Center for Computer Research in Music and Acoustics May 1984

Department of Music
Report No. STAN-M-17

PERCEPTION OF ATTACK TRANSIENTS IN MUSICAL TONES

by

John William Gordon

Research sponsored by

National Science Foundation
and
System Development Foundation

CCRMA
DEPARTMENT OF MUSIC
Stanford University
Stanford, California 94305

Center for Computer Research in Music and Acoustics May 1984

Department of Music
Report No. STAN-M-17

PERCEPTION OF ATTACK TRANSIENTS IN MUSICAL TONES

by

John William Gordon

A common twenthieth-century compositional technique is *klangfarbenmelodie*, accomplished by assigning successive notes of a melody to different instruments. Composers often find it necessary or desirable to realize such melodies synthetically on the computer by recording each instrument tone separately and then arranging the recorded tones in proper sequence. However, due to the way difference in attack characteristics are perceived, melodies realized in this manner will usually be found to display an uneven, rather than isochronous, rhythm. To obtain the desired isochronous rhythm, one must increase or decrease each tone's physical onset time so that its "moment of attack" is synchronous with the beat. The time between physical onset and this "moment of attack" is defined as *perceptual attack time* (PAT).

This dissertation poses three qustions concerning PAT: (1) How accurately can PAT be measured, (2) How well can PAT be predicted from a quantitative representation of the sound, and (3) Do individuals hear PAT differently? To address these questions, three experiments were run to obtain PAT measurements, one involving isochronous presentation of stimuli and two involving simultaneous presentations. Reliability of the data was confirmed, measurement accuracy was established, and subject consistency was evaluated.

The measurements were used to develop and test several PAT prediction models. Most models based the prediction of PAT on the time an absolute amplitude, relative amplitude, integration, or slope threshold was crossed. The slope-threshold models were the most successful, but it was necessary first to apply simple signal processing techniques to obtain useful calculations of envelope slope. A correlation of .995 between predicted and measure PAT values was obtained by using a model based on slope threshold plus a percentage of rise time and some minor adjustments.

Conclusions are as follows: (1) Measurement accuracy of PAT is dependent on the measurement procedure, and is limited by the discriminations of regularity and temporal order or by certain side effects, such as masking or fusion. (2) PAT predictions are accurate and can be practically applied. (3) Inconcsistencies among musically trained listeners in the way PAT is perceived are within expected limitations in discriminability.

This thesis was 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.

This research was supported by the National Science Foundation under Contract NSF BNS 77-22306-A2 and BNS 79-24646 and System Development Foundation under Grant SDF #345. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of Stanford University, any agency of the U. S. Government, or of sponsoring foundations.

PERCEPTION OF ATTACK TRANSIENTS IN MUSICAL TONES

John William Gordon, Ph.D.

Stanford University, 1984

Abstract

A common twentieth-century compositional technique is *klangfarbenmelodie*, accomplished by assigning successive notes of a melody to different instruments. Composers often find it necessary or desirable to realize such melodies synthetically on the computer by recording each instrument tone separately and then arranging the recorded tones in proper sequence. However, due to the way differences in attack characteristics are perceived, melodies realized in this manner will usually be found to display an uneven, rather than isochronous, rhythm. To obtain the desired isochronous rhythm, one must increase or decrease each tone's physical onset time so that its "moment of attack" is synchronous with the beat. The time between physical onset and this "moment of attack" is defined as *perceptual attack time* (PAT).

This dissertation poses three questions concerning PAT: (1) How accurately can PAT be measured, (2) How well can PAT be predicted from a quantitative representation of the sound, and (3) Do individuals hear PAT differently. To address these questions, three experiments were run to obtain PAT measurements, one involving isochronous presentation of stimuli and two involving simultaneous presentations. Reliability of the data was confirmed, measurement accuracy was established, and subject consistency was evaluated.

Responses of expert listeners were used to develop and test several PAT prediction models. Most models based the prediction of PAT on the time an absolute amplitude, relative amplitude, integration, or slope threshold was crossed. The slope-threshold models were the most successful, but it was necessary first to apply simple signal processing techniques to obtain useful calculations of envelope slope. Using a model based on slope threshold plus a percentage of rise time and some minor adjustments resulted in a correlation of .995 between predicted and measured PAT values.

Conclusions are as follows: (1) Measurement accuracy of PAT is dependent on the measurement procedure, and is limited by the discriminations of regularity and temporal order or by certain side effects, such as masking or fusion. (2) PAT predictions are sufficiently accurate to be practically applied. (3) Inconsistencies among musically trained listeners in the way PAT is perceived are within expected limitations in discriminability.

Acknowledgments

There are countless thoughts, impulses, opinions, and suggestions that influence one, either beneficially or detrimentally. I feel fortunate that I have learned, at least to some small degree, which of these to trust and which to reject. I have no doubt that my completion of this work and the Ph.D. degree is due primarily to a benevolent spiritual influence, beyond human interaction, that has guided me consistently from one step to the next. This influence has been perhaps too feebly and too infrequently glimpsed, but I give first and foremost credit to it.

I would be remiss if I did not also acknowledge human help I have received. I cannot express enough gratitude to John Chowning, my principal advisor. Without his efforts to establish the Center for Computer Research in Music and Acoustics (CCRMA), it is unlikely I would have undertaken the work described herein. The facilities and expertise available through CCRMA have been invaluable, and the assistantship made available to me through grant-awarded funds from NSF and System Development Foundation has been indispensable to my education.

Further gratitude goes to Dr. Earl Schubert, chairman of the Hearing and Speech Sciences Department at Stanford. Dr. Schubert has devoted much of his time as a second advisor to me, and has been especially helpful in the areas of statistics, experimental design, and auditory theory. I have found few people that are more kind, more available, or more responsive.

I would also like to thank the subjects that participated in the three timing experiments, and apologize for not being able to pay them for their services.

Finally, my deep and heartfelt thanks go to my dear wife, Marilyn, whose faith in my ability to finish this work has been at times stronger than my own. Marilyn has inspired me to work when nothing else could; her unflagging support, companionship, and love are the most delightful blessings a man could have.

Table of Contents

Chapter I

Introduction

In describing a musical sound, we might say that it has four principal characteristics: loudness, pitch, duration, and timbre (tonal color). These four terms describe specifically *perceptual* attributes of the sound. The first three have counterparts that describe *physical* attributes of sound, namely intensity, frequency, and (physical) duration; and although there is not an exact transformation from a sound's physical attribute to its perceptual one, we can say that loudness is primarily dependent upon intensity, pitch upon frequency, and perceputal duration upon physical duration.

Timbre has no such counterpart. In fact, as John Grey pointed out in his thesis [*Grey* (1975)], whereas loudness, pitch, and duration are essentially unidimensional measures, timbre is truly multidimensional. The timbre of a sound is perhaps most influenced by the spectral balance of its overtones (which in itself is really a multidimensional measure, since spectral balance can change throughout the sound's duration), but temporal aspects of the sound—especially attack characteristics,—can also influence the percept of its timbre.

Grey found that his subjects used three criteria (dimensions) in comparing the timbres of the brief orchestral instrument tones used in his studies; one dimension he interpreted to be related to spectral balance, and this interpretation was later verified quantitatively [*Grey and Gordon* (1978)]. The other two dimensions were interpreted as follows: one dimension appeared to be related to the presence or absence of low-amplitude, high-frequency onset noise preceding the attack; the

other dimension seemed to be related to whether or not all the harmonics changed in amplitude synchronously. As in the case of the first dimension, an attempt was made to develop models for predicting the coordinates of the set of tones along these other two dimensions. Such development proved difficult and complicated, however, due to the many parameters involved and the complex way they interacted; hence, the attempt to obtain any meaningful model was eventually abandoned. It became increasingly clear, however, that a better understanding is needed of the transient activity in a tone's attack and of how this activity influences timbre perception.

Some insight into transient perception was gained when Grey used the same set of recorded instrument tones to realize, on the computer, a composition entitled, "Loops," written by Robert Erickson, of San Diego. This piece employs the compositional technique known as *klangfarben-melodie,* or melody-of-changing-timbre, which is accomplished by assigning successive notes (or groups of notes) of a melody to different instruments. In the case of "Loops," the melody changed notes and instruments rapidly, and orchestral musicians found it difficult if not impossible to perform faithfully. Grey realized the piece synthetically by arranging his recorded tones of the various instruments in proper sequence. However, an unexpected phenomenon occurred in the result: though the rhythmic placement of each tone's physical onset time was executed carefully to the nearest millisecond, the melody displayed an uneven rhythm.

Since such care was given to physical detail, we can assume that the asynchronies arose due to the way the different attack characteristics of the individual instruments were perceived. In other words, arranging the physical onsets of certain tones isochronously to form an eighth-note pattern, say, does not necessarily imply that the tones will be perceived as even, or isochronous, eighth notes. If a tone's attack is quick, it might be heard as ahead of the beat, but a tone with a very gradual attack might be heard as behind the beat. To obtain the desired isochronous rhythm, one must increase or decrease each tone's physical onset time so that its "moment of attack," or perceived rhythmic emphasis, is synchronous with the beat.

1.1 Definition of PAT

Since perceptual "moment of attack" is not necessarily coincident with physical onset time, and since it is different for different instrument tones, it is a convenient focus for research in transient perception. The study of perceptual attack raises several questions at the outset: What is it about a sound that determines the relation between perceptual and physical attack? Is there a "moment of attack" for all sounds, or can a sound begin so gradually that only its presence is noticed, without any perception of rhythmic emphasis? Is perceptual attack objective, so that all listeners will agree as to when it occurs, or is it subjective? These and questions like them will be addressed in this dissertation.

Before proceeding, it is necessary to establish some definitions and conventions. The first definition is that of *physical onset time*. This definition can be somewhat arbitrary, so long as physical onset comes no later than *perceptual* onset, and so long as the definition is consistently applied. For the purposes of this dissertation, the physical onset of a tone will be defined as the time when the tone is first conceivably audible. In other words, physical onset time is essentially equated to perceptual onset time, though the actual temporal proximity of the two will depend on presentation level and possibly on envelope shape. *Perceptual attack time* (PAT) for a tone is defined as the time its perceptual "moment of attack" occurs, relative to the tone's physical onset. Thus, if a tone's rhythmic emphasis is perceived at a time 20 ms after its physical onset, its PAT is 20 ms. Note that a distinction is made between perceptual *attack* time and perceptual *onset* time. These two will often occur simultaneously, but for some tones it is possible to hear both an onset and a later "attack," especially in those produced by reeds, muted brass instruments, or bowed strings.

In some applications, it is important to know the absolute moment of perceptual attack; in other cases, all that is required is knowing when a tone's PAT occurs relative to the PAT for some other tone. For instance, to properly adjust the onsets of the tones used in "Loops," only relative information is needed; one tone can be chosen arbitrarily as a standard, with all other tones being adjusted relative to it. Let us then define *relative* perceptual attack time (RPAT) as the amount of time needed to shift a tone away from *physical* isochronism (with some standard tone) in order to attain *perceptual* isochronism (with that standard). *Absolute* perceptual attack time (APAT) is

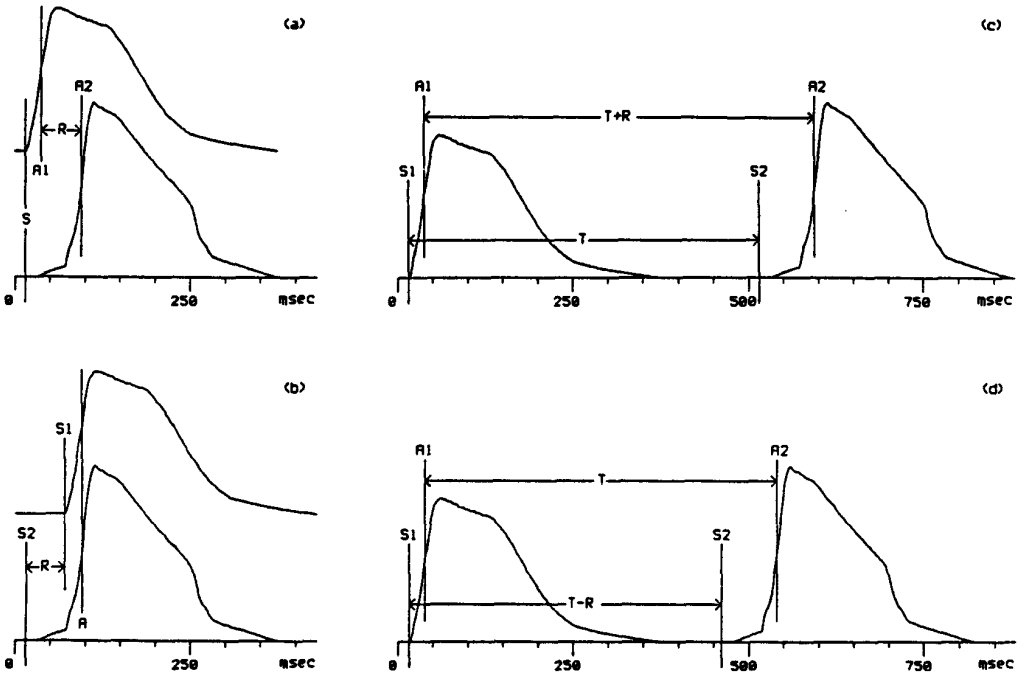

Figure 1.1 Example of amplitude envelopes for tones with different PATs. S1 & S2 (or S, when S1 = S2) represent the tones' physical onsets; A1 & A2 (or A) represent their APATs. (It is not apparent from the graphs that tone 2's envelope is > 0 for all $t > S2$.) (a): the tones are synchronized by physical onset, but their perceptual attacks are not perceived as simultaneous. R represents tone 2's RPAT relative to tone 1. (b): the same tones are now synchronized according to PAT, A, and physical onsets are thus separated by R. (c): Physical onsets are separated by the desired "beat period," T, but the perceived rhythm is not isochronous because the two PATs are separated by T+R. (d): Perceptual isochronism is attained by shortening the time between physical onsets to T-R.

given essentially the same definition as that for PAT, namely, the time a tone's perceptual "moment of attack" occurs, relative to its physical onset; however, APAT is a useful term for those instances when a specific distinction needs to be made between relative PAT and absolute PAT. A tone's APAT must be a nonnegative number, since perceptual attack never precedes physical onset. On the other hand, a tone's RPAT may be negative, positive, or zero, depending on what is chosen as the standard, and is also independent of the isochronal "beat rate." In Figure 1.1 can be seen examples of RPAT and APAT, as well as an illustration for each of the following four cases: physical synchrony, perceptual synchrony, physical isochronism, and perceptual isochronism.

Before going on to the next section, it should be clarified that PAT, as defined here, is not the same as what is usually referred to as "rise time." Rise time is a durational measurement, often

defined as the time it takes for a tone's amplitude envelope to go from 10% to 90% of maximum. PAT is not a duration, but a moment; and APAT is measured as the time-delay between physical onset and perception of attack.

1.2 Thesis Scope

Let us now consider three questions that will be addressed throughout this dissertation. The first is **(Q1)**: *How accurately can PAT be measured?* Since PAT is a subjective phenomenon, this question is essentially asking how accurately one can make a certain psychophysical measurement. Regardless of the particular experimental paradigm that is used, some limits of auditory discriminability will inevitably to come into play, and these in turn set a minimum margin of error around the desired measurement. One paradigm may have a smaller margin of error than another; but the margin can never really be made to go to zero. Of course, by increasing the sample size and analyzing the data statistically, one can raise the level of confidence concerning the measure's accuracy.

The second question is **(Q2)**: *How well can PAT be predicted from a quantitative representation of the sound?* (A quantitative representation might be a digital recording of the sound, stored in a computer as a discrete sequence of samples, or some data-reduced sequence such as amplitude envelope.) In other words, can a model be developed that will predict the PAT of a sound based on a set of parameters that describe the sound? The model should ideally be consistent with auditory theory so that it could be generalized to any sound; however, for certain applications, a practical model may be of more use than a purely theoretical one. Question 2 goes hand-in-hand with question 1, in that measurements of PAT are going to be used to develop the prediction model. If the accuracy of the measurements is in doubt, there is no way to determine the success of the model.

The third question is really an extension of the other two. It is, simply, **(Q3)**: *Do individuals hear PAT differently?* Clearly PAT is not completely free of subjectivity, if only because it is

perceptual in nature; what is at issue though is whether or not two or more individuals agree as to the time of a tone's perceptual attack. This question is related to question 1 in that accurate measurements of PAT will give some indication of the degree of agreement among subjects as compared to individual subject variation. It is also related to question 2 in that if there are consistent subject differences, reasons for the differences will need to be established and somehow incorporated into the prediction model. For instance, more than one aspect of a tone may influence the judgment of PAT, and individuals may weight these aspects differently. If so, the model becomes a weighted model with the weights tailored to individual preferences.

The remainder of this dissertation is focused on addressing the three questions just outlined. Chapter II presents an historical review of certain psychoacoustic topics, including models of PAT developed by other researchers, performance and perception of rhythm, auditory temporal acuity, and masking. Chapter III presents analysis of an experiment (Experiment I) designed to measure PAT my means of an isochronous paradigm, a discussion of the results, and initial attempts at formulating prediction models. The results from Experiment I indicated there was need for further research; Therefore, two more experiments (Experiments II and III) were run. These experiments employed a synchronous paradigm in an attempt to gain more accuracy. Chapter IV presents a thorough analysis of Experiments II and III, and culminates in the development of a prediction model whose output PAT values correlate very well with the measurement values.

As a means of verifying the prediction model, musical examples were constructed from the stimulus tones used in the three experiments, with physical onset times adjusted according to PAT values predicted by the model. Results from an informal experiment indicated that listeners agreed that the adjusted onset times yielded isochronous and synchronous rhythms. The prediction model proposed at the end of Chapter IV differs from one developed by another team of researchers; however, their model was not successful in predicting PAT for the set of tones used in this investigation. Therefore, to obtain further verification, the newly developed model was tested against their set of tones and was found to do quite well. Both of these verification procedures are discussed in Chapter V. Finally, Chapter VI presents conclusions and some suggestions for further research.

Chapter II

Historical Review of Relevant Psychoacoustic Research

The most directly relevant literature to this study is that dealing with models of PAT that have been developed by other researchers. It will be seen in the following section that the scopes of these models are all quite limited, and none of the models is based on experiments with real instrument tones. But these models serve as a convenient frame of reference, either as foundations on which to build, or as hypotheses that need to be tested or disproven.

Following the review of PAT models comes a discussion of psychoacoustic topics that seem relevant to the questions posed in Chapter I. These topics are categorized as: Perception and performance of rhythm, auditory temporal acuity, and masking and interaction effects. The final section of this chapter will deal with quantitative models of general auditory processing and transformation. A knowledge of these models will prove useful when assessing what aspects of a sound give rise to the perception of attack.

2.1 Models of Perceptual Attack Time

In the 1970's, several investigators conducted experiments that required the subject to judge whether or not the onset of one stimulus was synchronous with the offset of another stimulus. In an

attempt to explain the wide variance of subject response to this task, one investigator, M. J. Penner, developed a mathematical model for both PAT and perceptual *decay* time (PDT) that allowed for individual criteria differences [*Penner* (1975)]. Penner proposed a simple "running average" model, $y(t) = \int_0^\infty w(\tau)x(t-\tau)d\tau$, where $x(t)$ is the sensory input, $w(\tau)$ is an averaging mechanism that integrates sensory stimulation over time, and $y(t)$ is the time-varying perception. She then suggested that $PAT = \beta \cdot \Delta y$, where Δy represents the rise time (arbitrarily defined) of $y(t)$ and β is some percentage that is subject-dependent. (She also suggested a similar formula for PDT.)

Penner employed two subjects in her experiment, and found her model could predict the responses for both subjects, provided different values for β were used. She later repeated the same experiment with 19 subjects, and found wide subject variability [*Penner* (1978)]. The proposed model is useful, therefore, only in that it allows so many degrees of freedom (different values for two parameters for each subject). This implies that the answer to Q3, presented in Chapter I, is that no two individuals hear PAT in the same way,—which, if true, means that no general model for predicting PAT can be developed. However, the task involved in Penner's experiments (matching onset to offset) is rather unusual (not being important for either the perception or performance of musical rhythm), and considerably different from the task of synchronizing two onsets. In fact, it is both possible and plausible that the wide subject variance she found was due entirely to differences in the perception of offset. Also, Penner's stimuli were short synthetic bursts of either white noise or 1 kHz sine tones, with uniform attack envelopes; they can not be considered, therefore, as being representative of sounds heard in a conventional musical context. Hence, we should at this stage make no conclusion regarding the answer to Q3, and searching for a general prediction model for PAT is still worthwhile.

PAT is not a phenomenon limited to musical sounds; it also occurs in spoken syllables, and thus contributes to the perceptual rhythm of speech. Morton, Marcus, and Frankish studied PAT (they used the term "perceptual center," or "P-center") in spoken digits, and developed an experiment designed to measure it [*Morton, Marcus, and Frankish* (1976)]. (An application of the results from this experiment could be the synthesis of "spoken" phrases consisting only of digits (e.g., phone numbers), realized by concatenating recorded utterances of each digit. An accurate measurement of

each digit's PAT would enable one to properly adjust the physical onset time of each digit to yield natural sounding speech rhythm.) Stimuli in the experiment were recordings of spoken digits, and subjects were presented with stimulus pairs alternating periodically at a constant rate (A-B-A-B ...). The period between successive A's was fixed, but the delay between A and B (and thus between B and A) was variable and controllable by the subject. The task was to adjust this delay until A and B were perceived to be isochronous. All possible pairs of digits were presented.

A set of simultaneous differential equations was solved to obtain a P-center for each digit. The authors found that P-centers did not necessarily correspond to any articulatory, syllabic, or phonemic events in the utterance, and therefore appeared to be independent of strictly speech-related parameters. In a later paper, Marcus developed a formula for P-center that depended on the durations of initial consonant, vowel, and final consonant [*Marcus* (1981)]. The general formula was expressed as: $p = \beta C + \alpha V + k$, where p is P-center time, C represents initial consonant duration, and V represents the combined duration of vowel and final consonant. The constant k was included to indicate that only measurements of RPAT, and not APAT, were obtained for the set of stimuli. To obtain minimum error between his predicted and measured values, Marcus set β equal to 0.65 and α equal to 0.25.

Though the individual P-centers were found to be independent of specific speech-related events, the proposed model nonetheless includes phonemic durations as variables. It is questionable, then, whether this model can be applied to musical sounds, since such sounds cannot be partitioned into phonemic entities. However, the concept of PAT being proportional to stimulus duration (or a portion of stimulus duration) is one that can be considered in developing our own model for predicting PAT.

Very recently, Vos and Rasch conducted a study of PAT very similar to that of Morton, *et. al.*, using the identical A-B-A-B paradigm,—but with synthetic musical tones as stimuli instead of speech syllables [*Vos and Rasch* (1981)]. There were five stimuli, each being a complex, sawtooth-like waveform ($x(nT) = \sum_{k=1}^{20}(1/k)\sin(2\pi k f nT)$) modified by a simple amplitude envelope. The rise portion of each envelope was sigmoidal in shape (the second half of a cosinusoid), but the rise time for each stimulus varied from $\sim$5 to 80 ms. The decay portion was similar in shape to the rise portion; decay time was invariant across stimuli. All possible pairs of stimuli were presented

diotically through earphones.

Vos and Rasch hypothesized that PAT was attained when the rise portion of the amplitude envelope crossed a threshold that was relative to maximum amplitude. (This kind of threshold can be expressed either as a percentage of maximum amplitude, or as so many dB below maximum.) Vos and Rasch were able to calculate the value of the threshold for any two tones by means of a simple formula, whose dependent variables were the measured RPAT for the two stimuli and the rise times of their respective rise functions. The threshold thus calculated ranged from -14 to -18 dB (below maximum amplitude), though playing the stimuli at lower intensity levels brought the threshold somewhat closer to maximum (to -9 or -7 dB in some cases).

The Vos-Rasch model seems to be a good starting point, but it cannot be considered as the final word in PAT. First of all, the stimuli used in their experiments were so homogeneous in spectrum and envelope shape that further research would need to be done using real instrument tones, whose timbres and amplitude envelopes are not deterministic. Second, not only were the stimuli limited in number (5 in their first experiment, 3 in all others), but they were actually variations of the same stimulus waveform, differing only in rise time. Before any PAT model can be accepted as completely general, its validity should be tested against more (and spectrally different) tones. Finally, the measurement accuracy of RPAT in their experiments seems to be somewhat in question (this is discussed more fully in Section 3.2); this may be due to the particular paradigm used, but other paradigms (especially ones involving judgments of synchrony, *vs.* judgments of isochronism) should be tried in an attempt, at least, to make RPAT measurements more accurate.

2.2 Perception and Performance of Rhythm

PAT is integrally related to rhythm perception; indeed, PAT, as we have defined it, is what determines whether or not a rhythm is perceived as regular (even) or irregular (uneven). We might say that PAT is on a low level, while rhythm perception is on a higher level. In experimental paradigms such as those used by Morton, Marcus, and Frankish, and by Vos and Rasch, measurements of RPAT

are obtained based on judgments of isochronism, or regular rhythm. The greater the deviation from regularity that is tolerated,—in that an irregular rhythm is still perceived as regular,—the less accurate will be the measurements of RPAT thus obtained. Limits in the discrimination of temporal regularity, then, will set a limit to the accuracy of RPAT measurements,—at least when the stimuli are presented isochronously.

