Traditions and the Heartbeat of a Modern Nation*. Edited by Michael B. Bakan. ABC-CLIO World Music Series. Santa Barbara, CA: ABC-CLIO.

Davies, Matthew E. P. and Mark D. Plumbley. 2004. Causal Tempo Tracking of Audio. In *Proceedings of the 5th Annual International Conference on Music Information Retrieval (ISMIR)* (Barcelona): 164-169.

Davies, Matthew E. P. and Mark D. Plumbley. 2005. Beat tracking with a two state model. In *Proceedings of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*.

Desain, Peter. 1989. A connectionist and a traditional AI quantizer, symbolic versus syb-symbolic models of rhythm perception. In *Music, Mind, and Machine*, ed. Peter Desain and Henkjan Honing:83-98. Amsterdam: Thesis Publishers. Original edition, 1990 Music and the Cognitive Sciences Conference.

Desain, Peter. 1992. A (De)composable Theory of Rhythm Perception. In *Music, Mind, and Machine*, ed. Peter Desain and Henkjan Honing:101-116. Amsterdam: Thesis Publishers. Original edition, Music Perception Vol 9, No, 4, Summer 1992.

Desain, Peter and Henkjan Honing. 1992a. Tempo Curves Considered Harmful. In *Music, Mind, and Machine*, ed. Peter Desain and Henkjan Honing:25-40. Amsterdam: Thesis Publishers. Original edition, "Music and Time", edited by J. Kramer. Contemporary Music Review.

Desain, Peter and Henkjan Honing. 1992b. The Quantization Problem: Traditional and Connectionist Approaches. In *Music, Mind, and Machine*, ed. Peter Desain and Henkjan Honing:45-58. Amsterdam: Thesis Publishers. Original edition, Understanding Music with AI: Perspectives on Music Cognition, American Association for Artificial Intelligence.

Desain, Peter and Henkjan Honing. 1994. Advanced Issues in Beat Induction Modeling: Syncopation, Tempo, and Timing. In *Proceedings of the International Computer Music Conference* (Aarhus, Denmark): 92-94.

Desain, Peter and Henkjan Honing. 1999. Computational models of beat induction: the rule-based approach. *Journal of New Music Research* 28, no. 1: 29-42.

Desain, Peter, Henkjan Honing, and Klaus de Rijk. 1989. The Quantization of Musical Time: A Connectionist Approach. *Computer Music Journal* 13, no. 3: 150-167.

Dijkstra, E. W. 1959. A note on two problems in connexion with graphs. *Numerische Mathematik* 1: 269-271.

Disley, Alastair C. 2006. Timbral description of musical instruments. In *Proceedings of the 9th International Conference on Music Perception & Cognition (ICMPC9)*, ed. M. Baroni, A. R. Addessi, R. Caterina and M. Costa (Bologna, Italy): 61-68.

Disley, Alastair C., David M. Howard, and Andrew D. Hunt. 2006. Remote psychoacoustic testing using the Internet. *Journal of the Acoustical Society of America* 120, no. 5: 3125 (abstract only).

Dixon, Simon. 2001. Automatic Extraction of Tempo and Beat From Expressive Performances. *Journal of New Music Research* 30, no. 1: 39-58.

Dixon, Simon. 2006. Onset Detection Revisited. In *Proceedings of the 9th International Conference on Digital Audio Effects (DAFX)* (Montreal): 133-137.

Dixon, Simon and Werner Goebl. 2002. Pinpointing the Beat: Tapping to Expressive Performances. In *Proceedings of the 7th International Conference on Music Perception and Cognition (ICMPC)* (Sydney): 58-68.

Dixon, Simon, Werner Goebl, and Emilios Cambouropoulos. 2006. Perceptual Smoothness of Tempo in Expressively Performed Music. *Music Perception* 23, no. 3: 195-214.

Dixon, Simon, Fabien Gouyon, and Gerhard Widmer. 2004. Towards Characterisation af Music Via Rhythmic Patterns. In *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR)* (Barcelona): 509-516.

Dunlap, Knight. 1910. Reactions on Rhythmic Stimuli, with Attempt to Synchronise. *Psychological Review* 17: 399-416.

Duxbury, Chris, Juan Pablo Bello, Mike Davies, and Mark B. Sandler. 2003a. A combined phase and amplitude based approach to onset detection for audio segmentation. In *Proceedings of the 4th European Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS-03)* (London): 275-280.

Duxbury, Chris, Juan Pablo Bello, Mike Davies, and Mark B. Sandler. 2003b. Complex Domain Onset Detection for Musical Signals. In *Proceedings of the 6th International Conference on Digital Audio Effects (DAFx 03)* (London): 90-93.