Though discrimination of regularity might improve with training, it is presumed to be listener-independent and essentially limited only by auditory acuity. To the extent this is true, discriminability can be considered an objective influence on judgment of regularity, and will be considered in the next section. However, there are other influences on regularity judgment that are more subjective. One of these is the lack of perfect rhythmic regularity in any musical performance. Not only are there systematic and very noticeable deviations from regularity, such as rubato or the deliberately uneven Viennese waltz rhythm, but there are also random, subtle deviations that aren't necessarily intended but yet may be audible. A musically trained subject, being used to rhythmic nuances in human performances, may be willing to tolerate considerable irregularity,—even if it exceeds his physical limits of discrimination,—and still perceive the rhythm as regular.

These higher-level influences on regularity judgment are important to the study at hand, but it is difficult to obtain any quantitative measures for them. The literature that considers them does so almost exclusively from a qualitative viewpoint. This includes a vast number of papers on the perception of time and/or duration, summarized well in [*Allan* (1979)] and in [*Fraisse* (1978)], and also on the perception of rhythm and tempo, which are discussed in [*Fraisse* (1982)]. The relevance of this literature to the study of PAT, then, is rather limited.

Accuracy in Rhythm Performance

The study of rhythm performance gives us no direct insight into the temporal acuity of rhythm perception. Some performers seem to display certain systematic deviations from regularity when playing rhythms notated as regular; however, it is unclear whether these deviations are intended, being called for by the particular style involved, or unintended. If unintended, such deviations may give us a clue as to the size of some of the subjective tolerances mentioned above.

Gabrielsson conducted a study in which three musicians were asked to play simple, notated rhythms, and to play them evenly. Results showed a tendency for 2 eighth notes to be played with a short-long relationship, and for an eighth-sixteenth-sixteenth pattern to be played with a long-short-long relationship [*Gabrielsson* (1974)]. One problem with this study is that the error in measurement of onset (which appears to have been ~5 ms) seems to be larger than the deviations reported (which were on the order of 1–2 ms). But even if we assume great consistency in the measurement procedure, lending more credence to the findings, these deviations are small in terms of perceptual limits (which will be discussed in the next section).

Bengtsson and Gabrielsson continued this analysis by conducting a later study in which the performers were asked to execute the rhythms of certain well-known melodies in various styles [*Bengtsson and Gabrielsson* (1980)]. Again, systematic deviations from regularity were found; however, since some of the styles used (such as jazz and ragtime) are inherently uneven, whereas others are not, it is difficult determining to what degree these deviations were intended.

The Gabrielsson studies were of deviations from isochronism; Rasch, on the other hand, investigated deviations from synchrony by professional performers [*Rasch* (1979)]. Recordings were made of musical passages played by three different professional ensembles: a string trio, a recorder trio, and a woodwind trio. The moment of attack was measured for each note for each performer, and an overall statistical measure of asynchrony was calculated for each ensemble. What was desired for the attack measure was the time of *perceptual* attack (that is, PAT), and not physical onset; however, there was no extant model of PAT at that time. (The research by Vos and Rasch into perceptual onset time [*Vos and Rasch* (1981)] came later and was in all probability motivated by this study.) Rasch thus adopted an operational model for attack time that was 15–20 dB below maximum amplitude,—the very relative threshold later proposed by Vos and Rasch. The reason for this choice was simply that a measure corresponding to 3 dB below maximum was too variant among tones, and a lower threshold (below 20 dB) was sometimes buried too much in background noise.

The basic finding of Rasch's study was that sharp-attack instruments (recorders) seem to be more synchronous than slow-attack ones (strings). We might expect an attack to be more

perceptually salient when it is sharp (having a rapid rise time), since it carries more rhythmic impact than a slow one. If so, the placement of such an attack is likely to be more critical for the perception of synchrony, and performers of instruments that produce this kind of attack may feel the need to concentrate more carefully on precise synchronization (or may even find it physically easier to synchronize than string players). It could also be argued that if the measurement model were flawed, the error would be more pronounced in the case of slow attacks, since their rise times are longer.

Rasch put forth his own reason why string ensembles may display less synchrony than woodwind players; and this reason involves more of a musical issue than a physical one. In an earlier paper, Rasch had found found that for two complex tones identical in the shape of their temporal envelopes but different in frequency, slight temporal asynchrony between the tones' onsets foiled the percept of fusion much more than spectral differences alone (that is, when the tones were presented simultaneously and temporal cues were missing). He had also found that even when the asynchrony was as great as 30 ms, the sounds were still perceived as being synchronous [*Rasch* (1978)]. Since string sounds tend to blend with each other quite easily (moreso than woodwind sounds), Rasch argued that players may aim for slight asynchrony (whether consciously or subconsciously) in order that individual parts be more transparent, yet still perceived as synchronous.

As we will see in the next section, human perception of precise synchrony for certain kinds of sounds is quite acute. When the sounds are from musical instruments, however, there is the possibliity of their fusing into one perceptual timbre. Though complete fusion occurs rarely in musical performance, it is relatively common in computer synthesis, and can greatly confound the perception of simultaneous attack. This issue will be discussed in more detail in Chapter IV.

2.3 Auditory Temporal Acuity

There have been many studies made of temporal acuity, most of them on a fairly low level. The studies relevant to our examination of PAT can be grouped into three main areas. The first

of these, temporal order judgments, pertains to judgments of synchrony. The second, perception of duration and/or rhythmic regularity, pertains more to isochronous judgments. The third area is discrimination of rise time, which is a topic that can pertain to judgments of both synchrony and isochronism. In connection with all of these areas however, it is worthwhile to see how temporal events might be coded neurally. Hence, a fourth section dealing with this topic will follow the other three.

Judgment of Temporal Order

The term "temporal acuity," is often used in the literature to refer to the ability to resolve the order of two nearly simultaneous sounds. This usage will be adopted throughout this dissertation. A very extensive list of references on this topic can be found in [*Michon* (1979)]; below will be discussed the references that are specifically pertinent to the study of PAT.

The classical study on temporal order perception was reported by Hirsh [*Hirsh* (1959)]. Different kinds of synthetic stimuli were used (sine tones, noise bursts, clicks), with rise times being about 20 ms, and non-click stimuli having durations of 500 ms. To determine the durational threshold for confident perception of order, Hirsh paired stimuli in various combinations; physical offsets were kept synchronous, while onset asynchrony was changed in multiples of 10 ms. The basic finding was that ~20 ms onset asynchrony is needed for a subject to determine the correct order of the stimulus pair. This contrasted with earlier findings that 2 ms is enough onset asynchrony to prevent a stimulus pair from being heard as one sound. Hirsh studied the perceptual effect of shortening the rise time, and found that order acuity improved (the threshold was reduced to ~10 ms in some cases).

Because Hirsh varied onset asynchrony in 10-ms increments, it is conceivable that his subjects' acuity was even better than 10 ms. Shortly after Hirsh's study, Broadbent and Ladefoged published a letter to the editor, reporting that they found acuity to improve with training [*Broadbent and Ladefoged* (1959)]. This training factor was also noticed by Patterson and Green (see below).

Homick, Elfner, and Bothe studied 12-ms tone bursts in presence of masking noise, and found the same distinction that Hirsh had found, namely that the durational threshold for order exceeds

that for detection (in the presence of backward masking) [*Homick, Elfner, and Bothe* (1969)]. The order threshold was also found to increase as the amplitude of the tone burst decreased (e.g., the threshold was greater than 50 ms for barely audible bursts in the presence of $\sim$70 dB_{SPL} noise). The authors thus agreed with a hypothesis Hirsh had proposed, that there is a heirarchy of perceptual levels involved with temporal cues, with order judgments occurring at a higher level than detection judgments.

A series of studies by David Green, along with other researchers, found temporal acuity to be as small as 2 ms,—though this was possible only with considerable subject training. The main reason for this improvement over Hirsh's findings is apparently because the duration of their stimuli was very short (10 ms or less). In fact, Patterson and Green found that acuity improved with shorter durations (using both sine tones and clicks as stimuli), and concluded that long-duration stimuli obscured onset cues [*Patterson and Green* (1970)]. In later studies, Green found temporal acuity for sine tones to be essentially independent of frequency [*Green* (1973)], and Weir & Green found almost no acuity dependency on frequency difference,—judgments being limited as much by frequency resolution as by temporal resolution [*Wier and Green* (1975)].

Efron ran an experiment in which the stimuli, called "micropatterns," were two tones of different frequency, combined in various ways [*Efron* (1973)]. Rise times were on the order of 5–10 ms. The experiment's purpose was to see how discrimination of micropattern was affected by changing various temporal cues. The cues were overall micropattern duration, simple asynchrony (delay of one tone with respect to the other), onset asynchrony (with synchronous offset), and offset asynchrony (with synchronous onset). In general, Efron's conclusions were that for a given asynchrony, discrimination improves as overall duration is decreased (which supports the findings of Patterson and Green), and that offset asynchrony alone affords better discriminability than onset asynchrony alone. Temporal acuity was found to be as good as 2–4 ms in some cases.

Results of a study conducted by Pastore, Harris, and Kaplan further corroborated the finding that order threshold decreases as stimulus duration decreases [*Pastore, Harris, and Kaplan* (1982)]. In their case, rise times were 0.5 ms and offsets were synchronous. The threshold for temporal order judgment ranged from 5 ms (for 10-ms stimulus durations) to 12 ms (for 300-ms durations).

The authors also investigated the effect of lengthening the rise time of one of the stimuli (thereby lengthening its overall duration), keeping the other rise time constant at 0.5 ms. They found almost no effect until the rise time exceeded 12 ms, and only slight effect thereafter: With the rise time set at 100 ms, order threshold was 8 ms for the original 10-ms duration, and 24 ms for the original 300-ms duration. It should be noted, however, that perceptual onset may be later than physical onset when the rise time is 100 ms; if so, the order thresholds reported would actually be slightly smaller.

The general findings of this subsection can be summarized as follows:

- With sufficient training, it is possible for the order of two sounds to be perceived when onset asynchrony is as small as 2 ms, regardless of their frequency. However, this durational threshold must be increased as the durations of the sounds are increased, their amplitudes are reduced, or their rise times are increased.
- There is some indication that offset asynchrony is easier to detect than onset asynchrony.
- Temporal order judgments probably occur more centrally than detection judgments.

Judgment of Duration and Regularity

Whereas temporal order threshold pertains to judgments of simultaneity, duration and regularity discrimination pertain to judgments of isochronism. In this section, three kinds of discrimination will be discussed that need to be distinguished from each other: those of duration, intermittency, and regularity. Duration discrimination involves the estimation of a single time interval; that of intermittency involves the estimation of tempo or "beat rate"—or even frequency, if the tempo is fast enough; and regularity discrimination involves the assessment of evenness of rhythm.

In a paper by Michon, a distinction is made between duration and intermittency judgments, and several studies are cited that had explored the discrimination of one of these two temporal phenomena [*Michon* (1964)]. A preliminary comparison indicated that acuity is better for intermittency judgments than for duration judgments. Michon ran his own intermittency experiment, using pulses less than 1 ms long, varying the interval between pulses from 67 to 2700 ms. When

the inter-pulse interval was between 100 and 250 ms, discriminability of intermittency was found to obey a Weber law, with the ratio being ~1%.* (I.e., subjects were able to discriminate differences in intermittency of ~1–2.5 ms.) The Weber ratio jumped to 2% for inter-pulse intervals between 300 and 1000 ms, and otherwise the Weber relationship did not seem to apply.

Whereas Michon studied discrimination of intermittency (and indirectly, duration), Lunney ran an experiment to study regularity discrimination [*Lunney* (1974)]. An electronic metronome was devised in which every fourth beat was irregular by a controllable amount, and overall tempo was also variable. Discrimination appeared to obey a Weber law in this case also, with the Weber ratio being ~4% (the beat period was less than 300 ms).

Judgment of duration was explored by Getty, whose data seemed to obey his "generalized Weber Law" model [*Getty* (1975)]. The model is based on the theory that discrimination variance is the sum of a residual, stimulus-independent variance (constant), and a variance based on a Weber relationship to actual stimulus magnitude—in this case, duration. The values that best matched his data were a Weber ratio of ~6% and a residual variance of ~100 (implying a lower limit of ~10 ms for discrimination deviation).

The studies discussed so far employed equal-amplitude stimuli, but it may be of interest in studying PAT to consider judgments of isochronism in which the two sounds differ in amplitude,—after all, musical accent is quite prevalent, and accent is usually correlated with amplitude. Fraisse studied this question in an experiment which is discussed in a recent paper on time perception [*Fraisse* (1978)]. Fraisse found that when a sound receives a musical accent, whether it be subjective or objective, its duration is perceived to be longer than it would be without the accent.

To summarize, all three kinds of time discrimination seem to obey Weber laws, but they vary as to the size of the Weber ratio. Acuity for intermittency is perhaps best of the three, being about 1–2%. Next would be acuity for regularity, being about 4%, and judgment of duration is worst,

*Weber's law states that perceptual discriminability between two stimuli in regard to a particular physical attribute is proportional to the magnitude of that attribute. (Examples of attributes in this context are frequency, amplitude, or duration.) Mathematically, this is expressed as $\Delta x/x = k$, where x is the attribute being compared and k is a constant. When the law holds, $\Delta x/x$ is referred to as the Weber ratio or Weber relationship.

being on the order of 6% or higher. In terms of regularity or duration judgments, a mixture of accented with unaccented sounds could tend to perturb the accuracy of the judgment somewhat.

Discrimination of Rise Time

We have seen that rise time influences judgment-of-order acuity [*Pastore, Harris, and Kaplan* (1982)]; and, although there has been no direct study of how rise time influences judgments of duration or regularity, we can assume that such an influence exists. It seems plausible, then, that rise time also affects PAT in some way.

As was mentioned in the introduction, rise time is a temporal duration, while PAT is a "moment," or temporal instant. However, at least some rise times are so short that they would have to be perceived as being instantaneous, primarily because of temporal integration. This perceptual moment then would correspond to PAT. For longer rise times, integration can no longer accommodate the entire duration into one percept, and the ear would hear the increase taking place over time. In this latter case, it is important to know whether there is a clear moment corresponding to PAT, or whether PAT becomes a temporal "range." In other words, is PAT more precisely defined the quicker the rise time? If perceptual moment of attack *is* much less precise when rise time is long, then placing a sound with such a rise time within a rhythmic phrase should be less critical for the perception of rhythmic regularity (and perhaps also musically less satisfactory) than placing one with a sharp, near-instantaneous attack. Since it is less critical for the *performance* of rhythmic regularity [*Rasch* (1979)], we must conclude either that it is very difficult to control the time of attack of such sounds, or that the placing is less critical perceptually as well.

A review of the literature dealing with rise time discriminability may give us a better idea of how rise time affects the perception of attack, and more specifically, PAT.

In a letter to the editor discussing masking experiments, Wright commented that too quick a rise time for the probe tone may artificially lower its detection threshold, since its onset would be signaled by an audible "thud," or click [*Wright* (1960)]. The paper reviewed an experiment run by Wright that measured "thud/no-thud" threshold for a tone's rise time, varying intensity level as a parameter. The rise time threshold measured for a 250-ms, 1 kHz sine tone, was 1 ms at 50 dB_{SPL}

and 5 ms at 80 dB_{SPL}. Thus a quick rise time is more apt to give an audible "thud" the more intensely it is played, which implies that the audibility of clicks is related to temporal integration of spectral energy.

As in the case of the three kinds of time discrimination examined in the previous section, discriminability of rise time is usually expressed in terms of a Weber ratio,—though Weber's Law is sometimes seen to break down for rapid rise times. In a paper presented orally, Tenney reported on a study of rise time discrimination for 1 second, 1 kHz sine tones [*Tenney* (1962)]. For rise times (t) greater than 8 ms, the JND (Δt) was found to be a Weber ratio ($\Delta t/t$) of 50%; for shorter rise times, JND was a difference of 3–4 ms. (Comparison was always made between t and $t + \Delta t$, not $t - \Delta t$.) Rise functions were straight lines, and rise times were presumably measured from 0 to 100% of maximum amplitude.

Pollack's findings for 2-second long, white noise stimuli were similar to Tenney's [*Pollack* (1963)]. The JND (measured in the same way as Tenney's JND) was determined to be a ratio between 1.5 and 4 ($\Delta t/t$ = 50–300%) for rise times between 5 and 500 ms, and JND was better expressed as a difference of 3–10 ms for rise times between 1 and 5 ms. (Again, comparison rise times were always greater than standard rise times.) In Pollack's experiment, rise functions were sigmoidal in shape, and rise times were measured from 10 to 90% of maximum amplitude.

It was mentioned above that a quick rise time and a slow rise time will result in different percepts, one sounding more like an "instant" and the other sounding more like a "duration." It is useful to ask whether or not the boundary between the two is categorical, which implies a sudden shift in perception from "instant" to "duration" as rise time increases across the boundary. Cutting and Rosner explored this issue by using sine and sawtooth stimuli modulated by envelopes with variable rise times, the quick rise times sounding "plucked" while the long ones sounded "bowed" [*Cutting and Rosner* (1974)]. The "pluck/bow" boundary was found to occur at about a 40-ms rise time, and since the best discrimination between adjacent rise times was found for those straddling this boundary, Cutting and Rosner concluded that the boundary was indeed categorical. In a slightly later study, it was found that the perceptual boundary could be moved somewhat through selective adaptation [*Cutting, Rosner, and Foard* (1976)].

Shortly after these two studies, van Heuven and van den Broecke conducted a rise-time discriminability experiment of their own, using as stimuli 1-kHz sine tones and noise bursts lasting 400 ms [*van Heuven and van den Broecke* (1979)]. Rise times tested were between 20 and 70 ms, and discriminability was found to be a straight Weber ratio of 20–30%. (For 200-ms long stimuli, the Weber ratio was slightly less.)

Rosen and Howell attempted to replicate the Cutting and Rosner study, but could not replicate their results [*Rosen and Howell* (1981)]. By obtaining the original stimuli from Cutting and Rosner, they discovered that the rise times displayed an artifact: the stimuli were not evenly spaced according to rise time, and in fact the two stimuli whose rise times straddled the 40 ms boundary were the farthest apart. Hence the reason for these two rise times being the most discriminable. Rosen and Howell's data, in support of the findings of van Heuven and van den Broecke, indicated that rise time discrimination obeys a Weber Law.

There seem to be two main points then concerning rise time discrimination:

- Discriminability of rise time appears to obey a Weber law (at least for rise times greater than $\sim$10 ms), but the value of the Weber ratio evidently depends upon the durations of the stimuli, and rise times imposed on tones seem to be more discriminable than those imposed upon noise.
- Perception of rise time does not make any sudden or categorical jumps as rise time goes from very brief to very long.

Neural Firing

Whitfield has indicated that there are certain cells in the auditory nerve that respond only to pulses, and not to tone-stimuli [*Whitfield* (1978)]. The reason for this is because these cells are connected to many fibers spread throughout the length of the basilar membrane, and all (or most) of these fibers will fire only when the stimulus is impulsive, or near-impulsive, in nature. It is not clear how many fibers contributing to these cells need to fire in order for the cells to respond (i.e., what the threshold is), so it is difficult to know how impulsive a stimulus must be in order to activate this time-generating mechanism. It does seem reasonable, however, that these cells contribute to the coding of rhythmic events.

2.4 Masking and Interaction Effects

There is a very extensive literature on masking, which has been reviewed recently in [*Zwislocki* (1978)]. Its relevance to the study of PAT is twofold. First of all, when two sounds are played simultaneously, and their perceptual attacks are intended to be synchronous, there may be mutual masking, or one sound might dominate the other. Interactions could also be such that one or both attacks are not heard as they would be if played separately.

A second aspect of masking that deserves our attention is backward masking and/or interruption of perceptual processing. These concepts have been explored in many experiments that treated them as interactions between two stimuli, a masker and a probe; but in this dissertation we will also consider how these interactions might occur within a single sound.

These two areas are dealt with individually below, followed by a brief discussion of how perception of offset may influence PAT.

Mutual Interaction of Two Sounds

In one of the relatively few masking studies using sine tones as both masker and probe, Samoilova obtained backward masking effects for interstimulus intervals (ISI) as large as 1 second [*Samoilova* (1959)]. The masker was a 1200-Hz tone lasting 300 ms, while the 1 kHz probe was only 20 ms long. No specific values for rise times were given, though the author indicated that they were "smooth." In the case of simultaneous masking (onset of probe delayed 0–280 ms with respect to masker onset), Samoilova found the greatest masking to occur when probe onset was close to, or simultaneous with, masker onset. This implies that transients, even if relatively smooth, mask better than steady-states.

Chistovich and Ivanova conducted a simultaneous masking experiment in which probe and masker were both clicks (passed through a telephone) [*Chistovich and Ivanova* (1959)]. Varying the clicks' ISI resulted in three separate perceptions. For ISI $<$ 1.5 ms, only one click was heard, but two distinct clicks were not heard until ISI exceeded 50–100 ms. Within these limits was heard "one click of changing timbre." In other words, Chistovich and Ivanova found horizontal fusion (two

clicks being integrated into the percept of one source) could occur over a relatively large range of ISI. (*Horizontal* fusion is to be distinguished from *vertical* fusion, which is a complex overtone structure being heard as a single pitch and timbre.)

Deatherage and Evans conducted an extensive masking study using monotic, dichotic, diotic, and mixed presentation conditions [*Deatherage and Evans* (1969)]. As in Samoilova's experiment, both masker and probe were sine tones, though in this case they were also the same in frequency (1 kHz). Probe duration was either 25 or 100 ms, masker duration was 500 ms, and rise times were 5 ms. The masker was presented about 3 dB more intensely than the probe. Of interest to the study of PAT is fact that when probe and masker were contiguous in time, there was considerable horizontal fusion (the onset of the second tone was not heard).

Concerning the mutual interaction of two tones, then, the following points can be made:

- If horizontal fusion can occur for two sine tones of the same frequency (presumably if the rise time of the masked tone is at least 5 ms), it seems reasonable to assume that it can also occur for two complex tones whose harmonics coincide.
- The transient portion of a tone is apt to be a better masker than the steady-state portion.
- Since two impulsive attacks can fuse into one, even when separated by 50 ms (or more), we might expect this to be true for slower attacks also. However, the horizontal fusion phenomenon found by Chistovich and Ivanova may occur only if both stimuli are identical, since the auditory system processes two identical stimuli separated by less than $\sim$50 ms as a single source stimulus followed by an echo in a reverberant environment.

Interactions Within One Sound

When two sounds are arranged in an isochronous rhythm, there is no masking of one stimulus over the other. But is it possible for a portion of one stimulus to mask another portion of the same stimulus? More specifically, can perceptual attack mask perceptual onset, as these terms are defined in Chapter I? Or, can the phenomenon of backward masking be said to apply to the transient portion of a single stimulus? As mentioned in the previous subsection, Samoilova found even relatively smooth transients to be effective simultaneous and backward maskers, at least for brief tonal probes

[*Samoilova* (1959)]. It seems reasonable, therefore, to consider the onset portion of a tone as the probe, and its attack portion as the masker. In this subsection, then, will be reviewed experiments on backward masking and on interruption of perceptual processing.

A unique approach to masking was taken by Gol'dburt, using 100-Hz sine tones as both masker and probe [*Gol'dburt* (1961)]. The masker was presented at 100 dB_{SPL} and lasted 400 ms; various durations were used for the probe. Gol'dburt found not only the probe's detection threshold in the presence of masking, but also the intensity threshold for which the probe's perceptual duration was equal to its unmasked perceptual duration. In other words, the perceived duration of the probe was shortened in the presence of masking; moreover, this effect was more pronounced than simple masking. As masker intensity decreased, perceptual shortening also lessened, but the effect was still strong at brief (15–20 ms) ISIs.

Gol'dburt also tried making the physical duration of the probe longer until its perceptual duration (under masking) equalled that of an unmasked tone. The results were inconclusive, but much lengthening was required for ISIs $\leq$ 36 ms. The author concluded that auditory excitation persists after stimulation has terminated, and that such excitation can be interfered with by a masking stimulus.

In a standard simultaneous/backward masking experiment, Wright employed a masker of narrow-band noise, lasting 600 ms, and a 1-kHz sine tone as probe, which ended 100 ms before the termination of the noise (its duration thus varied). All rise times were 10 ms. In the simultaneous masking case (probe durations $<$ 100 ms), detection threshold was found to be dependent on masker SPL, with threshold increasing as probe duration decreased; this finding is consistent with the theory of temporal summation. In the backward masking case (probe duration $>$ 100 ms), two different situations arose. When onset asynchrony (OA) between probe and masker was less than 50 ms, detection threshold was still found to be dependent on masker SPL, though slightly lower than in the simultaneous masking case; however, there was no such dependency when OA was between 50 and 500 ms. For the latter case, since threshold was greater than that predicted by temporal summation alone, Wright postulated that the threshold shift was due partly to masking and partly to integration. He further reasoned that since threshold was independent of masker SPL, the effect

of the masker was solely to shorten the perceptual duration of the probe, the amount of decrease being such that integration over this shortened duration would account for the probe's threshold shift.

It is not clear what is contributing to threshold shift in the "short" backward masking region (OA $<$ 50 ms), but the fact that threshold is dependent on masker level in this region implies there is an integration effect. It is also possible, though, that the ear is having difficulty in segregating source information; that is, the probe might be heard as being another random frequency within the noise band.

A few years after the studies by Gol'dburt and Wright, some other papers appeared that corroborated the hypothesis of perceptual duration being shortened under a masking influence. Liss found that the perceptual duration of a visual stimulus could be shortened in the presence of backward masking [*Liss* (1968)]. Patterson and Green, as mentioned earlier, found that onset cues tended to be obscured by long-duration stimuli, which implies that a stimulus can interfere with the processing of an earlier excitation [*Patterson and Green* (1970)]. Finally, in a series of papers, Massaro discussed the phenomenon of what he called preperceptual auditory images, or the persistence of excitation after a short stimulus (20-ms tone burst) has terminated.

Massaro's experiments were similar to those on backward masking, but with the task being identification, rather than detection. He found processing of tone burst information could be hampered by following the burst with a masking tone several milliseconds (up to 250) after the burst had ended, and concluded that persistence of excitation facilitated perceptual processing [*Massaro* (1970)]. In another experiment, the effect lasted only up to 160 ms [*Massaro* (1971)]. It was later pointed out that an auditory image is necessary only for brief stimuli, and that interruption would occur only if the two stimuli can't be integrated or processed together as one [*Massaro* (1972*a*)]. Massaro also ran an experiment with probes longer than 20 ms, and found that processing time is more critical than stimulus duration for successful identification [*Massaro* (1972*b*)].

Hence, there is ample evidence to support the theory that one stimulus can interfere with the perceptual processing of an earlier stimulus, with the degree of interference depending on the later

stimulus's masking potential. Interference seems to occur whether or not there is a silent ISI, or even if the "two" stimuli are actually different portions of the same stimulus. Some musical tones display audible (suprathreshold) activity, such as tonguing noise, before the large increase in amplitude that defines the attack portion. In this case, it seems likely that the amplitude increase is masking (or at least interfering with) the sound that comes before it, and that onset and attack aren't really heard as separate entities. The perceptual duration of the onset activity might be shortened by the attack, or there may simply be horizontal fusion of the two. In either case, however, processing of the onset portion would be initialized prior to the actual attack, thereby "warning" the ear of the attack's imminence. In the case of onset and attack occurring simultaneously, no such "warning" would be given, and PAT would presumably come later than otherwise.

Perception of Offset

A final consideration should be advanced concerning the effect that offset or decay time may have on PAT. As was seen above, several investigators supported the theory that a perceptual auditory image persists after the stimulus has terminated (affording continuity of perceptual processing),—but this alone does not imply any effect on onset perception. However, when two stimuli occur simultaneously, order may be inferred by whichever stimulus is heard to terminate last. If this inference conflicts with the actual order of onsets, it could have a confounding influence on the perception of synchrony.

In a series of experiments, Efron had subjects synchronize the onset of one stimulus with the offset of another. Some stimulus pairs were bisensory (such as one visual stimulus and one auditory stimulus), and when both stimuli were auditory, they were presented dichotically. Efron found that perception of onset is independent of stimulus duration, but that perception of offset *was* duration dependent, in that there seemed to be a minimum perceptual duration of $\sim$130 ms for durations less than 130 ms [*Efron* (1970*a*, *b*)]. In a later experiment that employed overlapping (synchronous or near-synchronous) stimuli, Efron noticed a postdominance effect; the later-ending stimulus allowed the subject to infer temporal order [*Efron* (1973)].

Efron's findings were contradicted by Penner, who duplicated Efron's experiment but failed to

replicate his results [*Penner* (1975)]. This led her to develop her simple PAT model, discussed in detail in Section 2.1, which allowed for significant individual variation. In a thorough review on the subject of minimum perceptual duration, Allan also contradicted Efron's findings [*Allan* (1976)]. All in all, therefore, the literature on offset perception is of little relevance to our study of perceptual attack time.

2.5 Models of Auditory Processing

In studying psychophysical responses to certain stimuli, it is important to take into account auditory transformations that are known or hypothesized to take place either peripherally or centrally. These transforms are represented as models of auditory processing, which are useful in developing predictions about responses under study and in comparing predicted values with empirical data.

In our case, we are dealing with the perception of onsets and attacks. Presumably the most important aspect of processing to consider is how the ear responds to large changes in an amplitude envelope. Of lesser importance are pitch and timbre influences on attack perception, though these are inherent in some of the processing models under consideration.

The ear must detect the presence of a sound (its onset) in order to determine its moment of attack. Of course, in many instances the two percepts will occur simultaneously, but clearly perceptual attack cannot precede detection, or perceptual onset. Therefore, we must take into account auditory threshold models, and assure ourselves that predicted PAT is not earlier than time of detection.

Another question that needs to be addressed is whether PAT is based primarily on peripheral responses, whether it is a decision occurring at the central level, or whether it is perhaps determined somewhere in between. In other words, can we predict PAT from the amplitude (or power) envelope directly, or must we first process a representation of the sound according to some auditory model?

If it *is* necessary to use such a model, which model is appropriate?

Since PAT is so closely related to amplitude changes, a model of loudness perception seems to be a logical choice for consideration. In this section, therefore, certain loudness models will be discussed in addition to a brief review of threshold models.

Threshold Models

The classic paper on temporal summation was presented by Zwislocki almost 25 years ago [*Zwislocki* (1960)]. The basic finding was that the ear integrates energy at audibility threshold levels (assuming constant stimulus amplitude) with a 200 ms time constant. Zwislocki indicated that integration time might be shortened with increasing intensity, and that spectral effects are unclear,—though integration seems to take place only over energy within a critical band.

The temporal integration model has been used and refined by many researchers. Penner, for instance, used essentially the same model in developing a simple predictor of PAT (see Section 2.1). More recently, Penner ran a masking experiment whose results implied that more than one kind of temporal integration may be taking place in auditory processing [*Penner* (1980)]. Two maskers (surrounding the probe with a variable duration of silence in between them) were used, being 500-μs bursts of wide-band noise. The probe was a 100-μs click passed through a bandpass filter, and presented in the 9–12 dB_{SPL} range. Two-masker threshold was ~8–9 dB above one-masker threshold when interburst interval (IBI) was more than 3 ms, but when IBI was less than 3 ms, the threshold difference was less than 9 dB. To explain this, Penner proposed a model in which a brief temporal integrator, $z(\tau)$, precedes the conventional one, $u(\tau)$, such that when a and b are close together ($<$ 3 ms), $u(a+b) < u(a)+u(b)$.

In general, then, we can conclude that there is evidence of temporal integration at low intensity levels, though it isn't clear what the precise value for the time constant is.

Loudness Models

In 1955, S. S. Stevens reviewed several experiments studying the loudness of noise and 1-kHz sine tones, and proposed a power-law model for loudness: $\mathcal{L} = kI^{0.3}$ [*Stevens, S.S.* (1955)]. (The

power law can be interpreted as loudness doubling for every 10 dB increase in intensity.) The model was found to hold well for a 1-kHz sine tone, and also for noise above 50 dB_{SPL}; but for low intensities, loudness of noise seemed to grow more rapidly than that of the tone.

McGill took issue with the derivation of Stevens's power law, remarking that loudness estimation varied widely among subjects, each estimation obeying some power law if a different constant was added to each response [*McGill* (1961)]. Exponents then ranged (for different subjects) from 0.07 to 0.35. McGill claimed there was a product relation between loudness and reaction time: $\mathcal{L} \cdot \mathcal{R} = k$, where $\mathcal{L}$ represents loudness, $\mathcal{R}$ represents reaction time, and k is a constant. Since reaction time seemed to be a more consistent measure among subjects than loudness estimation, an experiment was run to measure subjects' reaction time to a 1 kHz sine tone at various intensities; the results then were used to determine that the power-law exponent for loudness was between 0.15 and 0.25.

S. S. Stevens's value of 0.3 for the exponent was used in J. C. Stevens and Hall's model, but also included in that model was a factor for durations less than the "critical duration" of ~150 ms [*Stevens, J.C. and Hall* (1966)]. This factor was also found to obey a power law, with an exponent ~1.25 times the one for energy. The Stevens and Hall model thus was: $\mathcal{L} = k_1 T^{0.375} E^{0.3}$ (T = duration, E = energy, and k_1 is a constant),—provided T was less than the "critical duration,"—and $\mathcal{L} = k_2 E^{0.3}$ otherwise.

Concurrently with Stevens and Hall, Zwicker and Scharf took a more thorough but complicated approach, developing a model that could be applied to complex sounds [*Zwicker and Scharf* (1965)]. The model begins with a representation of the sound's physical spectrum, and converts it into an excitation pattern, taking into account masking patterns, critical band summation, and frequency mappings of the basilar membrane. The formula for calculating loudness from the excitation pattern is rather complex (and involves integration over the entire spectrum of the pattern), but the fundamental power-law relationship is apparent, with an exponent = 0.23. Zwicker later extended the model so that it could be used to calculate time-varying loudness, and indicated that integration over spectrum should precede a smoothing integrator (low-pass filter) in the time domain [*Zwicker* (1977)].

Zwislocki gave further support to a power law for loudness (his exponent being 0.27) when he refined his auditory summation model so that it could be used to calculate loudness [*Zwislocki* (1969)]. This model, however, is not as practical to implement on the computer as Zwicker's models.

In a recent timbre study, Grey and Gordon used the Zwicker-Scharf model with great success [*Grey and Gordon* (1978)]. Since it had already been implemented at CCRMA and seemed to apply well to complex stimuli (especially as a time-varying model), it was chosen over others as a loudness model to use in the current study. The exponent of 0.23 is not too different from ones proposed by other investigators. It is also reassuring that all 16 tones used in the timbre study, having been perceptually equalized for loudness [*Grey* (1975)], yielded near-identical maximum loudness values when processed by the time-varying Zwicker model. (Graphs of some of these loudness envelopes are shown in the next chapter.)

Thus, the "Zwicker transform" of a complex stimulus, representing a time-varying sensation of loudness for that stimulus, is a viable alternative to direct time-varying amplitude or power measurements as an input sound representation to the various models that might be developed for predicting PAT.

Chapter III

Empirical Research: Part I

The literature reviewed in Chapter II has provided some insight into some of the issues associated with PAT, but the three main questions posed in Chapter I need to be studied and analyzed in more detail. A requirement for this study is a set of PAT measurements obtained empirically; this set of values not only can give insight into Q1, dealing with the accuracy of PAT measurements, and Q3, dealing with subject variance, but also can be used to develop and test various PAT prediction models. The focus of this chapter will be some initial empirical research (Experiment I) designed to obtain such a set of PAT values.

The measurement of PAT can be accomplished either through judgments of isochronism or through judgments of synchrony. There are problems unique to each measurement method, and there is no way of knowing *a priori* which method will yield more accurate results. Hence, choosing one method over the other is arbitrary. When a computer system is used to control the operation of the experiment, however, there may be software considerations that favor one of the two methods; this was indeed the case for Experiment I. Combining stimuli synchronously would have required a sophisticated algorithm for merging two sets of ordered commands for the digitial synthesizer into one, newly-ordered stream. Since it was desired to allow the subject to control the onset time of one stimulus relative to the other by means of a slide potentiometer (and to set it such that the stimuli were perceived as synchronous), this algorithm would have had to operate in real time, based on the current setting of the potentiometer. For isochronous judgments, no merging of the two command

streams is necessary,—only the amount of "rest" between stimuli needs to be controlled in real time. For this reason, then, the measurement method chosen for Experiment I was based on judgments of isochronism.

3.1 Method

- *Stimuli*

Stimuli were 16 instrument tones representing different orchestral families. These were used by Grey in his thesis work [*Grey* (1975)] and by Grey and Gordon in their study of how spectral modifications affect the perception of timbre [*Grey and Gordon* (1978)]. The instruments represented by these tones were: bassoon (BN), French horn (FH), English horn (EH), 2 different oboes (O6 and O9), E-flat clarinet (EC), B-flat clarinet (BC), flute (FL), trumpet (TP), muted trombone (TM), 2 tenor saxophones (X9 and X6), soprano saxophone (SS), and 3 violoncello tones, played by the same performer and on the same instrument, but with 3 different bowings (V6, V3, and V7). The abbreviations given with each instrument will be used henceforth in this dissertation.

The 16 tones were recorded digitally at a sampling rate of 25600, using a 14-bit ADC. They were then analyzed into time-varying harmonic functions, which in turn were approximated by contiguous line segments. The resynthesized tones, using line-segment approximations, were perceptually identical to the original ones. These line-segment analysis functions allowed for simple and efficient additive synthesis by means of a digital synthesizer, using 20 bits for internal arithmetic processing and 14 bits for its DAC output. The additive-synthesis format also made it easy to adjust certain parameters, allowing the tones to be equalized (perceptually) for pitch, duration, and loudness. The pitch of all tones thus corresponded to E-flat above middle C (about 311 Hz), durations were in general about .32 seconds, and presentation level corresponded to $\sim$90 dB_A. The first non-zero sample of a resynthesized stimulus tone was designated as its physical onset.

- *Procedure*

The paradigm, which was the same one adopted by Marcus [*Marcus* (1981)] and Vos and Rasch [*Vos and Rasch* (1981)], can be briefly described as follows. Two stimuli, represented as **A** and **B**, were presented alternately and repetitively in an indefinitely long loop pattern (**A-B-A-B...**), with a variable amount of silence separating adjacent tones. A slider controlled the physical onset time of **B** relative to that of **A**, and the subject was instructed to adjust the slider until **A** and **B** were perceived to be isochronous. The respective durations of **A** and **B** were kept constant, as was the overall loop time (**A** to **A**); thus, the setting of the slider determined both the silence between **A** and **B** and the silence between **B** and **A**.*

This procedure differs slightly from the one used by Vos and Rasch, in that the offsets of **A** and **B** in their experiments were physically isochronous, which had the side effect of changing **B**'s duration as its time of physical onset changed. In all probability, this side effect was not intended, but rather caused by equipment restraints; in either case, however, Vos and Rasch reasoned that the perceptual durations of **A** and **B** would be equal when their perceptual onsets were isochronous.

Since the tones used in Experiment I were already perceptually equal in duration (see below), it seemed desirable to preserve their physical durations. Also, the Vos-Rasch tones all had identical offset envelopes, while the tones in this study had different decay characteristics. There is no way of knowing *a priori* which procedure gives more accurate results, but it is unlikely that it really makes any difference, since PAT is supposedly independent of duration [*Efron* (1970*a*,*b*)].

- *Apparatus*

The subject took the experiment in a quiet listening environment, seated at a desk equipped with a computer terminal and slider. Tones were played through a single loudspeaker having a wide, flat frequency response and situated about 8–10 feet directly in front of the subject. All operational procedure was under the control of a program run on the computer terminal; this program determined which pair of tones to play according to the trial number, and controlled

*The loop time was roughly twice the sum of the durations of **A** and **B**; thus, there was ample margin within which to made the adjustment.

the synthesizer so that the tones were played in the A-B-A-B loop pattern. The program also communicated with a microcomputer that periodically read the setting of the slider by means of an 8-bit ADC; this value was then used to reset the A-B (and B-A) silence immediately. At the beginning of each trial, the computer program instructed the subject to move the slider to one of the two extremes, and would not begin playing the A-B pair until the subject had done so. Subjects indicated satisfaction with the setting of the slider by typing <RETURN> on the terminal keyboard; the computer then prepared the synthesizer for the next trial. The subject had the option to take a rest at any point and resume the experiment later; inspection of the data verified that this did not affect the consistency of a subject's answers.

- *Design*

The E-flat clarinet, which seemed to have average attack characteristics, was chosen as a standard tone; the two stimuli for any trial thus consisted of one of the 16 tones (including the E-flat clarinet) being paired with the standard. For statistical reasons, it was desired to have 15 replications for each instrument, for a total of 240 trials per subject; trials were thus grouped into 15 sets of 16, each instrument being represented once within each set. The order of trials within a set of 16 was determined randomly.

The overall loop time (duration between successive onsets of A) was 1200 ms; the "beat rate" was thus half of this, or 600 ms (corresponding to a metronome marking of 100). At the end of each trial, the value of the slider was used to calculate the time difference between the physical onsets of the two stimuli. Subtracting 600 ms from this value resulted in a Δt that represented the instrument's RPAT, relative to the E-flat clarinet. This value was accurate to within 2 ms.

- *Subjects*

Nine subjects took the experiment, all of whom were experienced in computer music. Various musical backgrounds were represented, including performing, composing, and acoustical researchers. All subjects were considered to have well-trained ears.

Table III.1 Analysis of variance for Experiment I. All factors are significant at the 99.9% level of confidence.

Factor	Mean Square	df	F	p
Instrument	29,956	15	208.029	< 0.001
Subject	3,000	8	20.834	< 0.001
Inst × Subj	266	120	1.847	< 0.001
Error	144	2016		

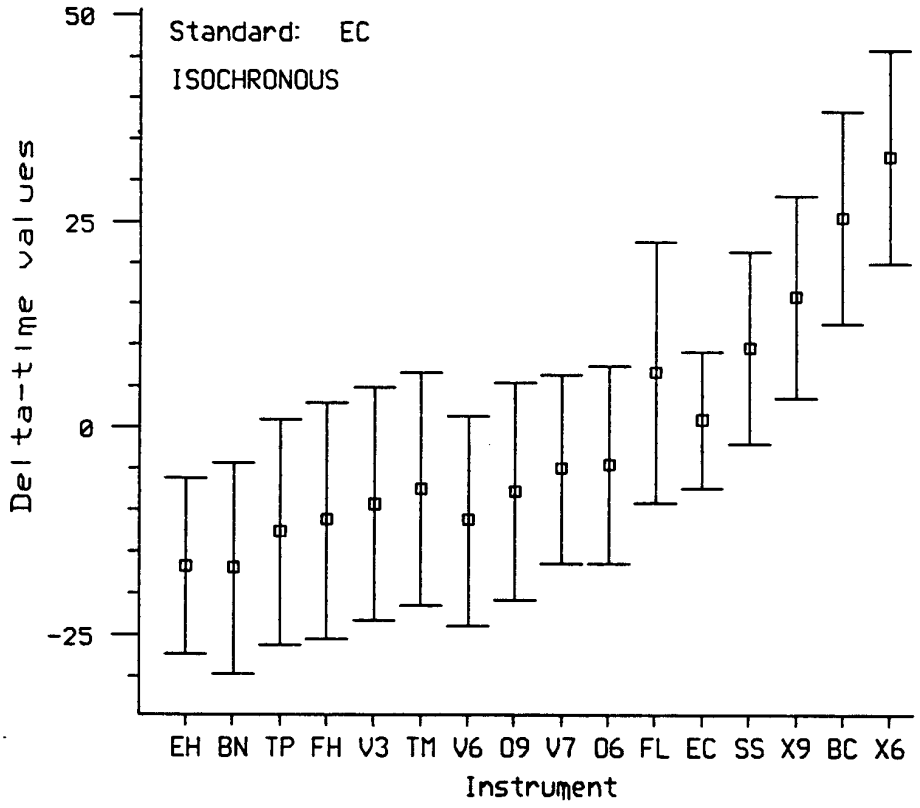

Figure 3.1 Mean Δt for 16 instrument tones, representing RPAT relative to the E-flat clarinet tone (EC). Vertical lines extend 1 standard deviation on either side of the mean.

3.2 Results

Initial examination of the data revealed a small number of outliers* that were more than 3 standard deviations away from the mean. These were replaced by the mean of that subject's

*_Outlier_ is a statistical term referring to a response separated from the main body of responses.

Table III.2 Mean Δt and standard deviation for the 16 tones in Experiment I. Two separate sets of physical measures of rise time are also given.

Instrument	Mean	Standard Deviation	Rise time	
			10–90%	0–100%
EH	−16.8	10.5	9.9	52.0
BN	−17.1	12.6	22.8	48.0
TP	−12.7	13.5	26.7	45.0
FH	−11.3	14.1	41.2	53.0
V3	−9.4	14.0	72.6	104.0
TM	−7.5	14.0	43.5	78.0
V6	−11.3	12.6	73.6	122.0
O9	−7.8	13.1	31.9	69.0
V7	−5.1	11.4	55.1	94.0
O6	−4.6	12.0	16.9	51.0
FL	6.6	15.8	75.7	105.0
EC	0.8	8.3	17.8	58.0
SS	9.5	11.7	11.8	63.0
X9	15.8	12.2	19.4	77.0
BC	25.3	12.9	31.3	113.0
X6	32.7	12.9	31.0	99.0

remaining responses. The data was then subjected to a 2-way analysis of variance (ANOVA), the factors being Instruments × Subjects. Results are shown in Table III.1.

The *F*-ratio* for instruments was 208 (df = 15,2016), indicating a very clear difference among them. The means for the 16 instruments are plotted in Figure 3.1, with marks to indicate ±1 standard deviation (SD) away from each mean. (The order of instruments along the X-axis is somewhat arbitrary; it coincides with the order used in later graphs and tables (in Chapter IV), in which the reasoning behind the chosen order is more apparent.) Values for the means and SDs are also listed in Table III.2.

Vos and Rasch observed that when the variable sound and standard were one and the same (all five tones were used as standards), their subjects tended to place the variable sound about 5

*Analysis of variance (ANOVA) is a common statistical procedure; if the reader is unfamiliar with ANOVA and the related concepts of *Mean Square*, *df* (degrees of freedom), *F-ratio*, and *p* (probability), an explanation can be found in practically any statistics text, such as [*Winer* (1971)].

ms too early, with respect to the standard. They indicated that this bias of the subject towards turning the knob consistently too far in one direction was found by several other investigators, and assumed that such a bias was also present when the variable and standard tones were not the same. They thus subtracted out the average bias from their set of Δt values, and used the corrected set of values to determine PAT [*Vos and Rasch* (1981)].

In Chapter II we saw that the Weber ratio for regularity discrimination is ~4% of beat period [*Lunney* (1974)]. That is, subjects tend to hear rhythmic phrases in which consistent deviations are less than 4% of beat period as regular. (Actually, 2% is probably a better value to use for our purposes—see the footnote on page 38.) Since the beat period in the Vos-Rasch experiment was ~400 ms (half the loop period between successive "A's"), we would expect the limit of discrimination in that case to be ~8–16 ms; a bias of 5 ms is thus not too surprising, and the corrections certainly seem justified. However, in the present experiment, as can be seen from Figure 3.1, the mean Δt value for EC, which is the same instrument as the standard, is very near 0 (0.8). We can conclude therefore that the group of subjects employed in Experiment I displayed no such bias phenomenon, and the Δt values do not need to be corrected.

Figure 3.2 shows each subject's set of Δt values (averaged over the 15 replications) individually. Table III.3 lists each subject's mean response (over all 240 trials) as well as his ranking among all subjects for each of the 16 instruments. It is clear that *individually,* subjects did show a bias towards being too early or too late, but that when taken as a group, the biases tended to average out. These biases are apparently the prime reason for the significant F-ratio (20.8, df = 8,2016) for subject variance. The reason for the significant instrument-subject interaction term can also be seen from Figure 3.2 and Table III.3: subject rankings across instruments are not consistent, though the inconsistency seems to be attributable to a small number of subject-instrument combinations.

Another point worth noting from Table III.2 and from figures 3.1 and 3.2 is that the variance from instrument to instrument is roughly constant. As one might expect, the standard deviation for instrument EC is the smallest (8.3); but the others range from 10.5 (EH) to 15.8 (FL), and most cluster around 12–13. We might conclude then that the discrimination was about the same regardless of which instrument was paired with the EC standard. Also, except for the responses of

Table III.3 Subjects are ranked in order of mean RPAT for each of the 16 tones used in Experiment I (isochronous judgments, with EC as the standard). Overall subject mean is given in parentheses after the subject's initials. The single character preceding each subject's initials is a key to the graph in Figure 3.2.