Duxbury, Chris, Juan Pablo Bello, Mark Sandler, and Mike Davies. 2004. A Comparison Between Fixed and Multiresolution Analysis for Onset Detection in Musical Signals. In *Proceedings of the 7th International Conference on Digital Audio Effects (DAFx 04)* (Naples, Italy): 1-5.

Duxbury, Chris, Mike Davies, and Mark B. Sandler. 2003. Temporal Segmentation and Preanalysis for Non-linear Time-scaling of Audio. In *Proceedings of the Audio Engineering Society 114th Convention*: Preprint #5812.

Duxbury, Chris, Mark Sandler, and Mike Davies. 2002. A hybrid approach to musical note onset detection. In *Proceedings of the 5th International Conference on Digital Audio Effects (DAFx)* (Hamburg, Germany): 33-38.

Eck, Douglas. 2001. A Network of Relaxation Oscillators that Finds Downbeats in Rhythms. In *Proceedings of the International Conference on Artificial Neural Networks* volume 2130: 1239-1247.

Eck, Douglas. 2002a. Finding Downbeats with a Relaxation Oscillator. *Psychological Research* 66, no. 1: 18-25.

Eck, Douglas. 2002b. Meter Through Synchrony: Processing Rhythmical Patterns with Relaxation Oscillators. Ph.D. thesis, Department of Computer Science, Cognitive Science Program Indiana University.

Eck, Douglas. 2007. Identifying Metrical and Temporal Structure with an Autocorrelation Phase Matrix. *Music Perception* 24, no. 2: 167-176.

Eck, Douglas and Norman Casagrande. 2005. Finding Meter in Music Using an Autocorrelation Phase Matrix and Shannon Entropy. In *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR)* (University of London): 504-509.

Eggermont, Jos J. 2001. Between sound and perception: reviewing the search for a neural code. *Hearing Research* 157: 1-42.

FitzGerald, Derry. 2004. Automatic Drum Transcription and Source Separation. Ph.D Thesis, Engineering and Applied Arts, Dublin Institute of Technology.

Friberg, Anders. 1995. Matching the rule parameters of Phrase Arch to performances of "Träumerei:" A preliminary study. In *Proceedings of the KTH Symposion on Grammars for Music Performance*, ed. & J. Sundberg A. Friberg (Stockholm, Sweden), volume 37-44.

Friberg, Anders and Johan Sundberg. 1993. *Perception of just noticeable time displacement of a tone presented in a metrical sequence at different tempos.*

Friberg, Anders and Andreas Sundström. 1997. *Preferred swing ratio in jazz as a function of tempo.*

Friberg, Anders and Andreas Sundström. 1999. Jazz drummers' swing ratio in relation to tempo. *Journal of the Acoustical Society of America* 105, no. 2: 1330 (abstract only).

Friberg, Anders and Andreas Sundström. 2002. Swing Ratios and Ensemble Timing in Jazz Performance: Evidence for a Common Rhythmic Pattern. *Music Perception* 19, no. 3: 333–349.

Frieler, Klaus. 2004. Beat and Meter Extraction Using Gaussified Onsets. In *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR)* (Barcelona): 178-183.

Gabrielsson, Alf. 1982. Perception and Perfomance of Musical Rhythm. In *Music, Mind, and Brain: The Neuropsychology of Music*, ed. Manfred Clynes:159-169. New York: Plenum Press.

Gordon, John W. 1987. The Perceptual Attack Time of Musical Tones. *Journal of the Acoustical Society of America* 82, no. 1: 88-105.

Gordon, John William. 1984. Perception of Attack Transients in Musical Tones. Ph.D. Thesis, CCRMA, Department of Music, Stanford.

Gordon, John William and John M. Grey. 1978. Perception of Spectral Modifications on Orchestral Instrument Tones. *Computer Music Journal* 2, no. 1: 24-31.

Gouyon, Fabien. 2005. A Computational Approach to Rhythm Description: Audio Features for the Computation of Rhythm Periodicity Functions and their use in Tempo Induction and Music Content Processing. Ph.D. dissertation, Department of Technology, Universitat Pompeu Fabra.

Gouyon, Fabien, Lars Fabig, and Jordi Bonada. 2003. Rhythmic Expressiveness Transformations of Audio Recordings: Swing Modifications. In *Proceedings of the 6th International Conference on Digital Audio Effects (DAFx 03)* (London, UK).

Greenwald, Jeff. 2002. Hip-Hop Drumming: The Rhyme May Define, but the Groove Makes You Move. *Black Music Research Journal* 22, no. 2: 259-271.

Grey, John M. 1975. An Exploration of Musical Timbre. Ph.D. thesis, Psychology, Stanford.

Grey, John M. 1977. Multidimensional perceptual scaling of musical timbres. *Journal of Acoustical Society of America* 61: 1270-1277.

Grey, John M. and James Andrew Moorer. 1977. Perceptual evaluations of synthesized musical instrument tones. *Journal of the Acoustical Society of America* 62, no. 2: 454-62.

Grubb, Lorin and Roger Dannenberg. 1997. A Stochastic Method of Tracking a Vocal Performer. In *Proceedings of the International Computer Music Conference* (Thessaloniki, Hellas): 301-308.

Grubb, Lorin and Roger Dannenberg. 1998. Enhanced Vocal Performance Tracking Using Multiple Information Sources. In *Proceedings of the International Computer Music Conference* (Ann Arbor, MI): 301-308.

Hainsworth, Stephen Webley. 2004. Techniques for the Automated Analysis of Musical Audio. Ph.D. dissertation, Department of Engineering, Signal Processing Group, University of Cambridge.

Hainsworth, Stephen Webley and Malcolm Macleod. 2003. Onset Detection in Musical Audio Signals. In *Proceedings of the International Computer Music Conference* (Singapore): 163-166.

Harsin, Charles Andrew. 1997. Perceptual-center modeling is affected by including acoustic rate-of-change modulations. *Perception and Psychophysics* 59, no. 2: 243-251.

Hawking, Steohen W. 1988. *A Brief History of Time: From the Big Bang to Black Holes*. New York: Bantam Books.

Hawkins, Jeff and Sandra Blakeslee. 2004. *On Intelligence: How a New Understanding of the Brain Will Lead to the Creation of Truly Intelligent Machines*. New York: Owl Books / Henry Holt and Company.

Hoffman, P., L. Masinter, and J. Zawinski. 1998. *RFC2368: The mailto URL scheme*. Accessed. Available from http://www.ietf.org/rfc/rfc2368.txt.

Honing, Henkjan. 2001. From Time to Time: The Representation of Timing and Tempo. *Computer Music Journal* 25, no. 3: 50 - 61

Honing, Henkjan. 2006. Evidence for tempo-specific timing in music using a web-based experimental setup. *Journal of Experimental Psychology: Human Perception and Performance* 32, no. 3: 780-786.

Honing, Henkjan and Olivia Ladinig. 2008. The Potential of the Internet for Music Perception Research: A Comment on Lab-Based Versus Web-Based Studies. *Empirical Musicology Review* 3, no. 1: 4-7.

Howell, Peter. 1988a. Prediction of P-center location from the distribution of energy in the amplitude envelope: II. *Perception and Psychophysics* 43: 99.

Howell, Peter. 1988b. Prediction of P-center location from the distribution of energy in the amplitude envelope: I. *Perception and Psychophysics* 43: 90-93.

Huron, David. 2006. *Sweet Anticipation: Music and the Psychology of Expectation*. Cambridge, MA: MIT Press.

Ivry, Richard B. 2004. The neural representation of time. *Current Opinion in Neurobiology* 14: 225–232.

Iyer, Vijay. 1998. Microstructures of Feel, Macrostructures of Sound: Embodied Cognition in West African and African-American Musics. Ph.D. thesis, Technology and the Arts in the Graduate Division, University of California at Berkeley.

Iyer, Vijay, Jeff Bilmes, David Lester Wessel, and Matthew Wright. 1997. A Novel Representation for Rhythmic Structure. In *Proceedings of the International Computer Music Conference* (Thessaloniki, Hellas): 97-100.

Janker, Peter M. 1995. On the Influence of the Internal Structure of a Syllable on the P-Center-Perception. In *Proceedings of the 13th International Congress of Phonetic Sciences* (Stockholm), volume 2: 510-513.

Janker, Peter M. 1996a. Evidence for the p-center syllable-nucleus-onset correspondence hypothesis. *Zaspil (ZAS Papers in Linguistics)* 7: 94-124.

Janker, Peter M. 1996b. The Range of Subjective Simultaneousness in Tapping Experiments with Speech Stimuli. In *Proceedings of the ESCA Workshop on the Auditory Basis of Speech Perception* (Keele University): 204-207.

Jehan, Tristan. 2004. Event-Synchronous Music Analysis/Synthesis. In *Proceedings of the 7th International Conference on Digital Audio Effects (DAFx 04)* (Naples, Italy): 561-567.

Jehan, Tristan. 2005. Downbeat Prediction by Listening and Learning. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (New Paltz, NY).

Jensen, Kristoffer and Tue Haste Andersen. 2003. Real-time Beat Estimation Using Feature Extraction. In *Proceedings of the Computer Music Modeling and Retrieval Symposium (CMMR 2003)*: 13-22.