Subject (Mean)	EH	BN	TP	FH	V3	TM	V6	O9	V7	O6	FL	EC	SS	X9	BC	X6
M = JRM (5.8)	9	9	6	9	8	9	8	3	7	8	9	7	7	9	9	9
X = CGR (3.9)	8	6	9	8	9	8	9	9	8	9	2	9	9	3	3	2
C = CC (0.1)	3	8	8	7	7	2	5	4	9	5	7	8	5	4	5	3
P = PW (−0.6)	6	7	7	3	6	7	6	6	6	3	4	3	4	6	2	8
W = WJ (−2.3)	2	5	1	4	3	6	1	7	4	2	3	6	8	7	6	6
J = JG (−3.0)	5	4	4	5	2	5	4	5	5	4	1	2	6	5	4	7
S = SCH (−3.0)	7	2	3	2	1	4	3	8	1	7	6	5	2	8	7	4
Z = JMS (−3.9)	4	3	5	6	5	1	2	2	2	1	5	1	3	1	8	5
K = KIP (−4.3)	1	1	2	1	4	3	7	1	3	6	8	4	1	2	1	1

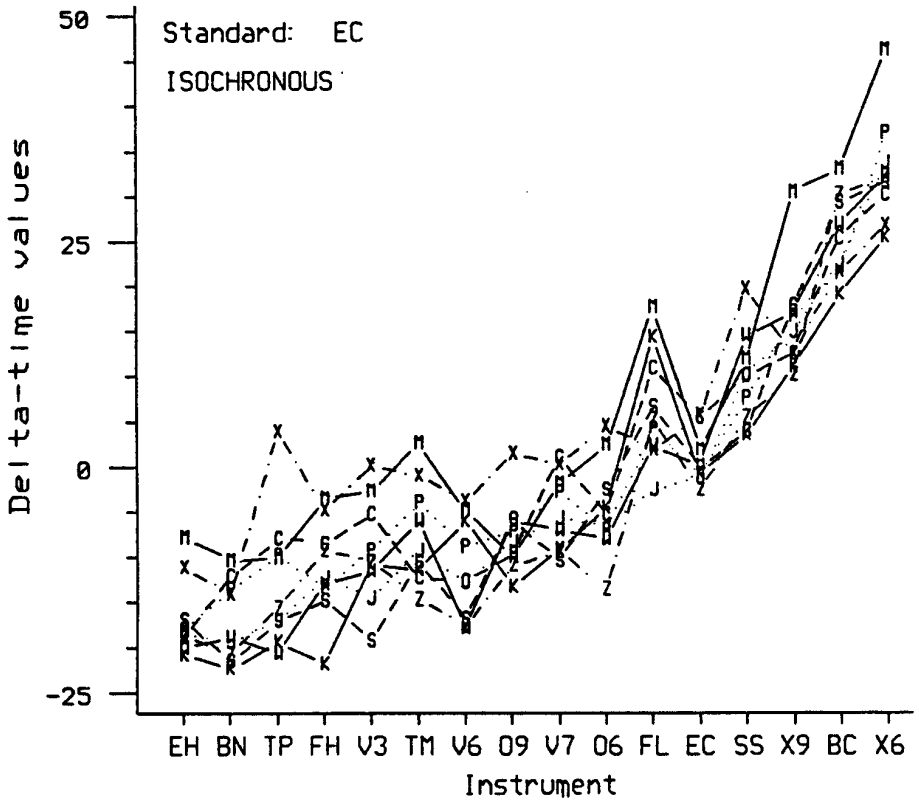

Figure 3.2 Mean Δt (RPAT) values for the 16 instrument tones used in Experiment I, plotted individually for each subject. Each letter corresponds to a different subject, as indicated in Table III.3.

subjects X (CGR) and M (JRM) in a few isolated instances, subject-instrument interaction seems to be an exchange of ranking more than a widening or narrowing of subject bias.

3.3 Discussion

We can now make an attempt to answer our three major questions, beginning with Q1, which deals with the accuracy of PAT measurements. If the standard deviation is used as an approximation of discriminability, we arrive at a 2% Weber ratio (12 ms = 2% of the 600-ms beat period). This is half of the 4% value obtained by Lunney, though the beat period in his study did not exceed 300 ms and his value was obtained by a different method* [*Lunney* (1974)]. Thus, the accuracy of Experiment I seems to be very good in terms of regularity discrimination, and may even be compared favorably with the discrimination of intermittency [*Michon* (1964)].

The measurements of RPAT from Experiment I, then, are probably as accurate as can be reasonably expected; however, they may not be precise enough for certain practical applications. For instance, we may wish to construct a musical phrase from the individual stimulus tones, intending to synchronize some of the tones' PATs (so that their attacks are perceived as being simultaneous). The accuracy of the RPAT, or Δt, values may not be good enough for this purpose, since temporal order can be perceived when perceptual onsets are separated by less than 10 ms in some cases, and asynchrony can presumably be noticed with even less separation [*Hirsh* (1959)].

The standard deviation of the RPAT values is relatively constant across the set of instrument tones. This implies that, at least for naturally produced tones, having temporally varying spectra and amplitude envelopes, the JND for PAT is essentially independent of rise time. Such independence can be seen from Table III.2, in which are listed the rise times of the 16 stimuli, measured according to two different, but common, criteria (10–90% of maximum, and 0–100%). For both criteria, there

*We might assume that Lunney's 4% value corresponds to the mean of a normally distributed set of responses, and that the spread of this distribution is such that the point of strict regularity is 2 standard deviations away from the mean (corresponding to the two-sided .05 level of significance). One SD would thus be equal to 2%.

is wide variability among the 16 rise times (10–75 ms in one case, 45–122 ms in the other). Both sets of rise time values correlate poorly with the RPAT set, resulting in product-moment correlation coefficients of .591 using the 10–90% criterion, and .334 using the 0–100% criterion.

Next, let us address Q3, dealing with individual differences. Table III.3 lists subject means, which range from −4.3 to 5.8. Despite the statistically significant differences among subject means (F-ratio of 20.8, df = 8,2016), the overall spread is relatively small—roughly half a standard deviation on either side of the instrument means (see Table III.2). This spread could easily be due to limits of regularity discrimination, although it does indicate that one subject tends to be always early while another tends to be always late. This tendency however can be explained by the nature of the paradigm involved. The continuous repetition of **A-B-A-B** can easily mesmerize one into believing that a slight irregularity is actually regular; indeed, at least one subject found that by switching his attention from listening to **A** as the "downbeat" tone to listening to **B** as the "downbeat" tone sometimes dramatically changed his judgment as to whether or not the rhythm was regular. Thus, we seem to have a simple bias phenomenon, though the magnitude and direction of the bias vary from subject to subject. These biases are within the theoretical limits of discrimination, and are more likely to result from different adjustment strategies than from any real differences in perceptual moment of attack. Furthermore, the spread among subjects, in comparison with the range of instrument means, seems to be small enough to warrant accepting the Δt values as valid data (that is, representative of the general population).

The F-ratio for the subject-instrument interaction term (1.847, df = 120,2016) is relatively small when compared to the other two F-ratios in Table III.1; nevertheless, this term is beyond the 0.001 level of significance. A possible reason for such interaction is that some instrument tones may display contrasting PAT cues and subjects are then assigning different weights to these cues in making their judgments of PAT. (Examples of contrasting cues are perceptual onset being separated by several milliseconds from perceptual attack, and salient spectral shifts not occurring simultaneously with large amplitude increase.) The weights a subject assigns would presumably be influenced by the subject's particular musical training, such as the instrument he plays or his conducting background.

Though there may be a slight indication of this kind of behavior in the data, there is definitely

nothing consistent. A more likely possibility for subject-instrument interaction is that certain instrument tones, specifically those with long rise time, might not lend themselves too well to precise placement within a rhythmic phrase. That is, for one of these tones, any value within a range of RPATs relative to the EC standard (rather than a particular Δt) might be equally satisfactory to subjects. If such is the case, the RPAT value selected by a subject would be arbitrary and dependent on adjustment strategy.

We will discuss these concepts more thoroughly in Section 3.5 and in Chapter IV. However, we might expect subjects to agree on RPAT values more readily if their judgments were based on instrument tones presented within a musical context rather than in isolation, and for now the interaction term really need not concern us.

The remaining question, Q2, asks how well PAT can be predicted from a quantitative representation of the sound. Before one can answer this question properly, two requirements must be met: (1) the formulation of at least one PAT prediction model, (2) reliable data against which the predicting capabilities of the model can be tested. The analysis of Experiment I indicates that the second of these needs (reliable RPAT measurements) has been met. In the next section will be presented several PAT models, as well as a discussion of their viability.

3.4 Prediction of PAT

Deriving a formula for predicting PAT involves making several decisions. The first decision is a selection of an appropriate representation of the sound. Is it sufficient to use the sound waveform itself (as it is stored in the computer), or should the formula be applied to some "transform" of the sound? If processing seems desirable, what kind of transform do we select?

Another decision that must be made involves choosing among the many models that might be appropriate for predicting PAT. Even after this decision has been made, however, one must still assign quantitative values to model parameters. That is, most models would predict PAT based on

the crossing of some threshold, for instance, but there may not be any *a priori* way of determining the value or even the nature of such threshold. Choosing an appropriate value is complicated by the fact that the empirical data are only relative values (examples of RPAT), whereas the models will be used to predict absolute values (APAT); if two thresholds predict sets of times (APATs) that are equal except for an additive constant, there is no way of choosing one threshold over the other (except by means of other criteria).

In short then, arriving at a successful prediction formula involves selecting:

- a proper representation of the sounds for which we wish to predict PAT;
- a model, presumably based upon the crossing of some threshold, which will accurately predict APAT when applied to the chosen sound representation,
- a quantitative value for the threshold that *best* predicts APAT.

The first of these issues is discussed in the following subsection on sound representations. A separate subsection will be devoted to the development of actual models for predicting perceptual attack time.

Sound Representations

It was mentioned in Chapter II that certain neural cells in the ear respond only to pulse-like stimuli, and that these may contribute to central coding of rhythmic events [*Whitfield* (1978)]. It seems reasonable to conjecture, then, that a sound's PAT is determined primarily, if not completely, by a change in its amplitude—or possibly change in intensity—that is rapid enough to trigger a firing of these cells. Let us focus then on the amplitude (intensity) characteristics of the 16 stimuli and—at least for the time being—ignore their timbral aspects.

Had the sounds used in this experiment been synthetic, their amplitude envelopes could have been easily formulated; however, since the stimuli were recorded examples of real instrument tones, a method had to be developed to obtain envelopes from these recordings. The intensity envelopes were approximated by squaring the amplitude envelope functions; this approximation is based on the proportionality of intensity and amplitude squared.

The algorithm used to generate amplitude envelopes was quite simple, and can be illustrated by referring to Figure 3.3. In the upper graph of this figure is a representation of the waveform from the French horn tone used in this study, somewhat typical of all 16 sounds. Its frequency (~313 Hz), and therefore also its period (~3.2 ms), is essentially constant throughout the entire duration of the tone; in the digital representation, this period translates into ~82 samples, since the sound was digitized at a sampling rate of 25600.

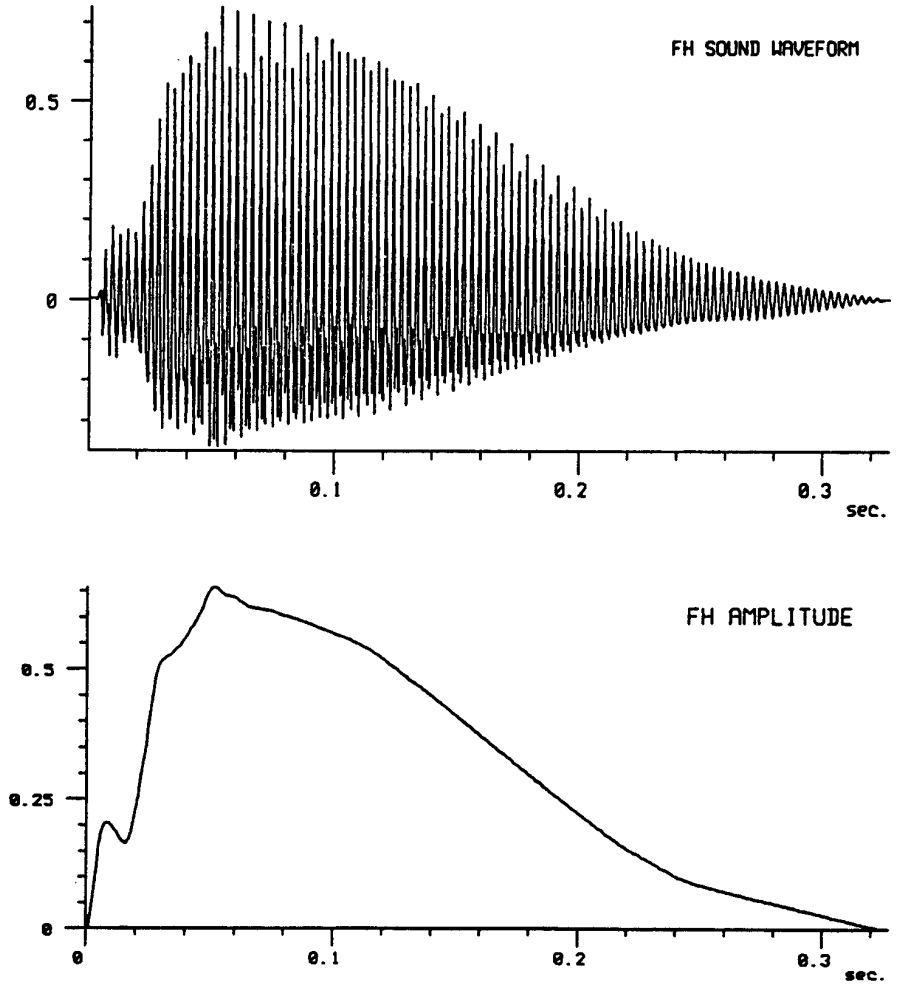

Figure 3.3 Two representations of the same French horn tone used as a stimulus in Experiment I. The upper graph shows the recorded tone as it might appear on a high-resolution oscilloscope; the tone's frequency is ~313 Hz, and its duration is .327 seconds. The outline of the same tone's peak amplitude is pictured in the lower graph. The units for both graphs are (normalized) amplitude *vs.* time (in seconds).

The basic idea behind the formulation of envelope from waveform representation was simply to connect maximum, or peak, values from period to period. However, as can be seen from Figure 3.3, the positive and negative peaks may not be symmetrical about equilibrium. The amplitude envelopes thus were formed by averaging the magnitudes of the maximum and minimum values within each period and connecting the averages together into a new digital sequence. This sequence had a sample rate of 312 (25600/82). For the convenience of having a sample every millisecond, this sequence was resampled to a new sampling rate of 1000, using a method suggested by Schafer and Rabiner [*Schafer and Rabiner* (1973)].

For comparison purposes, the amplitude envelope thus obtained for the French horn is displayed in the lower graph of Figure 3.3. The envelopes for all 16 stimuli are shown in Figure 3.4. It was more practical to use a normalized scale (0 to 1) for these functions than to try to scale them according to absolute sound pressure level. The maximum on this scale (1.0) corresponds to the greatest producible amplitude by the digital synthesizer; hence, as a protection against clipping (distortion), the maximum amplitude values for the individual instruments are somewhat less than 1.0. These maxima also range between $\sim$0.32 and 0.8; the variation is due to the tones being equalized perceptually, according to loudness, rather than physically, according to intensity.

A normalized scale was also used for the approximated intensity envelopes, which will be referred to as power functions. These functions are exact squares of the amplitude envelope functions, and are displayed in Figure 3.5.

Since rhythmic coding could be based more on loudness than on amplitude directly, it seems necessary to include a loudness model as one of our sound representations. In Section 2.5, certain loudness models were reviewed, with the "Zwicker transform" being chosen as the most appropriate one for use in this study.

The Zwicker transform converts a spectral representation of a sound into a loudness representation (either static or time-varying). Thus, to implement the transform, one must obtain a spectral representation from the sound waveform. Furthermore, to implement the *time-varying* Zwicker transform, a simple Fourier transform is insufficient—one must obtain a time-varying

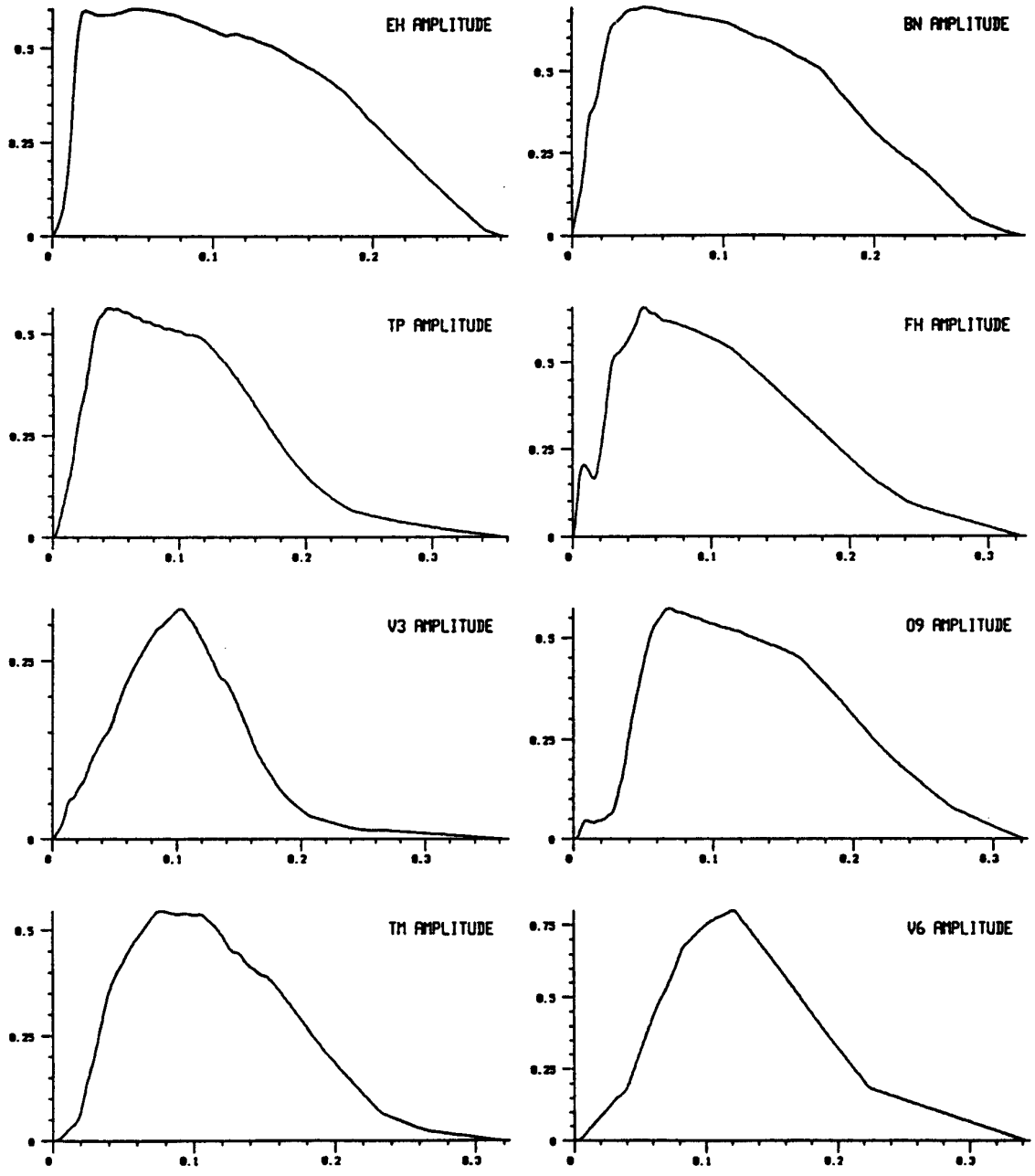

Figure 3.4 *Part 1.* Amplitude envelopes for the 16 stimulus tones, obtained from waveform representations by means of a simple algorithm (see text). Units are normalized amplitude *vs.* time in seconds. 8 of the envelopes are shown above.

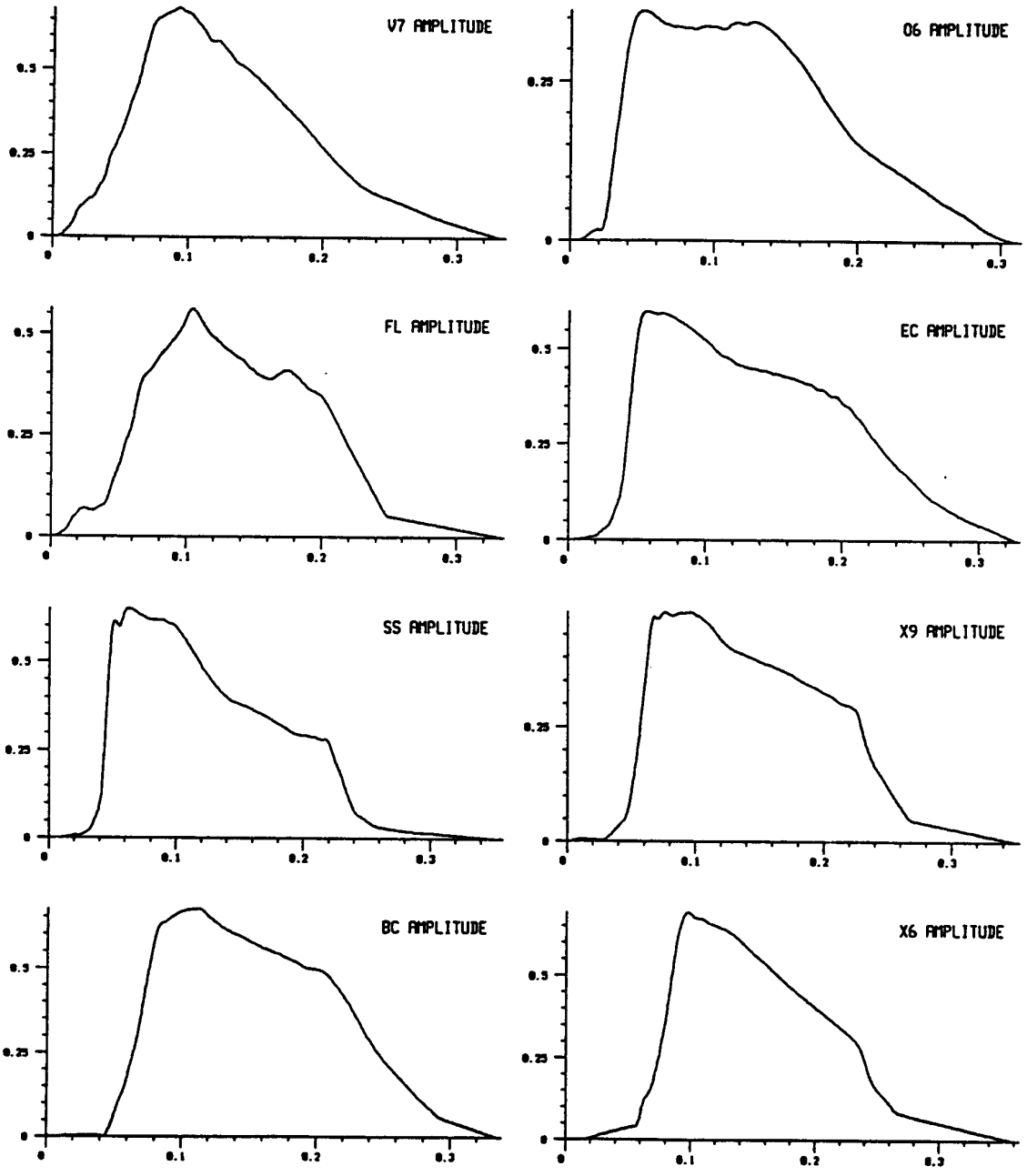

Figure 3.4 *Part 2.* The remaining 8 out of 16 stimulus envelopes are shown above.

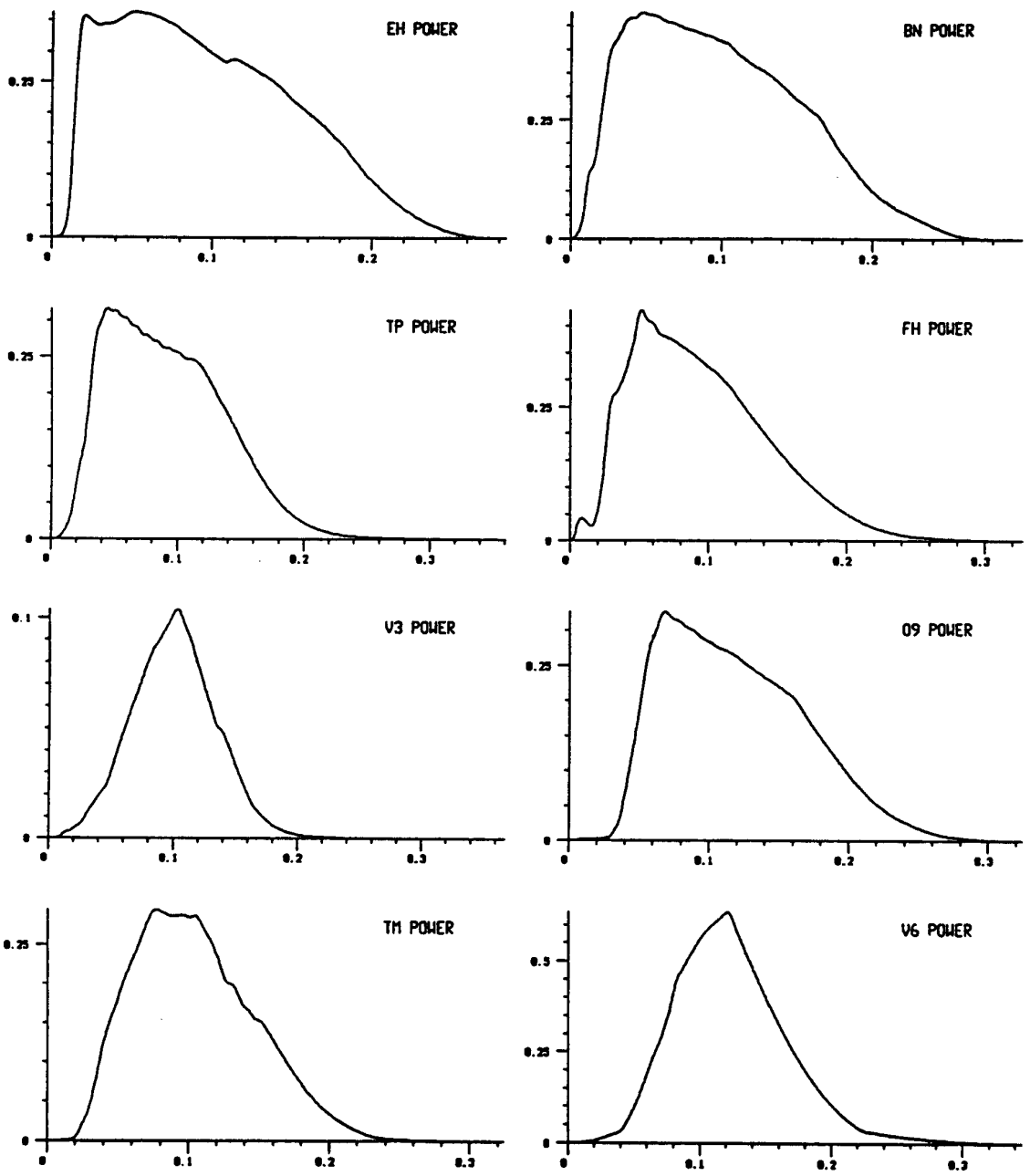

Figure 3.5 *Part 1.* Power (intensity) envelopes for the 16 stimulus tones. A stimulus's power envelope was obtained by squaring each value in the digital sequence corresponding to that tone's amplitude envelope. Units are normalized amplitude squared *vs.* seconds. 8 of the envelopes are shown above.