Klapuri, Anssi. 1997. Automatic Transcription of Music. Master's Thesis, Department of Information Technology, Tampere University of Technology, Finland.

Klapuri, Anssi. 1999. Sound onset detection by applying psychoacoustic knowledge. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (Phoenix, AZ), volume 6: 3089-3092.

Klapuri, Anssi. 2004. Signal Processing Methods for the Automatic Transcription of Music. Ph.D. dissertation, Department of Technology, Tampere University of Technology.

Kolinski, Mieczyslaw. 1973. A Cross-Cultural Approach to Metro-Rhythmic Patterns. *Ethnomusicology* 17, no. 3: 494-506.

Krumbholz, Katrin, Roy D. Patterson, Andrea Nobbe, and Hugo Fastl. 2003. Microsecond temporal resolution in monaural hearing without spectral cues? *Journal of the Acoustical Society of America* 115, no. 5: 2790-2800.

Lago, Nelson Posse and Fabio Kon. 2004. The Quest for Low Latency. In *Proceedings of the International Computer Music Conference* (Miami, FL): 33-36.

Large, Edward W. 1996. Modeling Beat Perception with a Nonlinear Oscillator. In *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society*: can't find (DS).

Large, Edward W. 2000. On synchronizing movements with music. *Human Movement Science* 19: 527-566.

Large, Edward W. and Mari Riess Jones. 1999. The Dynamics of Attending: How People Track Time-Varying Events. *Psychological Review* 106, no. 1: 119-159.

Large, Edward W. and John F. Kolen. 1994. Resonance and the Perception of Musical Meter. *Connection Science* 6, no. 2: 177-208.

Large, Edward W. and Caroline Palmer. 2002. Perceiving temporal regularity in music. *Cognitive Science* 26: 1-37.

Lee, Christopher S. 1985. The Rhythmic Interpretation of Simple Musical Sequences: Towards a Perceptual Model. In *Musical Structure and Cognition*, ed. Peter Howell, Ian Cross and Robert West:53-69. London: Academic Press.

Lerdahl, Fred and R. Jackendoff. 1983. *A generative theory of tonal music*. Cambridge, Mass: MIT Press.

Leveau, Pierre, Laurent Daudet, and Gaël Richard. 2004. Methodology and Tools for the evaluation of automatic onset detection algorithms in music. In *Proceedings of the 5th Annual International Conference on Music Information Retrieval (ISMIR)* (Barcelona): 72-75.

Lindsay, Kenneth A. and Peter R. Nordquist. 2006. A technical look at swing rhythm in music. *Journal of Acoustical Society of America* 120, no. 5: 3005 (abstract only).

Lindsay, Kenneth A. and Peter R. Nordquist. 2007. More Than a Feeling—Some Technical Details of Swing Rhythm in Music. *Acoustics Today*: 31-42.

Lindsay, Kenneth Alan. 2006. Rhythm Analyzer: A Technical Look at Swing Rhythm in Music. Master's thesis, Mathematics and Computer Science, Southern Oregon University.

London, Justin. 2004. *Hearing in Time: Psychological Aspects of Musical Meter*. Oxford: Oxford University Press.

Marcus, Stephen Michael. 1981. Acoustic determinants of perceptual center (P-center) location. *Perception and Psychophysics* 30, no. 3: 247-256.

Marolt, Matija, Alenka Kavcic, and Marko Privosnik. 2002. Neural Networks for Note Onset Detection in Piano Music. In *Proceedings of the International Computer Music Conference* (Gothenberg, Sweeden): 467-470.

Martinez, Wendy L. and Angel R. Martinez. 2002. *Computational Statistics Handbook with MATLAB®*. Boca Raton, FL: Chapman & Hall/CRC.

Master, Aaron S. 2006. Stereo Music Source Separation via Bayesian Modeling. Ph.D. Dissertation, Electrical Engineering, Stanford.

McAuley, J. Devin. 1995. Perception of Time as Phase: Toward an Adaptive-Oscillator Model of Rhythmic Pattern Processing. Ph.D. dissertation, Computer Science and Cognitive Science, Indiana University.

McGuiness, Andrew. 2005. Microtiming deviations in groove. Master's Thesis, Electronic Arts, Australian National University.

Moore, B. C. J. and B. R. Glasberg. 1996. A revision of Zwicker's loudness model. *Acta Acustica* 82: 335-345.

Moore, F. R. 1988. The Dysfunctions of MIDI. *Computer Music Journal* 12, no. 1: 19-28.

Moorer, James Andrew. 1975. On the Segmentation and Analysis of Continuous Musical Sound by Digital Computer. Ph.D. dissertation, CCRMA, Stanford. (http://ccrma.stanford.edu/STANM/stanms/stanm3)

Morton, John, Steve Marcus, and Clive Frankish. 1976. THEORETICAL NOTE: Perceptual Centers (P-centers). *Psychological Review* 83, no. 5: 405-408.

Nagai, Katsumi. 1996. *A study of a rhythm perception model*. Accessed 9/12 2006. Available from http://www.tsuyama-ct.ac.jp/kats/papers/kn8/kn8.htm.

Nava, Gabriel Pablo. 2004. Music beat and tempo tracking with Laplacian and Bayesian networks. Master's thesis, Information and Communication Engineering Department, Graduate School of Information Science and Technology, The University of Tokyo

Noll, A. M. 1969. Pitch determination of human speech by the harmonic product spectrum, the harmonic sum spectrum, and a maximum likelihood estimate. In *Proceedings of the Symposium on Computer Processing ing Communications*: 779–797.

Palmer, Caroline. 1997. Music Performance. *Annual Review of Psychology* 48: 115-138.

Palmer, Caroline. 2005. Time Course of Retrieval and Movement Preparation in Music Performance. *Annals of the New York Academy of Sciences*, no. 1060: 360-367.

Patel, Aniruddh D., Anders Lofqvist, and Walter Naito. 1999. The Acoustics and Kinematics of Regularly Timed Speech: A Database and Method for the Study of the P-Center Problem. In *Proceedings of the 14th International Congress of Phonetic Sciences* (San Francisco, California): xxx.

Paulus, J. and A. Klapuri. 2002. Measuring the similarity of rhythmic patterns. In *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR)*: 150–156.

Peeters, Geoffroy. 2005. Rhythm Classification Using Spectral Rhythm Patterns. In *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR)* (University of London): 644-647.

Pikovsky, Arkady, Michael Rosenblum, and Jürgen Kurths. 2001. *Synchronization: A universal concept in nonlinear sciences*. Edited by Boris Chirkov, Predrag Cvitanovi´c, Frank Moss and Harry Swinney. Cambridge Nonlinear Science Series. Cambridge: Cambridge University Press.

Pompino-Marschall, B. 1988. Acoustic Determinants of Auditory Rhythm and Tempo Perception. In *Proceedings of the Proceedings of the 1988 IEEE International Conference on Systems, Man, and Cybernetics*volume 2: 1184 - 1187.

Pressing, Jeff. 1983. Cognitive Isomorphisms between Pitch and Rhythm in World Musics: West Africa, the Balkans and Western Tonality. *Studies in Music* 17: 38-61.

Puckette, Miller S. 1991. Combining Event and Signal Processing in the Max Graphical Programming Environment. *Computer Music Journal* 15, no. 3: 68-77.

Rabiner, Lawrence R. and Bernard Gold. 1975. *Theory and Application of Digital Signal Processing*. Englewood Cliffs, New Jersey: Prentice-Hall.

Rapp-Holmgren, K. 1971. *A study of syllable timing*. KTH.

Reips, Ulf-Dietrich. 2002. Standards for Internet-based experimenting. *Experimental Psychology* 49, no. 4: 243-256.

Repp, Bruno Hermann. 1995. Expressive timing in Schumann's "Träumerei:"' An analysis of performances by graduate student pianistst. *Journal of the Acoustical Society of America* 98, no. 5: 2413-2427.

Repp, Bruno Hermann. 1998. A microcosm of musical expression. I. Quantitative analysis of pianists' timing in the initial measures of Chopin's Etude in E major. *Journal of the Acoustical Society of America* 104, no. 2: 1085-1100.

Repp, Bruno Hermann, W. Luke Windsor, and Peter Desain. 2002. Effects of Tempo on the Timing of Simple Musical Rhythms. *Music Perception* 19, no. 4: 565-593.

Roads, Curtis. 2001. *Microsound*. Cambridge, Ma: MIT Press.

Rossing, Thomas D., F. Richard Moore, and Paul A. Wheeler. 2002. *The Science of Sound (3rd edition)*. San Francisco.

Rothstein, William. 1989. *Phrase Rhythm in Tonal Music*. New York: Schirmer Books.

Rubine, Dean and Paul McAvinney. 1990. Programmable Finger-Tracking Instrument Controllers. *Computer Music Journal* 14, no. 1: 26-41.

Rushton, Julian. *Downbeat*. Accessed 13 December 2007. Available from http://www.grovemusic.com.