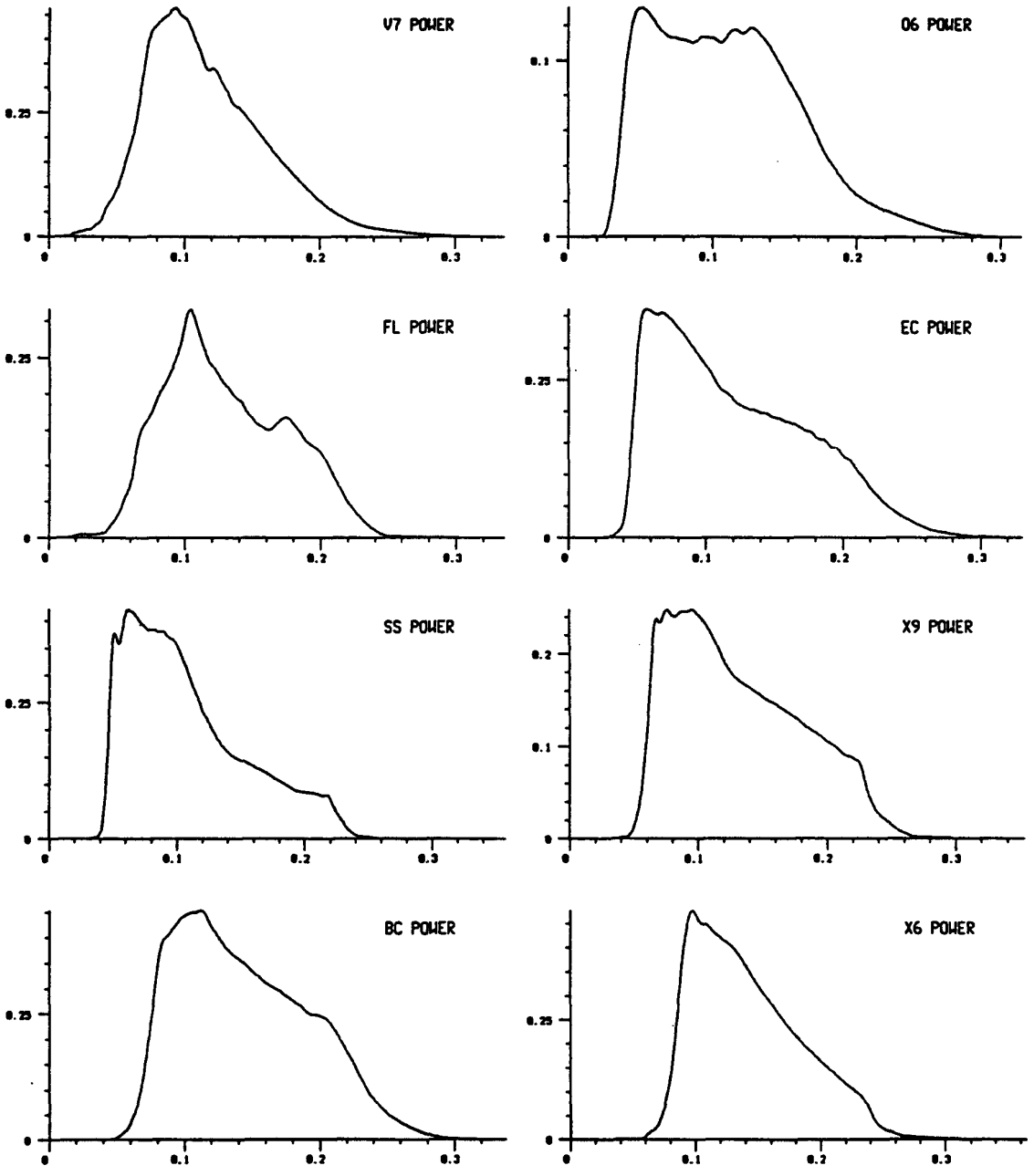

Figure 3.5 *Part 2.* **Shown above are the remaining 8 out of 16 power envelopes.**

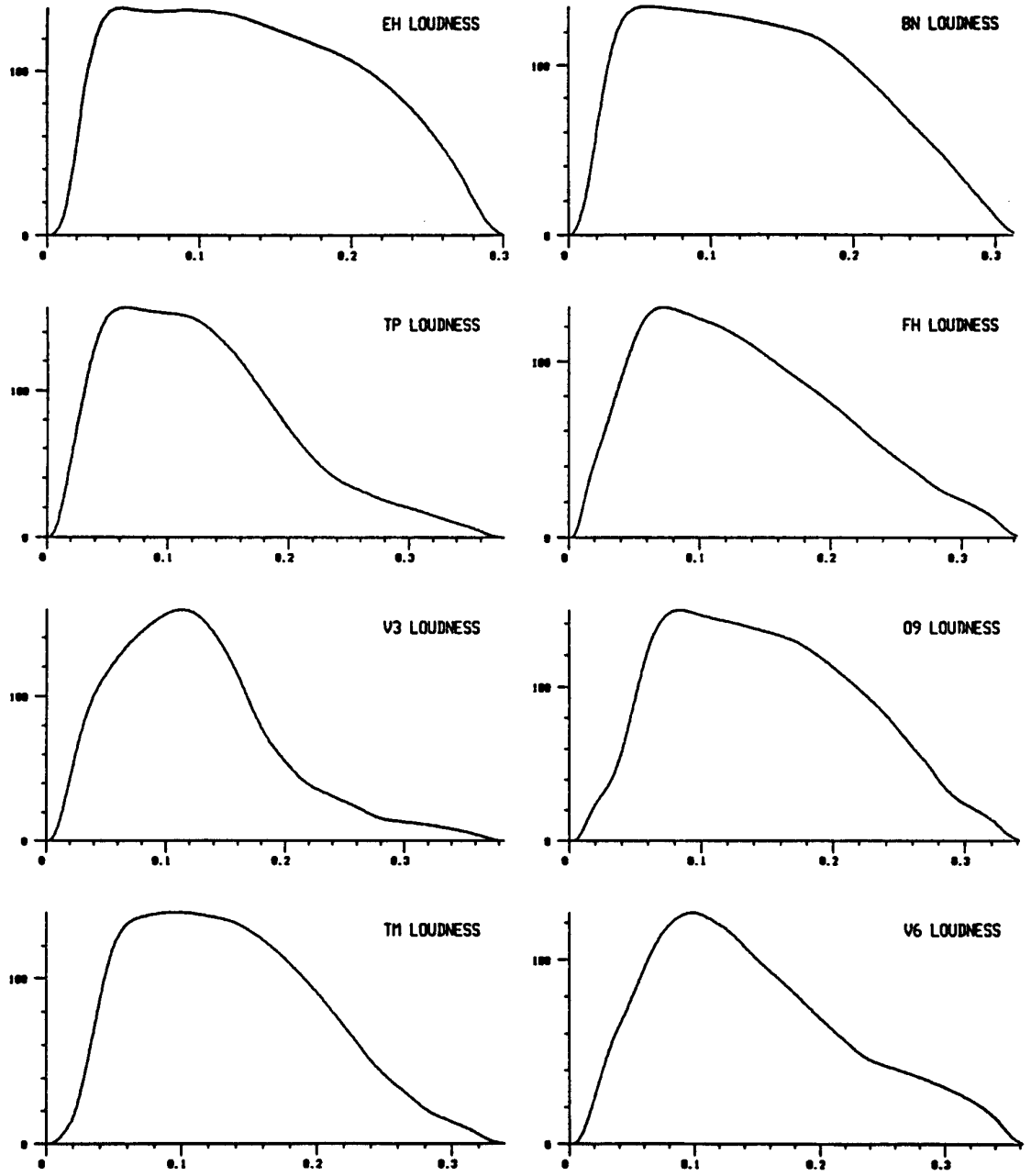

Figure 3.6 *Part 1.* Time-varying loudness envelopes for the 16 stimulus tones, obtained by applying the "Zwicker transform" to time-varying spectral representations for the tones. 8 of the envelopes are shown above. The units for the abscissa are *seconds*; those for the ordinate are somewhat arbitrary, but correspond approximately to *sones*.

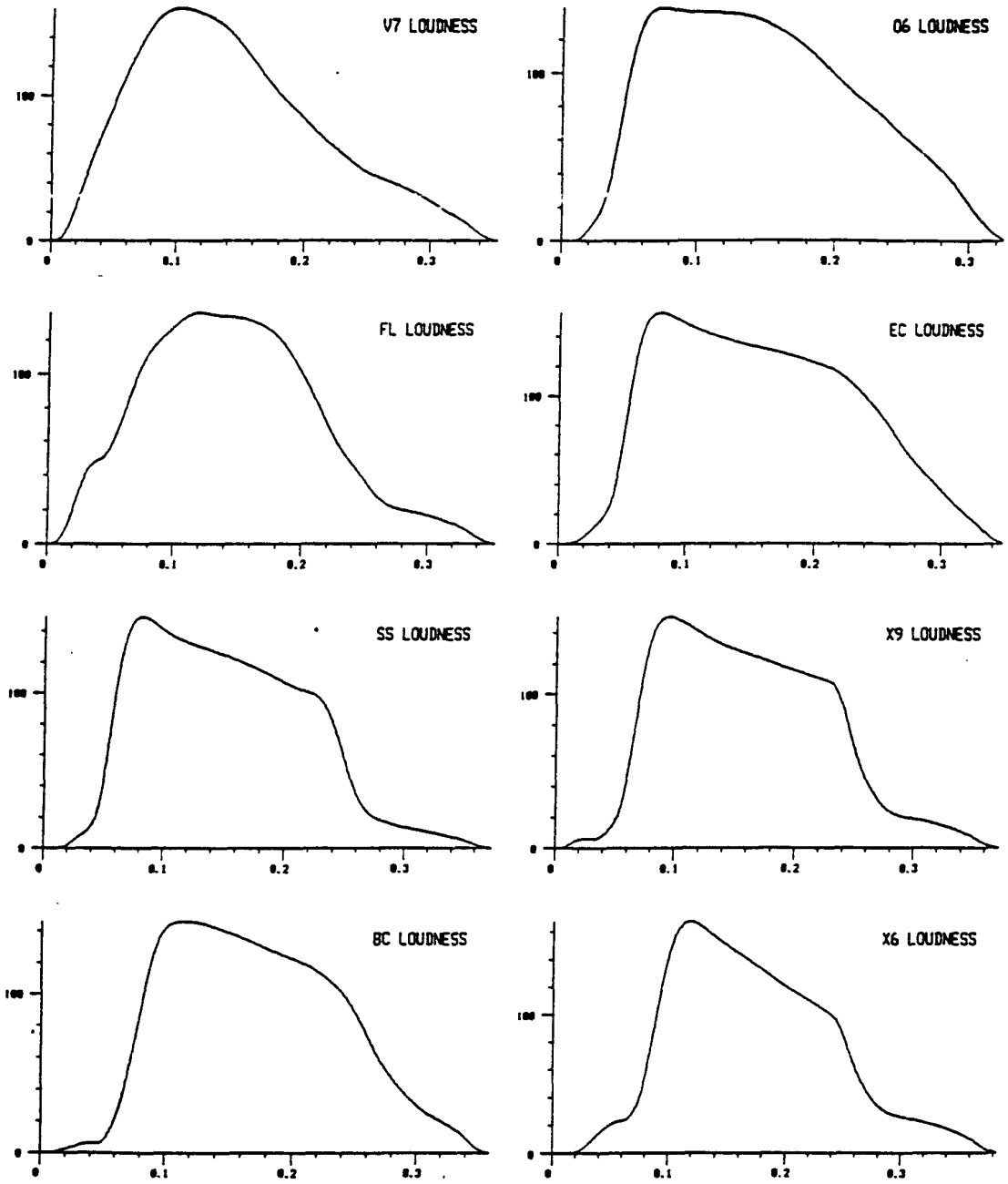

Figure 3.6 *Part 2.* The remaining 8 out of 16 loudness envelopes are shown above.

spectral representation with relatively fine time resolution (1–5 ms).

As indicated in Section 3.1, the 16 stimulus tones were resynthesized from such a time-varying spectral representation; hence, it was straightforward to realize the Zwicker-model conversion of these tones into time-varying loudness envelopes. These can be seen in Figure 3.6. Note that the maximum values for these functions (as compared to the amplitude or power envelopes) are all approximately equal. If we assume that the maximum of a time-varying loudness function is used by subjects as the measure of overall loudness (which may or may not be a valid assumption), then the near equality of maxima in Figure 3.6 indicates that the accuracy of the equalization process for loudness was quite good. (Alternatively, it is a check on the validity of the Zwicker model.)

We thus have three different sets of representations for our 16 instrument tones: the amplitude, power, and Zwicker envelopes. These were all used in testing the validity of the various models that will be discussed in the next subsection.

Prediction Models

Vos and Rasch hypothesized that PAT was based on a relative threshold model. Because of this, and because their rise functions were identical in shape (and differed only in rise time), it was straightforward to calculate PAT from their Δt values. Even if we were to accept the relative threshold hypothesis as being correct, the same formulas could not be applied to the RPAT (Δt) values obtained in this study. In our case, a trial-and-error approach seems the most practical way to explore the worth of various prediction models. It also seems useful to examine a more exhaustive set of models than just the relative threshold one.

The best model will fulfill three aims. The first aim, of course, is to have the model be a good predictor of PAT. But a second aim is to have the model be consistent with auditory theory. And a third aim is to be able to apply the model simply and practically to all types of sounds. These 3 criteria will help us determine a model's estimated validity.

One model that might be worth considering is one in which PAT is based on the time maximum amplitude is reached. It seems unlikely that the ear would wait until maximum were reached before

registering a perception of attack; on the other hand, if rise time were very short, and the maximum were a clearly defined point with decay beginning immediately thereafter (as in the case of sounds produced by many percussion instruments), perception of attack may very well coincide with the time of maximum amplitude.

A second approach would be to challenge the hypothesis that perceptual *attack* time and perceptual *onset* time are ever different—to aver rather that perception of attack is always coincident with perception of onset. In this case, the PAT prediction model becomes an onset prediction model; and we might expect perception of onset to occur when some absolute threshold is surpassed. This could be an amplitude threshold, or an energy (integration) threshold.

Another model to examine is one in which PAT is based on a relative threshold, as proposed by Vos and Rasch. In this model, threshold is expressed as so many dB below maximum, or as a percentage (ratio) of maximum. Vos and Rasch found a threshold 15 dB below maximum to best predict their Δt values, modifying the level slightly as overall intensity level changed. At first, it may sound as if this hypothesis implies that the ear "knows" what the maximum value is going to be before it is reached. What it really implies, however, is that the ear becomes acclimated to a certain listening level (determined by the maximum amplitude of the sounds it is listening to), and that its threshold for perception of attack shifts accordingly.

Finally, it seems likely that PAT would correlate well with the time the *slope* of the amplitude reaches a maximum. Indeed, if PAT corresponds to the firing of neural cells that are triggered by pulse-like stimuli, it seems appropriate to focus attention on that portion of the amplitude envelope that is most like a pulse, or where its most rapid increase takes place.

Models for all these hypotheses were developed, and tested against the data obtained from Experiment I. One or more of the three sets of envelope functions (amplitude, power, loudness) were inputs to the models, and most of the models allowed for a variable threshold as a parameter. (That is, no prior assumption was made regarding the quantitative value of the threshold.)

With this kind of model, for any particular setting of the parameter, the predicted PAT values can be tested to see how well they correlate with the Δt values obtained empirically. However, a

simple product-moment correlation coefficient is not a sufficient indication of the model's validity. The set of predicted values (predictions of APAT) should also be equal to the set of empirical values (measurements of RPAT) except for an additive constant. In other words, there should be a linear relationship between the predicted and empirical values, with a slope of 1.0. The correlation coefficient measure is based on the assumption that such a linear relationship exists, and will give an indication of how well the APAT-RPAT coordinates fall on a straight line; but it should also be determined how close the *slope* of the best-fitting line comes to 1.0.

Such a check can be derived from a standard linear regression analysis, which determines the best-fitting line to a set of XY-coordinates by means of a least-squares criterion (minimizing the standard deviation of the Y-values from the line). By using the same least-squares criterion with the added constraint of fixing the slope equal to 1.0, we can calculate the standard deviation of the Y-values from this restricted line, thereby obtaining a measure of our desired linear relationship. For any particular PAT model, therefore, the most preferable value for the parameter involved will be the one for which this standard-deviation measure is smallest *and* the product-moment correlation coefficient is largest.

The various PAT prediction models that were tested against the data will be discussed individually below.

- *Time of Maximum (MAX)*

The time-of-maximum-amplitude model (MAX) has no parameter; hence, correlation and linearity can be tested directly. Also, the times of maximum for the power envelopes are equivalent to those for the amplitude envelopes, and can thus be disregarded for the purposes of reviewing this model. In Figure 3.7, the empirical means are plotted against times of maximum for both the set of amplitude functions and the set of Zwicker functions. A segment embracing $\pm 2/3$ standard deviation from the mean is marked for each instrument to give a rough indication of response spread over all subjects. In addition, the best-fitting line with slope = 1.0 is drawn in each plot. As mentioned above, the ideal model would predict APAT values such that all 16 points would fall on or very near this line.

These graphs clearly show that the MAX model is a poor one for both sets of envelope functions, though the Zwicker transform seems to do slightly better than the direct amplitude measurements.

- *Absolute Threshold (ABS)*

The absolute-threshold model (ABS) predicts APAT as the time an instrument's envelope goes above a fixed value. The parameter for the model is the level of the threshold itself, referenced to the threshold of hearing. On a normalized scale for amplitude, one can use 0 as a convenient reference point, provided the sounds are played at a high enough intensity level. Since the quantization of the sounds used in Experiment I was 14 bits (84 dB dynamic range) and the sounds were played at 90 dB_A, one can safely assume that any non-zero values in the envelope functions were above hearing threshold (for the equivalent steady-state sound). If the intensity level were lower, one would expect the reference level (on the normalized scale), and hence the threshold itself, to be higher.

For this model, as for the previous one, power functions can be disregarded, since a PAT determined according to the time a power envelope crosses some absolute threshold will be equivalent to a PAT determined according to the time the corresponding amplitude function crosses the square root of that threshold.

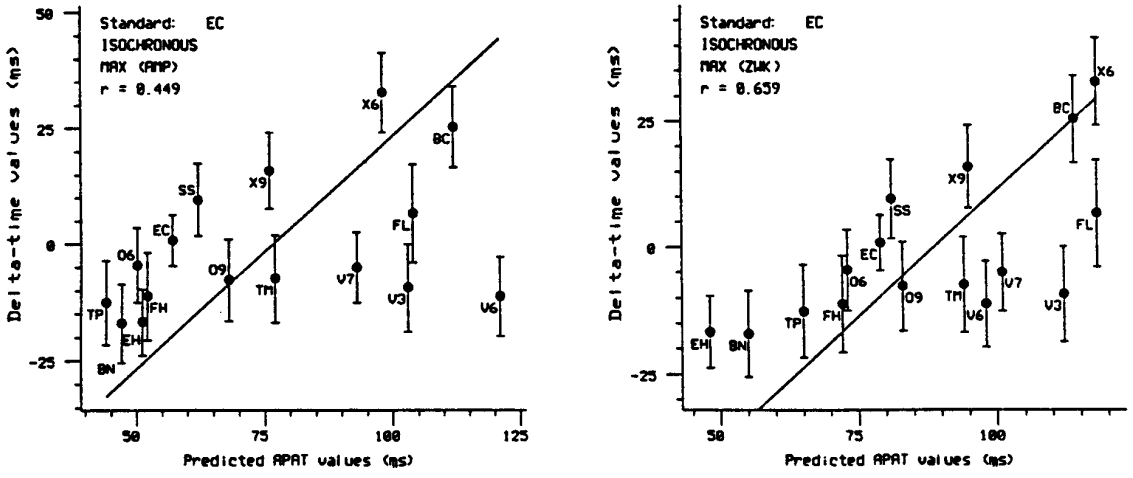

Figure 3.7 RPAT (Δt) values obtained empirically from Experiment I, correlated with predicted APAT values according to the MAX model. The left graph shows results using the set of 16 amplitude envelopes as input to the model; the right graph shows results using the set of loudness (Zwicker transform) envelopes. Product-moment correlation values are represented by r (.449 and .659).

It should be recalled that two measures were proposed for testing the validity of any PAT model; one was a correlation coefficient, which should be as close to 1.0 as possible, and the other was a standard deviation measure, which should be as small as possible. (The latter measure tests how well the APAT-RPAT coordinates fall on a line with slope equal to 1.0.) When a model includes a parameter in its formulation, it is useful to graph these measures as functions of the parameter as it varies over a range of values; these graphs then can help us determine the appropriate parameter setting that will yield optimum results.

Examples of this procedure, as applied to the ABS model for both the amplitude and Zwicker sets of envelopes, can be seen in Figure 3.8. In this figure, the independent variable for all four functions is the absolute threshold parameter, although it is an amplitude threshold parameter in the left graph and a loudness threshold parameter in the right graph. The dependent variable for the "CORRELATION" functions is simply the product-moment correlation coefficient between the set of 16 empirical Δt (RPAT) values and the corresponding set of values (APATs) predicted by the ABS model. The dependent variable for the "LINEAR FIT" functions is essentially the inverse of the standard-deviation measure mentioned above; rather than find the minimum of the standard-deviation measure, it seemed more appropriate to find the maximum of its inverse (since we're also

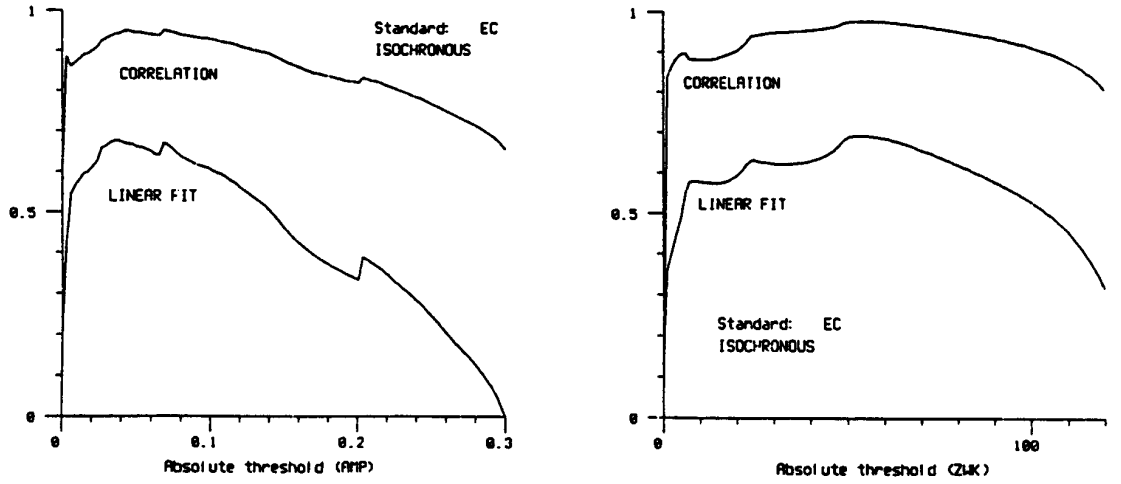

Figure 3.8 Correlation and linearity measures as absolute threshold varies over a range of amplitude (left graph) or loudness (right graph) values. Abscissa scales correspond to the ordinate scales of figures 3.4 (amplitude envelopes) and 3.6 (loudness envelopes).

finding the maximum of the correlation measure).

The left graph of Figure 3.8 shows two values of the amplitude threshold for which maximum (or near-maximum) correlation is obtained: 0.047 (∼−26 dB) and 0.0687 (∼−23 dB). (These values were found by hand with the aid of the computer.) The model's best prediction, then, will occur when the parameter is given one of these two values.

In the graph for the Zwicker functions, however, though a maximum is reached at 52, there is a wide range of threshold values for which correlation hardly changes. This is probably due to the fact that the shapes of the Zwicker functions are much more homogeneous in their attack than the amplitude envelopes, causing a wide range of parameter values to yield a set of predicted APAT values that are essentially equal except for an additive constant.

We can now plot the empirically obtained values for RPAT against the APATs as predicted by this model, with the parameters set at the specific values mentioned above. Such plots for the two amplitude thresholds (0.047 and 0.0687) are displayed in Figure 3.9; the plot for the loudness threshold is shown in Figure 3.10.

We can see that when the threshold parameter for amplitude envelopes is set to .047, the fit

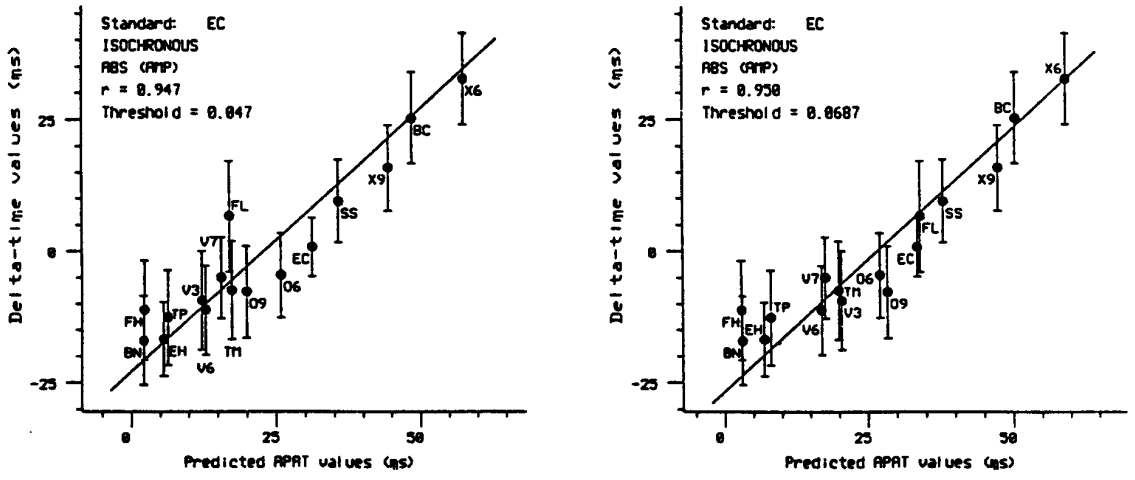

Figure 3.9 RPAT (Δt) values correlated with predicted APAT values according to the ABS model. Amplitude envelopes were used as input sound representations to the model. The two graphs show correlations obtained by setting the theshold parameter at two different levels (0.047 and 0.687); product-moment coefficients are approximately equal for the two cases (.947 and .950).

is quite good except for the flute, whose APAT is predicted too early. The flute's envelope has a "flat" spot in this region (see Figure 3.4, *Part 2*), making its predicted PAT quite sensitive to slight threshold shifts. When the threshold parameter is raised to .0687, the flute's predicted PAT comes much closer to the best-fitting line; but the PATs for other instruments (namely, O9, V3, V6) move farther away from this line. Choosing between these two thresholds is therefore problematic.

Let us now assess the model's worth in terms of the three criteria we have established. First, the model is certainly practical to use; hence one criterion is met. Second, the thresholds that best predict PAT are about 20 dB below maximum (though the maxima range between ~.32 and ~.8 for the 16 instruments). This threshold level is significantly above what one would expect the hearing threshold to be for the listening level used in Experiment I; however, it is difficult to assess the influence temporal integration has in determining the time of perceptual onset. The model's

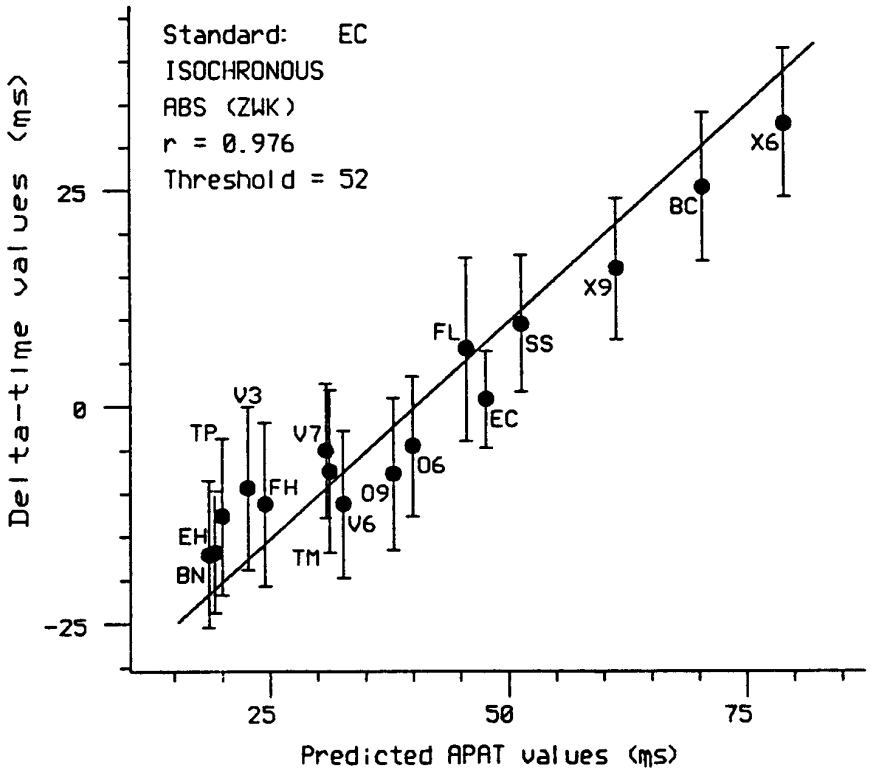

Figure 3.10 RPAT-APAT correlation according to the ABS model, using the set of loudness functions as input. The threshold parameter was set to 52, and the product-moment correlation (r) is 0.976.

consistency with auditory theory is therefore a moot point, if we make the assumption that PAT coincides with perception of onset (using these terms as they are defined in Chapter I).

Our third criterion deals with accuracy. The correlations for both threshold settings are relatively quite high; but the graphs need to be examined in more detail. The distance of a point from the best-fitting line (as plotted in Figure 3.9), can be used as a measure of the model's inaccuracy. For a few cases, this distance is 10 ms or more, which is audible in judgments of simultaneity (Patterson and Green cite 2 ms as the limit [*Patterson and Green* (1970)]), and probably also in judgments of isochronism if the beat period is less than 300–400 ms (assuming a Weber ratio of 2–4%). Even if we were to postulate that measurement errors in determining RPAT were such that the "true" Δt values weren't represented by the mean, but rather fell closer to the best-fitting line, it is highly unlikely that such errors would be great enough to account for all of the inaccuracies. Also, the "early" instruments (EH, BN, TP, FH) tend to be predicted *too* early by the model, while those falling in the middle range tend to be predicted slightly late. This suggests an exponential, or at least non-linear, relationship between the predicted and empirical PATs, which must be considered inappropriate unless some lawful reason for the non-linearity appears.

The set of Zwicker envelopes seems to do much better (see Figure 3.10). Correlation with predicted APATs is even higher than for the set of amplitude functions ($r = .976$), and distances of points from the best-fitting line are smaller. However, we still see the tendency to predict the quick instruments too early and the average ones too late. Also, the threshold is about half of maximum on the loudness scale, which again implies that the model is predicting APAT to occur after perception of onset. In terms of practicality, since the Zwicker transform is rather complicated and computationally expensive to implement, it would be preferable to find a model based on direct amplitude measurements whose accuracy is at least as good as that obtained here.

- *Relative Threshold (PCT)*

Vos and Rasch found that their Δt values were predicted quite well by the relative threshold model they developed, which set PAT to the time the amplitude envelopes crossed a threshold about 15 dB below (or $\sim$18% of) maximum. However, rather than fix the threshold at this value in testing

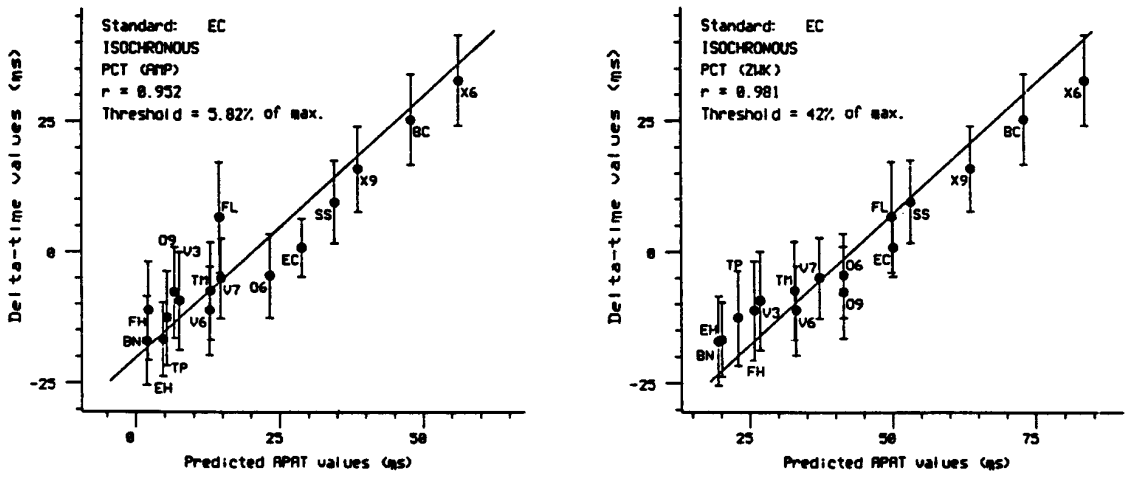

Figure 3.11 RPAT values correlated with APAT values predicted by the PCT model. Examples are given for both amplitude (left graph) and loudness (right graph) envelopes.

the model against our data, we should allow the threshold to vary as a parameter, as has been done in the ABS case. Also, the dB measure is the same regardless of whether amplitude or power is used: $\text{dB} = 20\log_{10}(A/M) = 10\log_{10}(A^2/M^2)$, where A represents a value on the amplitude envelope function, A^2 represents the corresponding value on the power envelope function, and M is maximum amplitude (and M^2 is maximum power). Therefore, the power envelopes can again be disregarded in our testing procedure, since a relative threshold applied to them can be converted into some equivalent threshold for amplitude envelopes.

As in the case of the ABS model, correlation and linearity plots were obtained for both the amplitude and Zwicker functions, the parameter ranging from 1 to 99%. (Note that 100% would correspond exactly to the MAX model.) Since the shapes of the correlation and linearity curves are very similar to the ones obtained for the ABS model (a single peak for the amplitude functions and a "flat" region for the Zwicker envelopes—see Figure 3.8), these graphs will not be reproduced here. The peak for the amplitude functions occurred for a parameter setting of 5.82%; the best setting for the Zwicker envelopes was ~42%. The RPAT-APAT plots for these parameter settings are shown in Figure 3.11.

The plot for amplitude envelopes is quite similar to the corresponding ABS-model plot (with

the parameter set to 0.047); overall inaccuracies and trends are about the same. In the Zwicker-envelope case, however, the fit is even better than for the ABS model, though the skew for quick and average instruments is still present. In fact, though the relationship between RPAT and APAT values appears to be essentially linear, the slope of this line is somewhat less than 1.0. Therefore, even though the relative threshold model (PCT), as applied to the set of Zwicker functions (and with the threshold set to 42% of maximum), is the best model we have yet tested, it is nevertheless less than ideal.

For comparison purposes, the threshold value proposed by Vos and Rasch was used to predict PAT from the set of amplitude envelopes. The result is plotted in Figure 3.12. Many of the instrument means fall very close to the best-fitting line, and thus the model is quite good for this

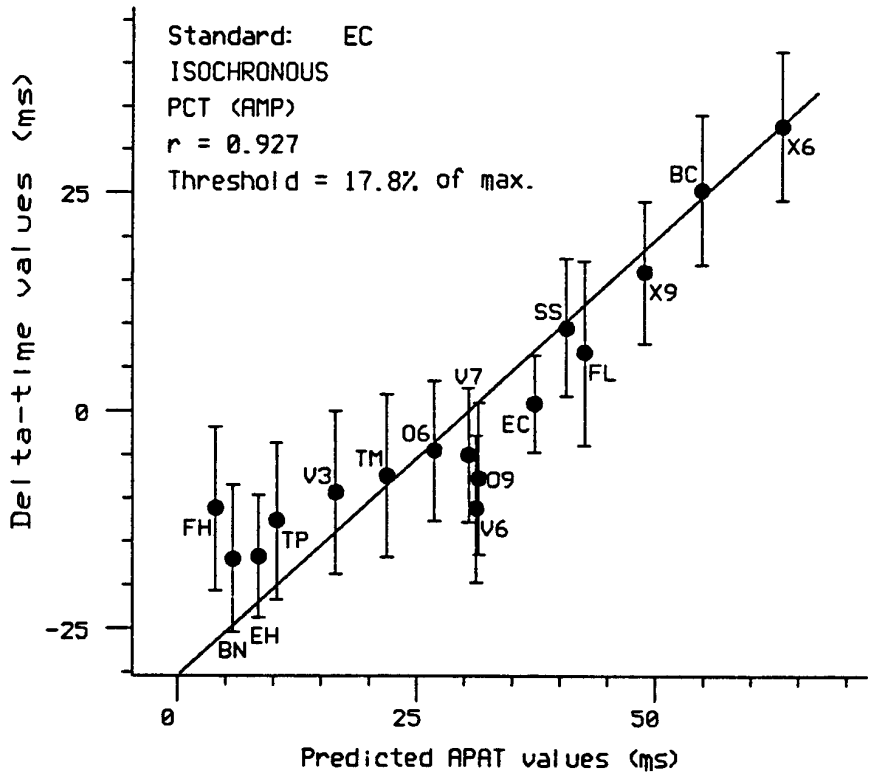

Figure 3.12 Empirical RPAT (Δt) values correlated with APAT values predicted according to the Vos-Rasch relative threshold model (θ = 17.8% of maximum amplitude).

subset of instruments. But the fact that some points stray quite far from the line shows that the model cannot be universally applied to all instruments. It also illustrates the usefulness of using stimuli obtained from real instrument tones, rather than from synthetic ones.

For the sake of completeness, a different kind of relative threshold model was also tested. It measured the relationship between threshold and maximum as a difference in amplitude or loudness, rather than as a ratio. This model, however, did more poorly than the ABS and PCT models, and was abandoned without further consideration.

- *Integration Threshold (ENE)*

Since temporal integration influences auditory detection at, and loudness above, hearing threshold levels [*Zwislocki* (1960)], it is appealing to examine its possible influence in determining PAT. A simple approximation to temporal summation (ignoring critical band separation, adaptation, and other effects), is to accumulate the values of a single envelope function at each sample (or ms in our case). PAT can then be determined by the time this cumulative sum crosses some threshold. This model will be referred to as the integration threshold model (ENE), and the threshold value will be allowed to vary as a parameter. ("ENE" refers to energy, which is the time-integral of power.)

For this model, amplitude and power functions will not yield equivalent results; hence, both

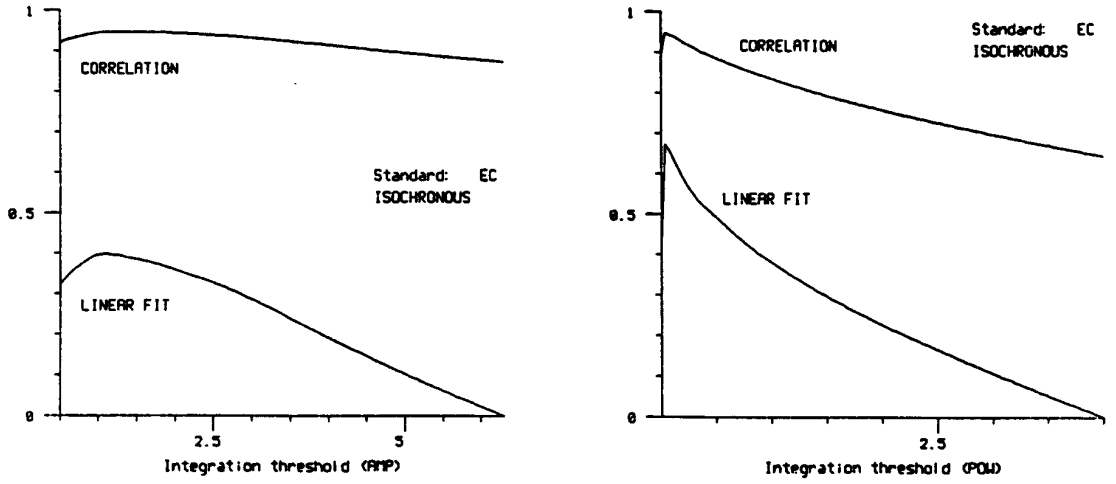

Figure 3.13 RPAT-APAT correlation and linearity measures for both amplitude (left graph) and power (right graph) envelopes as absolute integration threshold varies over a range of values. (APAT values were predicted according to the ENE model.)

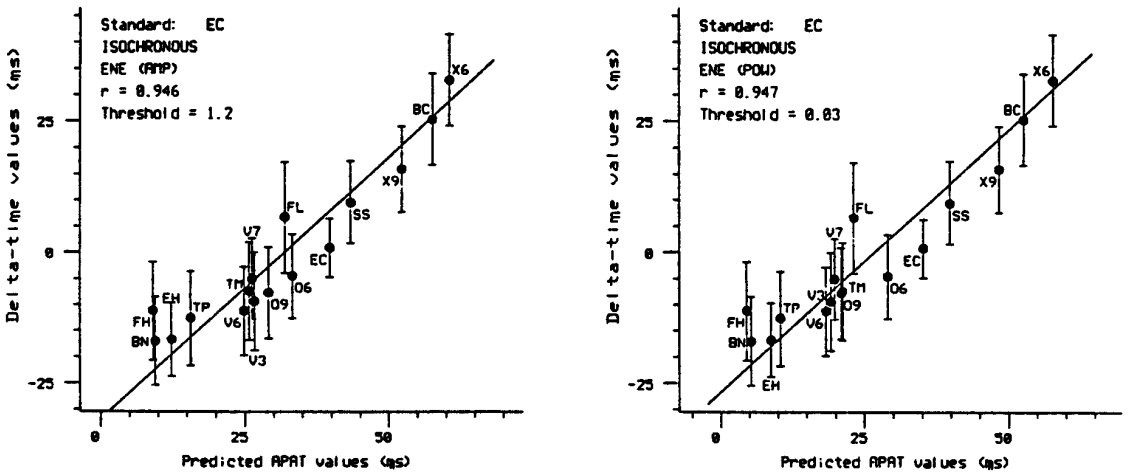

Figure 3.14 RPAT values correlated with APAT values predicted by the ENE model. Examples are given for both amplitude (left graph) and power (right graph) envelopes.

should be tested. However, since temporal integration is included in the time-varying Zwicker transform (actually, both spectral and temporal integration occurs, though at different levels), it seems inappropriate to test the ENE model using the set of Zwicker envelopes.

Figure 3.13 shows the correlation and linearity curves for both sets of envelopes as the threshold parameter varies over appropriate values. Correlation for the amplitude set is virtually flat over the entire range; in this case, the linearity curve was used to determine the "best" threshold value. For the power functions, the correct threshold to use was more obvious.

The corresponding RPAT-APAT plots are shown in Figure 3.14. The two plots are very similar, and the same kinds of inaccuracies that were found for the ABS and PCT models are present in these graphs as well. The inaccuracies are large enough in some cases to warrant concluding that the ENE model is not a valid one in predicting PAT for this set of data.

- *Slope Threshold (ENE)*

The final model that will be considered in this chapter is one based on the slope, or derivative, of the various envelope functions. It was hypothesized that the time the slope reached a maximum would correlate well with the set of RPAT values, and would thus be a good predictor of APAT. There is no theoretical reason for not including all three sets of envelopes, and since the slope of x^2

($2xdx$) is different from the slope of x (dx), the sets of power and amplitude envelopes should yield different results.

To obtain time-varying slope functions from the envelope curves, a simple first-order difference equation ($y(n) = x(n) - x(n-1)$) was used to approximate the derivative (and this approximation preserved sampling rate). In examining these slope functions, it was noticed that some of them exhibited more than one maximum, with these maxima equal or near-equal in value. Also, small irregularities in the source (envelope) functions could cause large excursions in the derivative (slope) functions, since these irregularities tended to happen in the space of one or two samples. Furthermore, a preliminary check (by eye) of all 48 slope functions (3 sets of 16) showed clearly that the proposed hypothesis for predicting APAT was not a valid one. It was therefore abandoned at that point.

It still seemed appropriate to derive a prediction model for APAT based in *some* way on envelope slope. But at the same time, it seemed appropriate to gather more measurements of RPAT—to check in some way the accuracy of the Δt values that were used to test the various prediction models proposed thus far.

3.5 Further Discussion

It is useful now to review our three major questions again to see how well they can be answered. It is more convenient to take them in reverse order, beginning with Q3.

- Q3: *Do individuals hear PAT differently?* We can at this point conclude neither yes nor no to this question. The ANOVA from Experiment I has shown a significant difference among subjects in their responses (their particular settings of the knob), and also in their interaction with the set of instruments. But the overall spread of subject means ($\sim$10 ms) is roughly equal to what we would expect as the limit in the discrimination of regularity (1–2% of 600 ms), and the interaction is too inconsistent to permit interpretation. So differences among subjects may well be procedural, and

cannot be generalized into specific perceptual differences.

Also, if subjects really *did* hear PAT differently, the differences would in all likelihood be exaggerated for instruments with long rise times. It is clear from Figure 3.2, however, that the spread of responses is virtually constant for all instruments (whereas the instrument rise times show wide variation—see Table III.2).

- Q2: *How well can PAT be predicted from a quantitative representation of the sound?* We have seen that three models (ABS, PCT, ENE) predict sets of PAT values that are virtually the same, when the predictions are based on the amplitude or power envelopes. In all three cases, predicted APAT for a few isolated instruments appears to be erroneous enough that we would expect to hear the discrepancy in certain special listening conditions. The two models (ABS, PCT) tested on the Zwicker transform envelopes are also essentially equivalent to each other, and inaccuracies are much smaller than for the amplitude (power) envelopes.

It seems that we have sufficient evidence, therefore, for stating that PAT can be predicted reasonably well, especially if the chosen sound representation is the Zwicker loudness model. However, there is as yet insufficient evidence for choosing one model over another; in fact, all of the models are less than satisfactory, since—despite the high correlations between RPAT and APAT,—none achieves the precise linear relationship between RPAT and APAT that we desire (illustrated by quick instruments being predicted too early and average ones being predicted too late).

To correct this deviation, it may be necessary to modify or extend our models to account for different weightings subjects may give to various PAT cues. However, the deviation from strict linearity does not necessarily imply that the models are at fault; it could well mean that the measurements of Δt are in error such that the instrument means don't correctly represent RPAT. This seems even more likely when we consider that the sets of predicted APAT values for all models display approximately the same relationship with the set of RPAT values. On the other hand, it is also possible that our sound representations are inadequate; for instance, there may be rapidly changing spectral aspects of the sounds that influence PAT as much as (or even more than) amplitude characteristics. But before making the prediction models more sophisticated, there seems to be a

need to examine the measurements from Experiment I more closely. This brings us to a discussion of our last major question:

- Q1: *How accurately can PAT be measured?* As mentioned in section 3.3, measurement accuracy depends on auditory temporal acuity, which in turn depends on the nature of the appointed task. In Experiment I, we found discriminability of regularity to be the main limiting factor. A different experimental design can be chosen to reduce or even eliminate this factor, and running such an experiment might yield more accurate measurements of Δt. This approach was indeed taken, and will be discussed in detail in Chapter IV. But a topic still remains that needs to be covered here, and that is whether or not the Δt values obtained in Experiment I are best represented by the instrument means.

In section 3.2 it was reported that an outlier more than 3 standard deviations away from the mean was replaced by the mean of that subject's remaining responses. However, this still left a spread of responses that was quite wide, and, in general, there was no reason to assume that the response spread could be approximated by a normal or even symmetric distribution. This implies that the median or mode may not be the same as the mean, for instance, or that there could even be more than one mode.

To get a better idea of how the responses were distributed, it was desired to obtain something akin to a histogram of responses. However, a simple discrete histogram would have been an inadequate approximation to the frequency distribution, since the actual number of response values was so great that there was no straightforward way to group values into discrete classes. Rather than take this approach, a method based on probability distributions and densities was chosen. A digital sequence was formed that approximated the ogive for each set of responses (16 curves in all, one for each instrument). This curve was smoothed by passing it through a low-pass, zero-phase, digital filter. A probability density curve was then derived by differentiating the smoothed ogive; the derivative was approximated by a simple first-order difference equation.