Ryan, Kenneth John, Joseph V. Brady, Robert E. Cooke, Dorothy I. Height, Albert R. Jonsen, Patricia King, Karen Lebacqz, David W. Louisell, Donald W. Seldin, Eliot Stellar, and Robert H. Turtle. 1979. The Belmont Report: Ethical Principles and Guidelines for the protection of human subjects of research, ed. Education Department of Health, and Welfare: The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (http://ohsr.od.nih.gov/guidelines/belmont.html)

Sachs, Curt. 1953. *Rhythm and Tempo: A Study in Music History*. New York.

Samsudin, Ng Boon Poh, Evelyn Kurniawati, Farook Sattar, and Sapna George. 2006. A Unified Transient Detector for enhanced aacPlus Encoder. In *Proceedings of the 120th Audio Engineering Society Convention (AES)* (Paris, France): Convention Paper 6807.

Sayed, A. H. and T. Kailath. 2001. A Survey of Spectral Factorization Methods. *Numerical Linear Algebra with Applications* 8: 467-496.

Scheirer, Eric David. 1995. Extracting Expressive Performance Information from Recorded Music. Master's thesis, Program in Media Arts and Sciences, School of Architecture and Planning, Massachusetts Institute of Technology.

Scheirer, Eric David. 1996. Bregman's Chimerae: Music Perception as Auditory Scene Analysis. In *Proceedings of the International Conference on Music Perception and Cognition*: 317-322.

Scheirer, Eric David. 1997. Pulse Tracking with a Pitch Tracker. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (Mohonk, NY).

Scheirer, Eric David. 1998. Tempo and Beat Analysis of Acoustic Musical Signals. *Journal of the Acoustical Society of America* 103, no. 1: 588-601.

Schloss, W. Andrew. 1985. On the Automatic Transcription of Percussive Music: From Acoustic Signal to High-Level Analysis. Ph.D. dissertation, Program in Hearing and Speech Sciences, Stanford.

Scott, Sophie. 1998. The point of P-centres. *Psychological Research* 61, no. 1: 4-11. (http://www.springerlink.com/content/lab4y564a93e0gkj/)

Seifert, Frank, Katharina Rasch, and Michael Rentzsch. 2006. Tempo Induction by Stream-Based Evaluation of Musical Events. In *Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR)*: 344-345.

Seppänen, Jarno. 2001a. Computational Models of Musical Meter Recognition. Master's Thesis, Department of Information Technology, Tampere University of Technology.

Seppänen, Jarno. 2001b. Tatum Grid Analysis of Musical Signals. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (New Paltz, New York): 131-134.

Shlain, Leonard. 1991. *Art & Physics: Parallel Visions in Space, Time & Light*. New York: Quill / William Morrow.

Slaney, Malcolm. 1993-1994. *Auditory Toolbox: A MATLAB Toolbox for Auditory Modeling Work*. Apple Computer.

Slaney, Malcolm. 1997. A Critique of Pure Audition. In *Computational Auditory Scene Analysis*, ed. Dave Rosenthal and Hiroshi Okuno:15. Mahwah, NJ: Lawrence Erlbaum Associates.

Slaney, Malcolm and William White. 2007. Similarity Based on Rating Data. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR)* (Vienna, Austria): 479-484.

Smith, Julius O., III. 1982. Synthesis of bowed strings. In *Proceedings of the International Computer Music Conference* (Venice): 308-340.

Smith, Julius O., III. 1983. Techniques for Digital Filter Design and System Identification with Application to the Violin. Ph.D. thesis, Electrical Engineering, Stanford University (CCRMA). (http://ccrma.stanford.edu/STANM/STANM/stanm14)

Smith, Julius O., III. 2007a. *Introduction to Digital Filters with Audio Applications*: W3K Publishing, http://books.w3k.org. (http://ccrma.stanford.edu/~jos/filters)

Smith, Julius O., III. 2007b. *Mathematics of the Discrete Fourier Transform (DFT) with Audio Applications*: W3K Publishing, http://books.w3k.org. (http://ccrma.stanford.edu/~jos/mdft)

Smith, Julius O., III. 2007c. *Physical Audio Signal Processing, August 2007 Edition*: online book. (http://ccrma.stanford.edu/~jos/pasp)

Smith, Julius O., III. 2007d. *Spectral Audio Signal Processing, March 2007 Draft*: online book. (http://www-ccrma.stanford.edu/~jos/sasp)

Smith, L. S. 1994. Sound segmentation using onsets and offsets. *Journal of New Music Research* 23, no. 1: 11-23.

Smith, Leslie S. and Dagmar S. Fraser. 2004. Robust Sound Onset Detection Using Leaky Integrate-and-Fire Neurons With Depressing Synapses. *IEEE Transactions on Neural Networks* 15, no. 5: 1125-1134.

Soraghan, Christopher J., Tomas E. Ward, Rudi Villing, and Joseph Timoney. 2005. Perceptual centre correlates in evoked potentials. In *Proceedings of the 3rd European Medical and Biological Engineering Conference (EMBEC 05).*

Stewart, Alexander. 2000. 'Funky Drummer': New Orleans, James Brown and the rhythmic transformation of American popular music. *Popular Music* 19, no. 3: 293-318.

Stockhausen, Karlheinz. 1957. ...wie die Zeit vergeht... ("...How Time Passes..." trans. Cornelius Cardew 1959). *Die Reihe* 3: 13-42 or 99–139.

Strogatz, Steven. 2003. *Sync: The Emerging Science of Spontaneous Order*. New York: Theia/Hyperion.

Sundberg, J., A. Askenfelt, and L. Frydén. 1983. Musical performance. A synthesis-by-rule approach. *Computer Music Journal* 7: 37–43.

Sundberg, Johan, Anders Friberg, and Roberto Bresin. 2003. Attempts to Reproduce a Pianist's Expressive Timing with Director Musices Performance Rules. *Journal of New Music Research* 32, no. 3: 317-325.

Supper, Ben, Tim Brookes, and Francis Rumsey. 2006. An Auditory Onset Detection Algorithm for Improved Automatic Source Localization. *IEEE Transactions on Audio, Speech & Language Processing* 14, no. 3: 1008-1017.

Takeda, Haruto, Takuya Nishimoto, and Shigeki Sagayama. 2004. Rhythm and Tempo Recognition of Music Performance from a Probabilistic Approach. In *Proceedings of the 5th Annual International Conference on Music Information Retrieval (ISMIR)* (Barcelona): 357-364.

Tanghe, Koen, Micheline Lesaffre, Sven Degroeve, Marc Leman, Bernard De Baets, and Jean-Pierre Martens. 2005. Collecting Ground Truth Annotations for Drum Detection in Polyphonic Music. In *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR)* (University of London): 50-57.

Temperley, David. 2000. Meter and Grouping in African Music: A View from Music Theory. *Ethnomusicology* 44, no. 1: 65-96.

Todd, Neil P. McAngus. 1994. The Auditory "Primal Sketch": A Multiscale Model of Rhythmic Grouping. *Journal of New Music Research* 23, no. 1: 25-70.

Todd, Neil P. McAngus. 1995. The Kinematics of Musical Expression. *Journal of the Acoustical Society of America* 97, no. 3: 1940-9.

Toiviainen, Petri and Tuomas Eerola. 2005. Classification of Musical Metre with Autocorrelation and Discriminant Functions. In *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR)* (University of London): 351-357.

Toussaint, Godfried. 2002. A Mathmatical Analysis of African, Brazilian and Cuban Clave Rhythms. In *Proceedings of the BRIDGES: Mathematical Connections in Art, Music and Science* (Towson University, Towson, MD): 157–168.

Toussaint, Godfried. 2003. Classification and Phylogenetic Analysis of African Ternary Rhythm Timelines. In *Proceedings of the BRIDGES: Mathematical Connections in Art, Music and Science* (University of Granada, Granada, Spain): 25-36.

Toussaint, Godfried. 2005. Mathematical features for recognizing preference in Sub-Saharan African traditional rhythm timelines. In *Proceedings of the 3rd International Conference on Advances in Pattern Recognition* (University of Bath, Bath, United Kingdom): 18-27.

Tufte, Edward R. 2001. *The Visual Display of Quantitative Information*. Cheshire, CT: Graphics Press.

Tzanetakis, George, Georg Essl, and Perry Cook. 2001. Automatic musical genre classification of audio signals. In *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)*: 205-210.

Tzanetakis, George, Ajay Kapur, W. Andrew Schloss, and Matthew Wright. 2007. Computational Ethnomusicology. *Journal of Interdisciplinary Music Studies* 1, no. 2: 1-24.

Villing, Rudi, Tomas Ward, and Timoney. 2007. A review of P-Centre models. In *Proceedings of the Rhythm Perception and Production Workshop (RPPW)* (Dublin): (abstract and presentation slides only). (http://cspeech.ucd.ie/twiki/bin/view/Main/VillingAbstract07)

Villing, Rudi, Tomas Ward, and Joseph Timoney. 2003. P-Centre Extraction from Speech: the need for a more reliable measure. In *Proceedings of the Irish Signals and Systems Conference* (Limerick): 6.

Vorberg, Dirk and Rolf Hambuch. 1978. On the temporal control of rhythmic performance. In *Attention and Performance*, ed. J. Requin. Hillsdale, NJ: Lawrence Erlbaum.

Vos, Joos and Rudolf Rasch. 1981. The Perceptual Onset of Musical Tones. *Perception and Psychophysics* 29, no. 4: 323-335.