The density graphs for six of the instruments are shown in Figure 3.15, and are representative of all sixteen instruments. The responses for EC (upper left) approximate a normal distribution

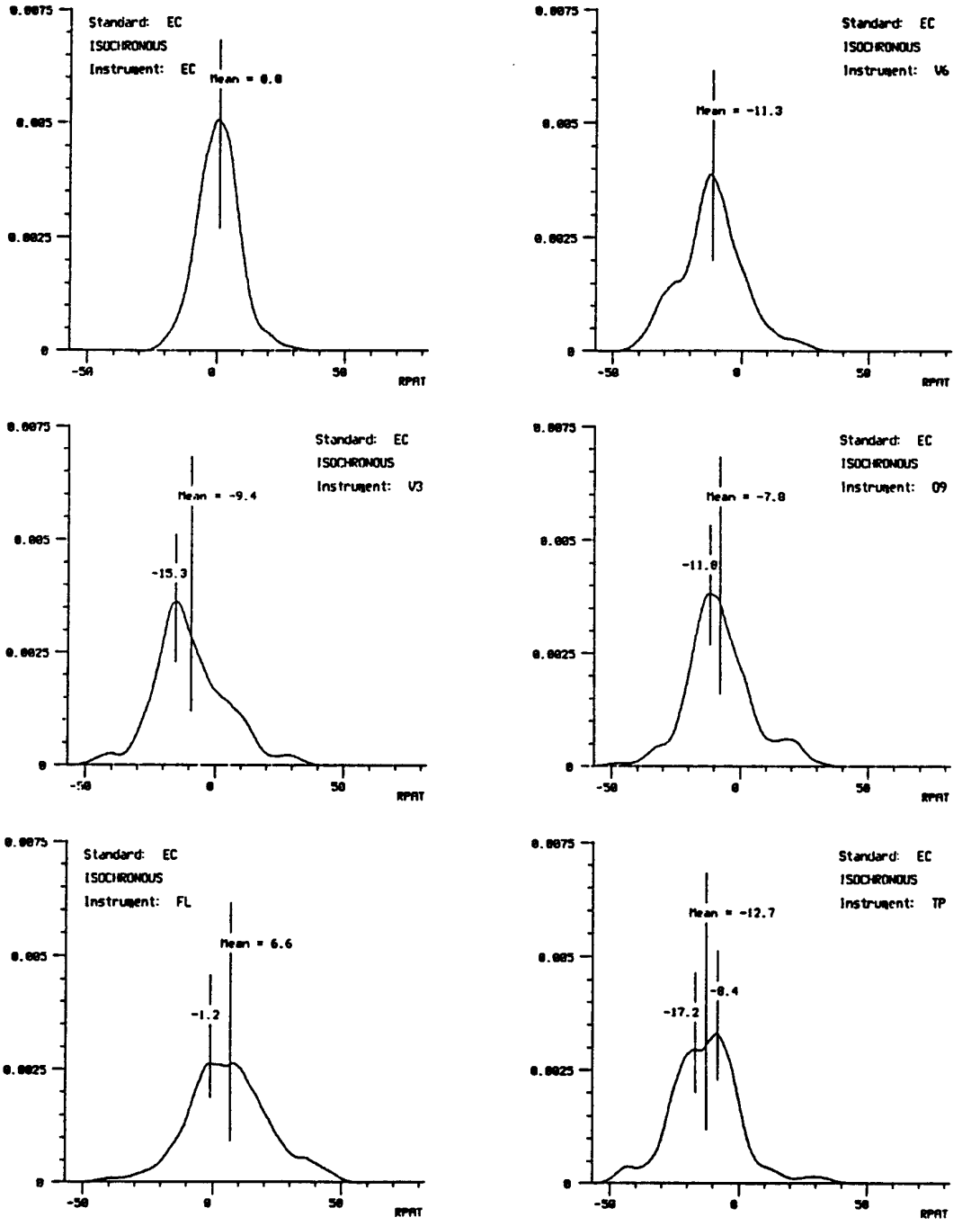

Figure 3.15 Probability density curves for 6 of the 16 stimulus tones used in Experiment I. The curves were obtained by differentiating ogives representing response distributions for all subjects. Abscissa coordinates represent empirical RPAT (Δt) values. The ordinate scale is such that the complete integral of any density curve equals 1.0. Vertical lines marking the mean response are shown for each instrument; other vertical lines mark certain modes (see text).

quite well, and the mean is clearly the best representative for RPAT. The curve for instrument V6 (upper right) is not quite as symmetric as the curve for EC, and is also more spread out; nonetheless, the mean seems to best represent Δt in this case as well. Four other instruments (EH, TM, V7, and SS) had density graphs similar to V6 and EC. The middle two graphs of Figure 3.15 (instruments V3 and O9) show more asymmetry and the influence that outlying values have on the mean. In both cases there is a mode a few milliseconds less than the mean that would probably be a better choice for representing RPAT. (These modes are indicated in the figure by the lines that intersect the peaks of the graphs, accompanied by their X-axis values.) There are six other instruments (BN, FH, O6, X9, BC, and X6) whose graphs display this behavior, all of the modes being less than the mean.

The densities for the remaining two instruments (FL and TP) are special cases, and are shown in the lower two graphs of Figure 3.15. The curve for the flute exhibits a "plateau," which suggests that the flute's RPAT may be better represented by a range of values instead of just a single Δt. 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.

The trumpet tone's density graph is special in that it is almost bimodal, with the modes falling on either side of the mean (taking on the values of -17.2 and -8.4, as indicated in Figure 3.15). The mean is still probably the best choice for RPAT in this case, though either of the two modes is also a plausible candidate. What is interesting, however, is that there is no spectral behavior in the trumpet tone (as there is in the flute tone) that could account for the bimodal response.

We can conclude, therefore, that the set of Δt values that were used to represent RPAT may need to be altered slightly before using them in testing the accuracy of our various prediction models.

Chapter IV

Empirical Research: Part II

At the beginning of Chapter III, it was mentioned that judgments of either isochronism or synchrony could be used to obtain measurements of RPAT. In Experiment I, isochronous judgments were chosen based primarily on software considerations. At the end of Chapter III, the accuracy of these Δt measurements was placed in question, and it was suggested that the set of RPAT values might need to be altered somewhat before the set could be used for further testing of PAT prediction models.

In this chapter, two additional experiments, Experiments II and III, are discussed; both of these involve judgments of synchrony rather than isochronism. It was hypothesized that having subjects synchronize instrument tones simultaneously, rather than isochronously, would yield more accurate RPAT measurements than those obtained in Experiment I; however, the problem of merging command streams for the synthesizer was still present (this problem was also mentioned at the beginning of Chapter III). Rather than attempt the development of a real-time merging algorithm, a different approach to playing simultaneous tones was taken. Sample-data sequences for the instrument tones were generated from the original command files by having the synthesizer write its output to disk instead of to its DAC; these sequences could be loaded into the synthesizer's delay memory quickly (much more quickly than the duration of a single trial). The synthesizer was instructed to read samples from delay memory for each sequence, and to output them to the DAC. There was enough room in delay memory to store four such sequences (two for the current trial and

two for the next trial), each sequence separated by 100 ms of "silence" (zero-valued samples). For a two-sequence trial, the starting sample of one sequence relative to the starting sample of the other sequence could be controlled in real time by the subject; thus, by adjusting this control, the two sequences could be made to sound more or less synchronous in attack.

Using this means of presentation had the desired effect of allowing real-time control of synchrony, and it was therefore adopted for both Experiment II and Experiment III. The methods and results from these two experiments will be described in the next sections, followed by an overall discussion of Q1 (measurements) and Q3 (subject consistency). Discussion of Q2 (predictability) will be expanded into its own section, in which the PAT prediction models developed in Chapter III will be reviewed and retested using a revised set of RPAT (Δt) values. Also in this section will be developed some models based on slope and/or rise time; these models will be seen to predict PAT more accurately than earlier models. This section will then end in an overall discussion of Q2.

4.1 Experiment II, Method

- *Stimuli*

The stimuli were the same set of 16 instrument tones used in Experiment I. Three of the tones were chosen as standards: EC, BN, and V6. EC was the standard used in Experiment I and exhibited average attack characteristics. BN had a very quick attack; V6 had a long, drawn-out attack.

- *Procedure*

The two stimuli, **A** and **B**, were presented together and repeated at 1-second intervals. Certain keys on the subject's teletype keyboard controlled the physical onset time of **B** relative to that of **A**; the subject was thus instructed to exercise such control until **A** and **B** were perceived to be synchronous. The subject was told to concentrate on the attack portions of the two tones and to pretend to be a conductor, trying to get the two "players" to perform exactly together on the beat.

- *Apparatus*

The listening environment was the same as that used in Experiment I. A computer program was run to control all operational procedures, including determining which pair of tones to play according to the trial number and controlling the synthesizer so that both tones were played at the same time. Typing an appropriate key on the teletype keyboard interrupted the computer program, which in turn caused the relative onset times between **A** and **B** to be updated with the desired value (different keys changed the relative onset times by different amounts). Each trial began with the relative onset times $\sim$120 ms apart; the subject was thus forced to synchronize a pair of tones that initially was obviously asynchronous.

A number was displayed on the terminal screen that was the sum of a random integer and the relative onset times between **A** and **B**. This displayed number thus changed as the subject adjusted the difference in relative onset times; but the subject could not use it to determine his answer (such as setting this number to 0), because of the random element. In other words, the subject was able to use the displayed number as an anchor in comparing various **A-B** relationships, but the anchor itself was arbitrarily determined from trial to trial. Subjects indicated satisfaction with the synchrony of the tones by typing <RETURN> on the terminal keyboard; the computer then proceeded to the next trial. The subject had the option to take a rest at any point and resume the experiment later.

- *Design*

A trial consisted of one of the 16 tones paired with one of the three standards, EC, BN, or V6. The standards differed among each other in their attack characteristics, especially rise time. The reason more than one standard was used was to introduce more variety into the judgment task and to test how the set of RPAT values depended on the rise time of the standard. Each possible pair was replicated a total of 6 times, including standard-standard pairs (such as EC-BN). In actuality, each standard was paired with a set of 15 instrument tones (all 16 tones minus one of the other two standards), the order of trials within a set being determined randomly. Thus, the total number of trials was $3 \times 15 \times 6$, or 270 trials. Standards were rotated every two complete sets, or every 30 trials.

Table IV.1 3-way Analysis of variance for Experiment II.

Factor	Mean Square	df	F	p
Instrument	37,595	15	641.757	< 0.001
Standard	76,229	2	1301.264	< 0.001
Subject	198.8	7	3.393	< 0.005
Inst × Stan	855	30	14.595	< 0.001
Inst × Subj	301.8	105	5.152	< 0.001
Stan × Subj	619	14	10.567	< 0.001
Inst × Stan × Subj	260	210	4.438	< 0.001
Error	58.581	1920		

The subject had control over the relative onset time between **A** and **B** to within a sample period ($\sim 40\mu s$), if desired. In practice, the subject chose no finer increment than 16 samples, or $\sim$0.6 ms; this was still much finer control than what was available in Experiment I. The measured RPAT for each trial was simply the sample delay (converted into milliseconds) between the physical onset of **A** and the physical onset of **B**.

- *Subjects*

Eight subjects performed the experiment, all of whom were experienced in computer music and considered to have well-trained ears. Six of the eight subjects participated in Experiment I.

4.2 Experiment II, Results and Discussion

The data from Experiment II was subjected to a 3-way ANOVA, the factors being Instruments × Standards × Subjects. Results are shown in Table IV.1. Again, the factor due to instruments was highly significant (F-ratio of 642, df = 15,1920), indicating that subjects could clearly distinguish among the instrument tones' attack times.

Each instrument's mean and standard deviation is plotted in Figure 4.1. Means for standard BN have been increased by 10 ms to avoid considerable overlap with means for standard EC. For a

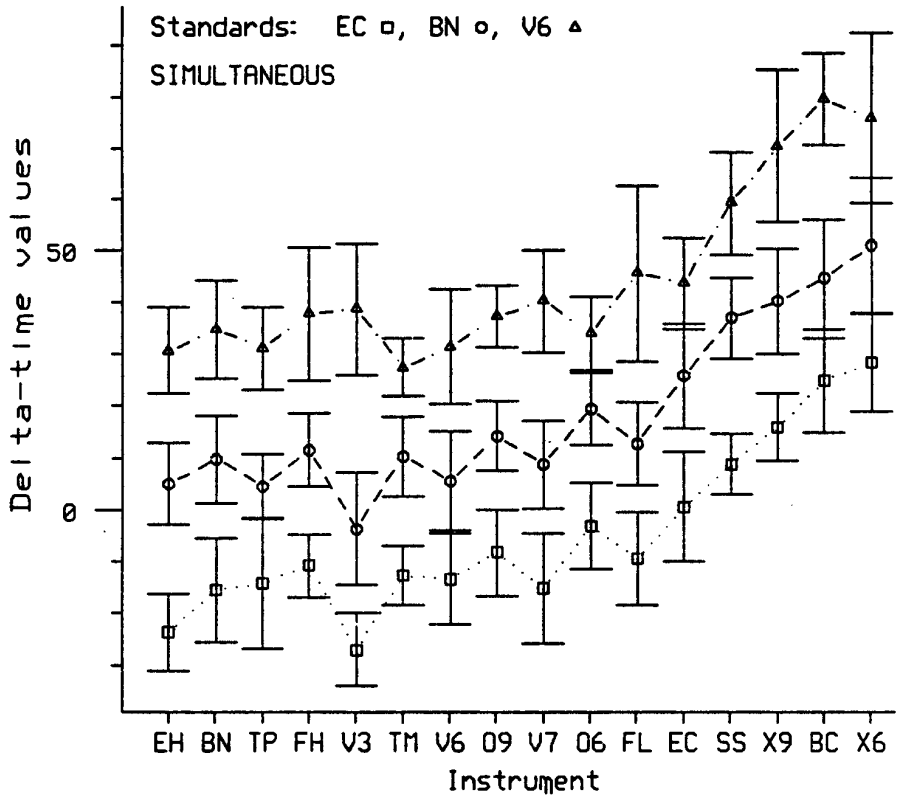

Figure 4.1 Mean Δt for 16 instrument tones, representing RPAT relative to three different standard tones (EC, BN, V6). Means relative to the BN standard are all offset by a constant value of 10 ms; means relative to V6 are all offset by a constant of 30 ms. Vertical lines extend 1 standard deviation on either side of the mean.

similar reason, means for standard V6 have been increased by 30 ms. The actual values are listed in Table IV.2. Standard deviations are roughly constant across instruments and standards, and, at least for standards EC and BN, they are generally smaller than those from Experiment I. Note, however, that a comparison of a standard with itself (EC with EC, BN with BN, or V6 with V6) yields about the same SD as other comparisons; it should be recalled that a comparison of EC with itself in Experiment I resulted in a SD noticeably smaller than all other comparison standard deviations.

The F-ratio for standards (1301, df = 2,1920) is highly significant, but this is probably due

Table IV.2 Mean Δt and standard deviation for the 16 tones in Experiment II, as measured against three different standards: EC, BN, V6.

Instrument	EC standard		BN standard		V6 standard	
	Mean	(SD)	Mean	(SD)	Mean	(SD)
EH	−23.8	(7.4)	−5.1	(7.8)	0.4	(8.3)
BN	−15.6	(10.1)	−0.4	(8.4)	4.6	(9.5)
TP	−14.4	(12.5)	−5.6	(6.1)	0.9	(7.9)
FH	−10.9	(6.1)	1.4	(7.0)	7.6	(13.0)
V3	−27.2	(7.0)	−13.8	(10.8)	8.5	(12.7)
TM	−12.9	(5.7)	0.0	(7.6)	−2.8	(5.6)
V6	−13.5	(8.8)	−4.6	(9.5)	1.2	(11.1)
O9	−8.4	(8.4)	4.0	(6.8)	7.1	(5.9)
V7	−15.4	(10.6)	−1.5	(8.4)	10.0	(10.0)
O6	−3.3	(8.3)	9.2	(7.0)	3.8	(7.1)
FL	−9.7	(8.9)	2.5	(8.0)	15.4	(17.1)
EC	0.4	(10.6)	15.6	(10.1)	13.5	(8.8)
SS	8.6	(5.8)	26.7	(7.9)	29.1	(10.0)
X9	15.7	(6.5)	30.1	(10.2)	40.3	(14.9)
BC	24.7	(10.0)	34.5	(11.5)	49.5	(8.9)
X6	28.1	(9.5)	41.0	(13.3)	45.8	(16.6)

primarily to the large differences among average Δt's for each standard. In other words, the average Δt for standards BN and V6 was 15–20 ms greater than the average Δt for standard EC. This was because BN's attack is relatively quick and most of the RPATs measured against BN were positive, while EC's attack is slower and many of the RPATs measured against EC were negative. (Though V6's rise time is relatively long, its attack was nevertheless perceived as quick; hence, almost all of the RPATs measured against V6 were positive.)

The set of RPAT values for each standard, then, are separated from each other by a relatively large constant, and this separation accounts for the significance of the F-ratio for standards. Reasons for the significant instrument-standard interaction term can be seen from Figure 4.1. The connecting lines between instruments for each standard show that the overall trends for standards EC and BN are almost identical; however, standard V6 exhibits a somewhat different pattern. V6 is thus seen as a special case. In fact, if a general comparison is made between isochronous and simultaneous

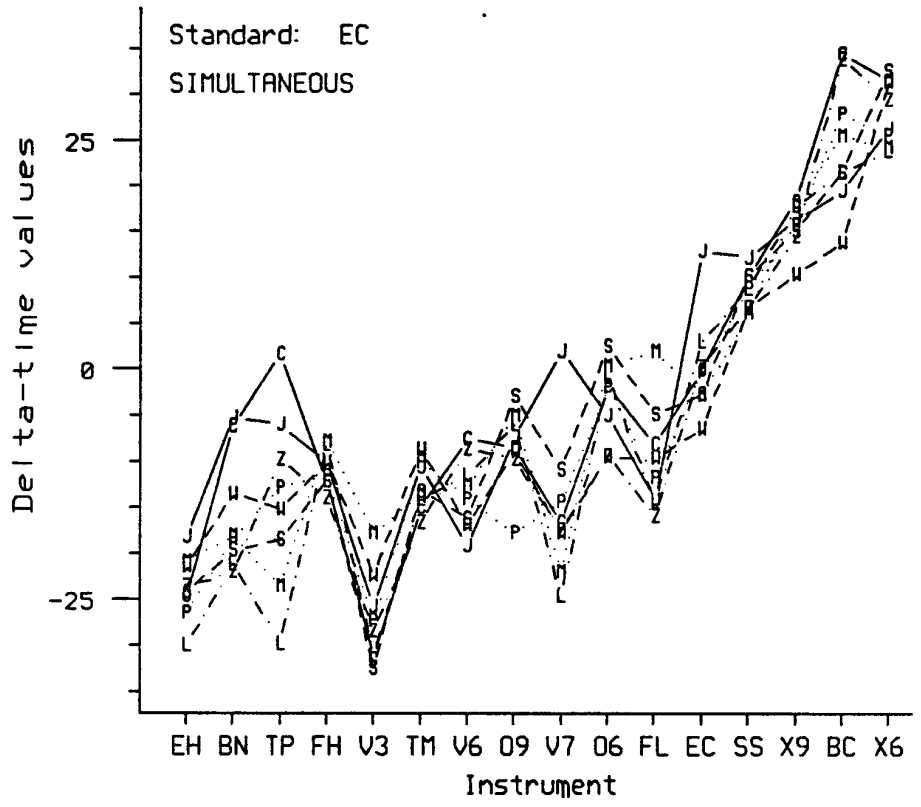

Figure 4.2 Mean Δt (RPAT) values for the 16 instrument tones used in Experiment II, as compared with standard EC. Mean values are plotted individually for each subject.

judgments (see Figure 3.1), it can be seen that subjects respond in a fundamentally different way to *all* of the string tones (V6, V7, V3). This is apparent not only from the difference in trend for standard V6 as compared to the other two standards in Experiment II, but also from the two different trends for standard EC (isochronous and synchronous—Figure 4.1 *vs.* Figure 3.1).

Subject differences are best assessed by examining Figures 4.2, 4.3, and 4.4. The first two of these graphs show small subject differences; on the other hand, Figure 4.4 illustrates several cases in which there was significant subject disagreement. Thus we see that subject responses to the string tones (or at least V6 in this experiment) are different from responses to the other tones—not only in a general way, as seen from Figure 4.1, but also on an individual level.

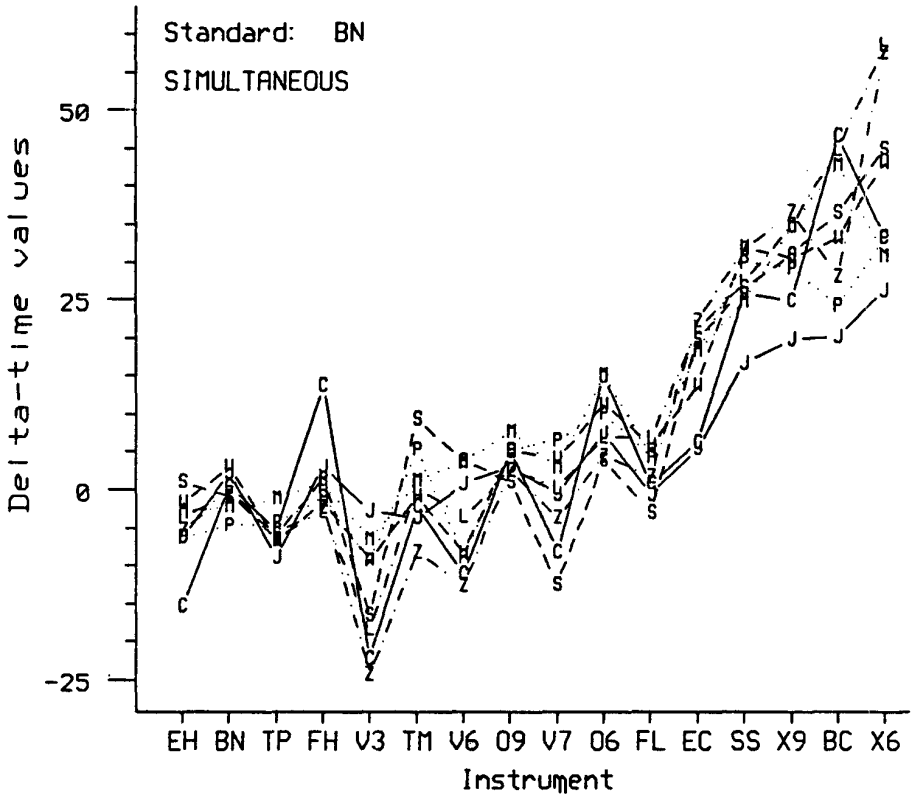

Figure 4.8 Mean Δt (RPAT) values for the 16 instrument tones used in Experiment II, as compared with standard BN. Mean values are plotted individually for each subject.

The *F*-ratio for subjects (3.4, df = 7,1920) is significant at the 99.5% level of confidence; however, the mean square for subjects (198.8) is smaller than the mean squares for all the interaction terms (855, 302, 619, 260). In other words, inherent differences among subjects aren't any more significant than interactions of subjects with standards and with instruments. This supports the inclination in Chapter III to conclude that subjects really do *not* hear PAT differently.

As noted in Chapter III, response distributions may not be normal and symmetric, and thus modal Δt values may not coincide with their respective means. It is therefore useful to examine representations for the frequency distributions of all the responses, as was done for Experiment I. The same algorithm used to obtain probability densities for the data from Experiment I was applied

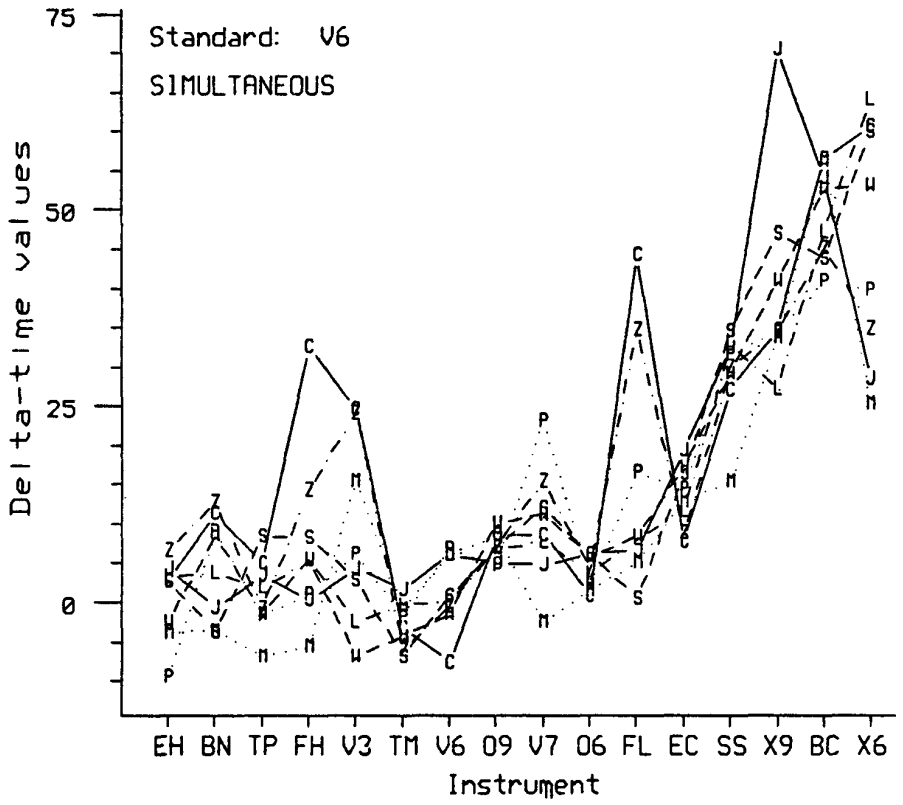

Figure 4.4 Mean Δt (RPAT) values for the 16 instrument tones used in Experiment II, as compared with standard V6. Mean values are plotted individually for each subject.

to the RPAT responses from Experiment II. Some of these density graphs are shown in Figures 4.5 and 4.6.