Vos, Piet G., Jircaroní Mates, and Noud W. van Kruysbergen. 1995. The Perceptual Centre of a Stimulus as the Cue for Synchronization to a Metronome: Evidence from Asynchronies. *The Quarterly Journal of Experimental Psychology Section A* 48, no. 4: 1024 - 1040. (http://www.informaworld.com/smpp/content~content=a771340246~db=all)

Waadeland, Carl Haakon. 2001. It Don't Mean a Thing If It Ain't Got That Swing – Simulating Expressive Timing by Modulated Movements. *Journal of New Music Research* 30, no. 1: 23-37.

Waadeland, Carl Haakon. 2003. Analysis of Jazz Drummers' Movements in Performance of Swing Grooves – A Preliminary Report. In *Proceedings of the Stockhom Music Accoustics Conference (SMAC)* (Stockholm, Sweden): 573-576.

Wessel, David Lester. 1979. Timbre space as a musical control structure. *Computer Music Journal* 3, no. 2: 45-52.

Wessel, David Lester and Matthew Wright. 2002. Problems and Prospects for Intimate Musical Control of Computers. *Computer Music Journal* 26, no. 3: 11-22.

Wessel, David Lester, Matthew Wright, and Shafqat Ali Khan. 1998. Preparation for Improvised Performance in Collaboration with a Khyal Singer. In *Proceedings of the International Computer Music Conference* (Ann Arbor, Michigan): 497-503.

Widmer, G. 2002. Machine discoveries: A few simple, robust local expression principles. *J. New Music Res.* 31, no. 1: 37–50.

Widmer, Gerhard and Werner Goebl. 2004. Computational Models of Expressive Music Performance: The State of the Art. *Journal of New Music Research* 33, no. 3: 203–216.

Williams, Arthur B. and Fred J. Taylor. 2006. *Electronic Filter Design Handbook*. New York: McGraw-Hill.

Wing, A. M. and A. B. Kristofferson. 1973a. Response delays and the timing of discrete motor responses. *Perception and Psychophysics* 14: 5-12.

Wing, A. M. and A. B. Kristofferson. 1973b. The Timing of Interresponse Intervals. *Perception and Psychophysics* 13, no. 3: 455-460.

Wright, Matthew. 2002. Problems and Prospects for intimate and satisfying sensor-based control of computer sound. In *Proceedings of the Symposium on Sensing and Input for Media-Centric Systems (SIMS)* (Santa Barbara, CA): 1-6.

Wright, Matthew. 2005. Open Sound Control: an enabling technology for musical networking. *Organised Sound* 10, no. 3: 193-200.

Wright, Matthew and Edgar J. Berdahl. 2006. Towards Machine Learning of Expressive Microtiming in Brazilian Drumming. In *Proceedings of the International Computer Music Conference* (New Orleans, LA): 572-575. (http://ccrma.stanford.edu/~eberdahl/Projects/Microtiming)

Wright, Matthew, Ryan J. Cassidy, and Michael F. Zbyszynski. 2004. Audio and Gesture Latency Measurements on Linux and OSX. In *Proceedings of the International Computer Music Conference* (Miami, FL): 423-429.

Wright, Matthew, Amar Chaudhary, Adrian Freed, Sami Khoury, and David Wessel. 1999. Audio Applications of the Sound Description Interchange Format Standard. In *Proceedings of the Audio Engineering Society 107th Convention*: preprint #5032.

Wright, Matthew and Adrian Freed. 1997. Open Sound Control: A New Protocol for Communicating with Sound Synthesizers. In *Proceedings of the International Computer Music Conference* (Thessaloniki, Hellas): 101-104.

Wright, Matthew and Julius O. Smith, III. 2005. Open-Source Matlab Tools for Interpolation of SDIF Sinusoidal Synthesis Parameters. In *Proceedings of the International Computer Music Conference* (Barcelona): 632-635.

Yoshii, Kazuyoshi, Masataka Goto, and Hiroshi G. Okuno. 2005. Inter:D: A Drum Sound Equalizer for Controlling Volume and Timbre of Drums. In *Proceedings of the 2nd European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology (EWIMT)* 205-212.

Zanon, Patrick and Giovanni De Poli. 2003. Estimation of Parameters in Rule Systems for Expressive Rendering of Musical Performance. *Computer Music Journal* 27, no. 1: 29-46.

Zicarelli, David. 1998. An Extensible Real-Time Signal Processing Environment for Max. In *Proceedings of the International Computer Music Conference* (Ann Arbor, Michigan): 463-466.

Zwicker, Eberhard and Hugo Fastl. 1999. Information Processing in the Auditory System. In *Psychoacoustics: Facts and Models*: 23-60. Berlin: Springer.