Figure 4.5 shows some bimodal distributions (upper two graphs), flattened "plateau" distributions (middle two graphs), and skewed distributions (lower two graphs) for standards EC and BN. However, most of the densities (24 out of 32) for these two standards were similar to a normal distribution or only slightly skewed. For standard V6, normal-like distributions were obtained for only eight of the sixteen tones; the other densities were similar in shape to the four illustrated in Figure 4.6.

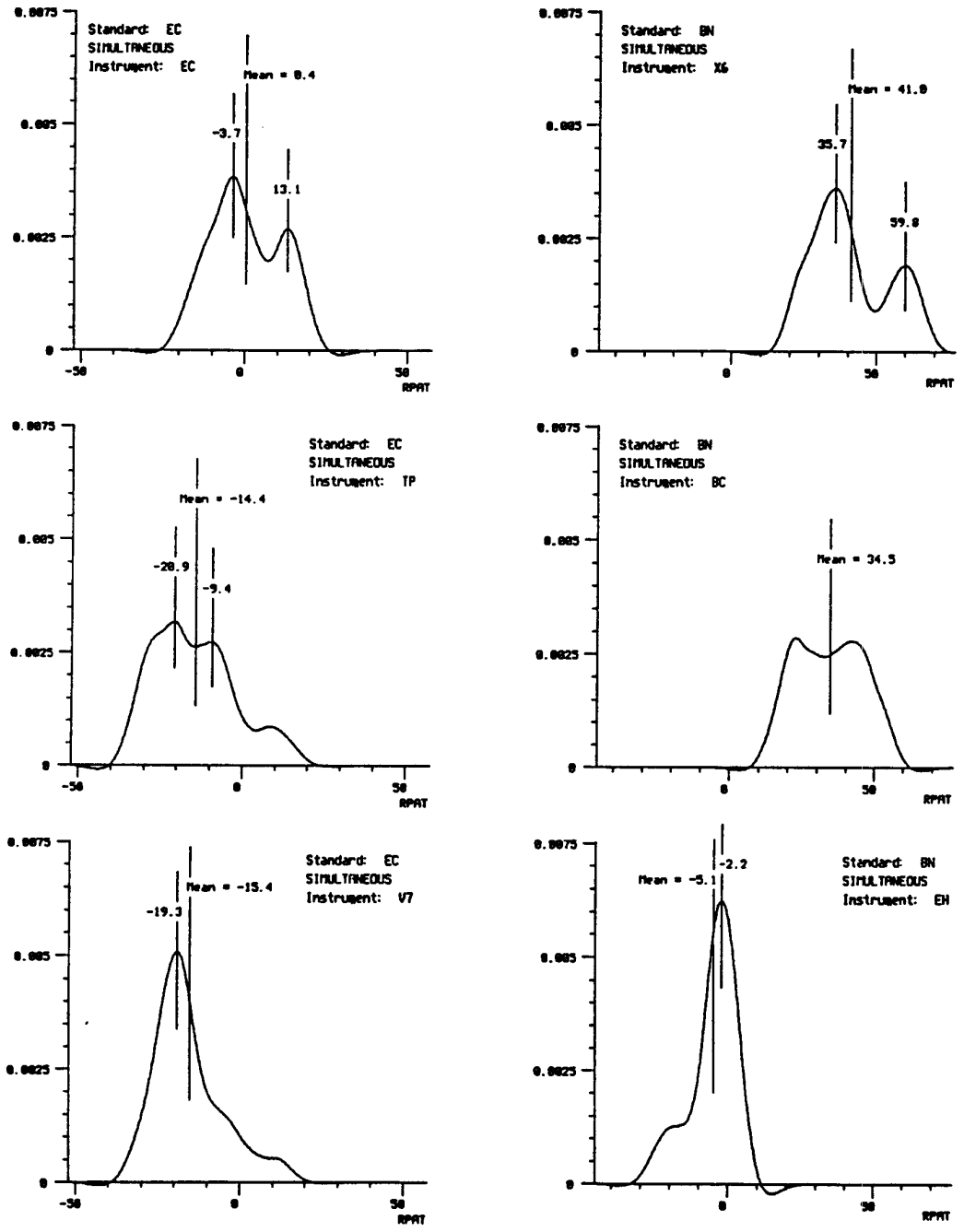

Figure 4.5 Probability density curves for some of the tones used in Experiment II. The left graphs represent 3 of the 16 tones synchronized with standard EC; the right graphs represent 3 tones synchronized with standard BN. Abscissa coordinates represent empirical RPAT (Δt) values. Vertical lines marking the mean response are shown for each instrument. In addition, other vertical lines mark certain modes.

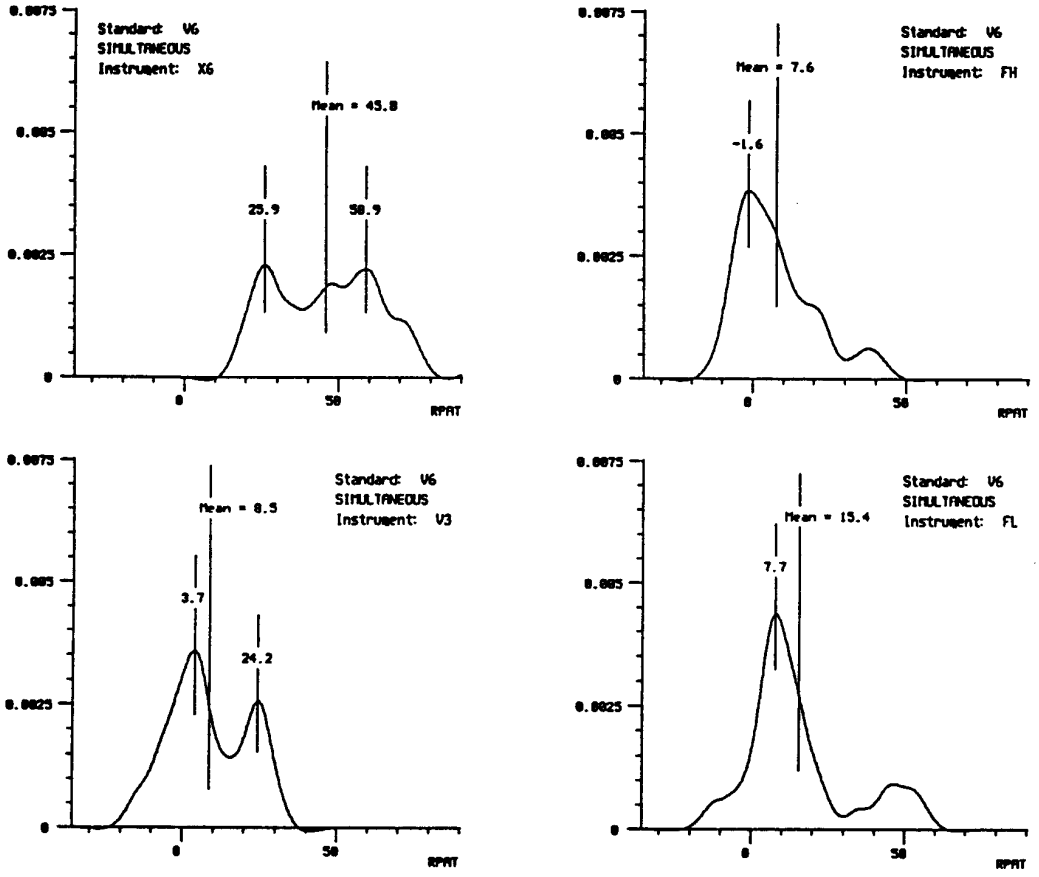

Figure 4.6 Probability density curves for 4 of the 16 stimulus tones synchronized with standard V6 in Experiment II. Coordinates and vertical lines are as in Figure 4.5. Each distribution in both Figure 4.5 and 4.6 represents 48 judgments.

The bizarre shapes of the density graphs seen in Figures 4.5 and 4.6 can be accounted for by an unexpected side effect of the experimental paradigm involving simultaneous judgments. This side effect was that many of the stimulus pairs fused perceptually into a single tone, with a single attack, over a wide range (50 ms in some cases) of relative onset times. Sometimes the fused timbre of the pair changed continuously over the entire course of the range of onset times; in other instances the timbre remained essentially constant.

Bimodal distributions, such as the top two graphs of Figure 4.5 and the lower left graph of Figure 4.6, seem to be a result of fusion combined with response strategy. Most subjects adopted a

strategy of moving the standard and variable tones closer together just until the two were perceived as simultaneous. When the stimulus pair fused into one attack over a range of onset-time values, this strategy resulted in a response mode for each end of the range. That is, responses tended to cluster around one of two modes, depending on whether the standard tone preceded the variable tone or followed it. Bimodal behavior is evident for the E-flat clarinet paired with itself (see the upper left graph of Figure 4.5), a combination for which one would ordinarily expect a normal response distribution.

Flattened or squashed distributions, such as the middle two graphs of Figure 4.5 and the upper left graph of Figure 4.6, also seem to be a result of fusion, but in these cases responses were spread out over the entire range of relative onset times for which fusion took place. There may be other side effects to the synchronous paradigm (discussed later in this chapter) that could account for the skewed and other non-normal, non-symmetric distributions from Experiment II; but it is possible that they too could be explained by fusion effects. However, regardless of the reasons for strange distribution shapes, it is apparent that responses to instruments compared with standard V6 were more affected by these side effects than those of instruments compared with standards EC and BN.

The motive for Experiment II was to obtain more accurate RPAT measurements; but, because of fusion and possibly other side effects, the use of the synchronous paradigm did not accomplish this goal. This does not mean the results from Experiment II need to be discarded; indeed, at least some of the RPAT measurements from this experiment seem to be *as* accurate as the RPATs from Experiment I—but not *more* accurate. It was hypothesized at this point that using a drum sound as a standard would prevent the stimulus pair from fusing when the attacks of the pair were synchronized, thus affording more accurate RPAT measurements under the synchronous paradigm. Therefore, a third experiment was designed that was identical to Experiment II except that only one standard was used, this standard being a brief drum sound produced by a quick slap of a hand on a drum head. This experiment will now be described, followed by an overall discussion of the results from both Experiment II and Experiment III.

4.3 Experiment III, Method

- *Stimuli*

The stimuli were the same set of 16 instrument tones used in Experiments I and II, with the addition of a drum sound as standard. The drum sound was produced by a human hand slapping the head of a conga drum; this sound was recorded digitally into the computer. The format for the recording was a sample-data sequence, the same format used for the stimuli in Experiment II. The drum sound will henceforth be referred to as SLAP, and is illustrated in Figure 4.7.

- *Procedure and Apparatus*

Procedure and apparatus for Experiment III were identical to those used in Experiment II.

- *Design*

A trial consisted of one of the 16 tones paired with the SLAP standard. The order of trials within a set of 16 was determined randomly, and each set was replicated a total of 5 times. There were thus 80 trials for each subject. Control over relative onset time between **A** and **B** was the same as in Experiment II; the accuracy of each RPAT measurement is therefore within ±0.5 ms.

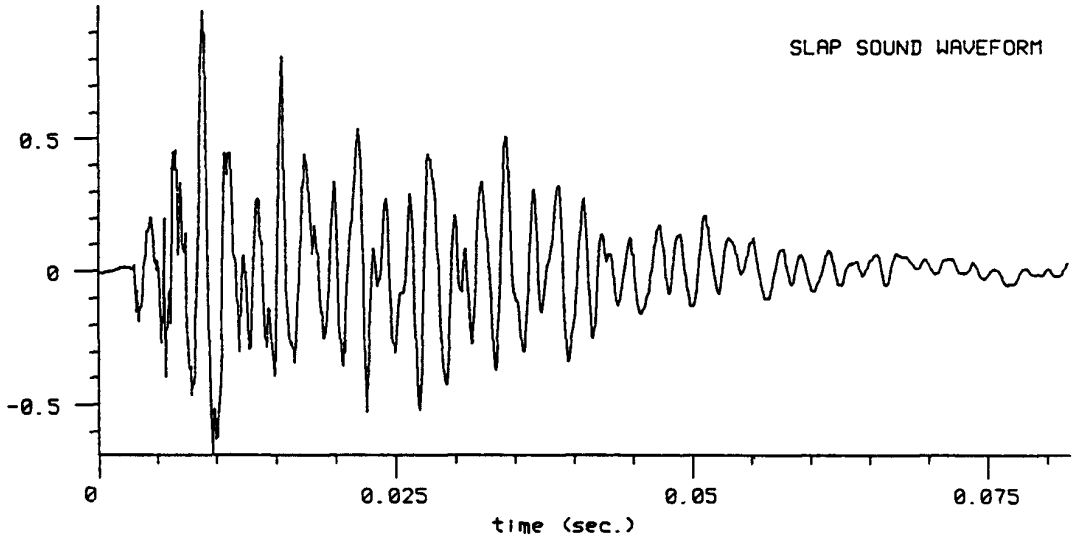

Figure 4.7 Sound waveform representation of the drum sound (hand slap) used as a standard in Experiment III. Abscissa values are in seconds (total duration is 80+ ms) and ordinate values represent amplitude normalized on a scale of −1.0 to 1.0.

- *Subjects*

Ten subjects participated in the experiment, including the eight subjects from Experiment II. All of the subjects were experienced in computer music and considered to have well-trained ears.

Table IV.3 2-way Analysis of variance for Experiment III. All factors are significant at the 99.9% level of confidence.

Factor	Mean Square	df	F	p
Instrument	5,059	15	37.712	< 0.001
Subject	1,732	9	12.912	< 0.001
Inst × Subj	221	135	1.647	< 0.001
Error	134	640		

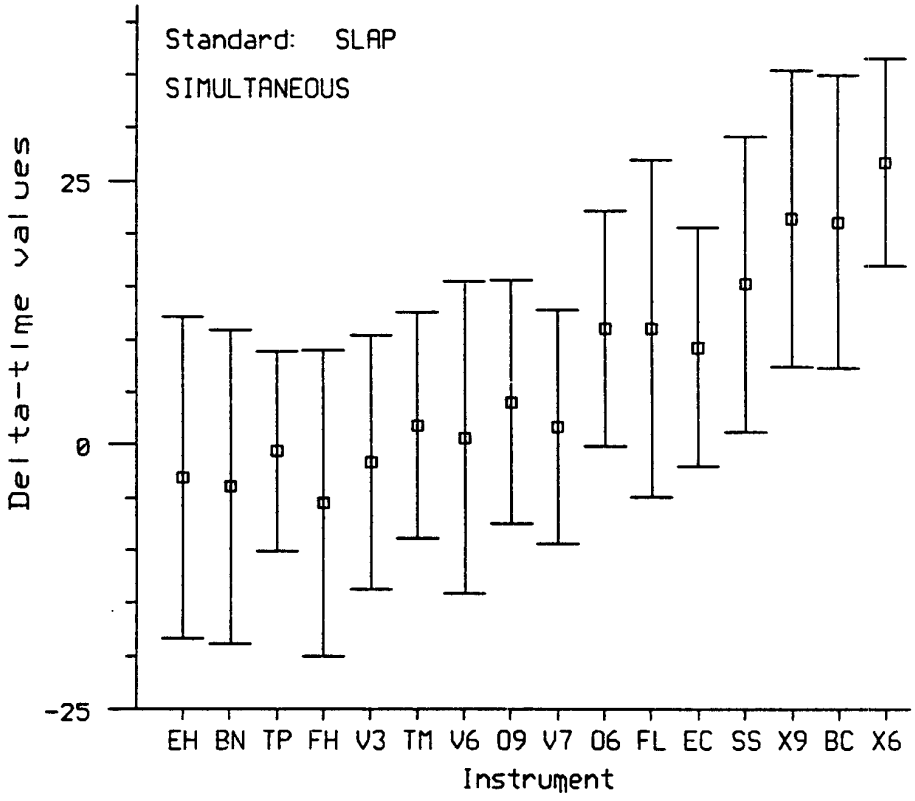

Figure 4.8 Mean Δt for 16 instrument tones, representing RPATs relative to the standard SLAP. Vertical lines extend 1 standard deviation on either side of the mean.

4.4 Experiment III, Results

Results of a 2-way ANOVA applied to the data from Experiment III are shown in Table IV.3. We can see that the factor due to instruments is again significant (F-ratio of 37.7, df = 15,640), as are the factors due to subjects and to instrument-subject interaction. Contrary to the findings for Experiment II, the variance attributable to subjects far exceeds the variance for the interaction term.

Overall instrument means and standard deviations are displayed in Figure 4.8. Individual subject means for each instrument are shown in Figure 4.9. As can be seen from these graphs, there was wide variability of response, especially when compared to previous findings.

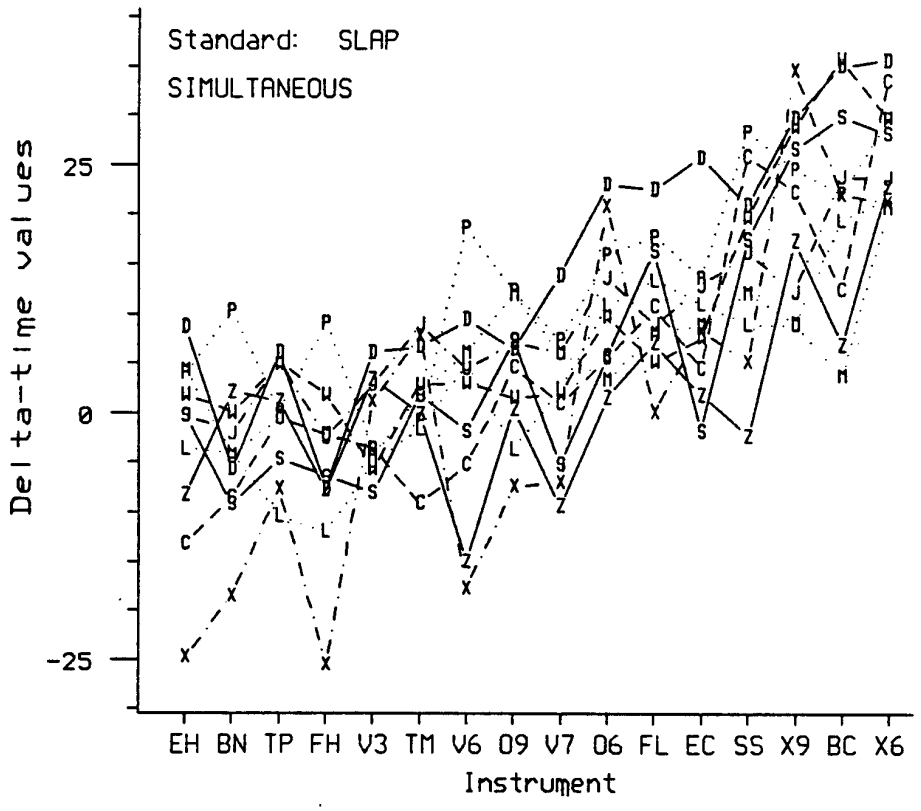

Figure 4.9 Mean Δt (RPAT) values for the 16 instrument tones used in Experiment III, as compared with standard SLAP. Mean values are plotted individually for each subject.

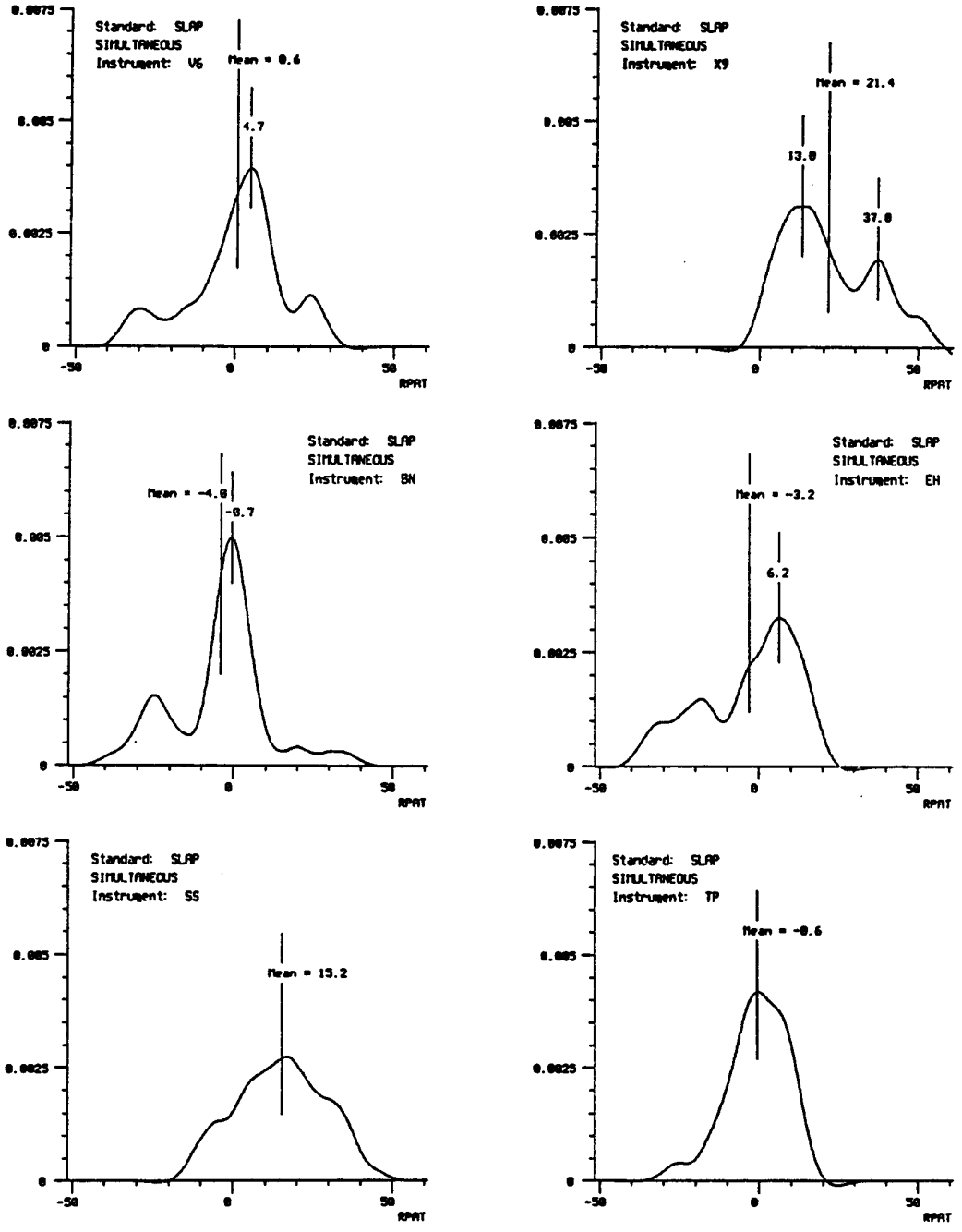

Figure 4.10 Probability density curves for 6 of the 16 tones synchronized with standard SLAP in Experiment III. Abscissa and ordinate values, as well as vertical line markings, are as in Figures 4.5 and 4.6. The strange shapes are typical of all 16 stimulus distributions. Each distribution represents 50 judgments.

Figure 4.10 gives us more insight into the results of Experiment III. Shown are density graphs representing response distributions for six of the 16 stimulus tones. The wide spread of responses for instrument V6 (upper left graph of Figure 4.10) was also obtained for instruments O9, TM, O6, and V3. Densities for instruments FH, FL, and V7 exhibited gross skews similar to those for BN and EH (middle two graphs). The response for TP (lower right graph) is the one that best approaches a normal distribution, but it is still less than ideal. Instruments EC and X6 had similar density graphs.

Subjects that took all three experiments agreed that Experiment III was the most difficult. The fusion problem evident in Experiment II was avoided, but a worse problem took its place. Auditory acuity of temporal order became a factor; subjects could usually tell that the percepual attacks of a stimulus pair were not synchronous, but they could not tell which tone of the pair preceded the other. This confusion was probably enhanced by the inherently different attack characteristics of the SLAP 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 (see Figure 4.7), whereas the quickest rise times (measured analogously) among the other 16 stimuli were 45–50 ms (see Table III.2). Some subjects reported simply giving up on several of the trials.

It is curious that subjects should find it more difficult to synchronize instrument tones with a percussion instrument than with another non-percussion instrument tone. After all, the drum is the standard rhythm-setting instrument to which all other instruments synchronize. However, in a musical context, a mental sense of the rhythm is also available to the performer; synchronization in that case seems easier than the task involved in Experiment III. It is worth noting that Morton *et al.*, applying an experimental paradigm to spoken-digit stimuli that was virtually identical to that used in Experiments II and III, found higher variance in their data when a click was used as the standard than when another spoken syllable was used as the standard [*Morton, Marcus, and Frankish* (1976)]. Another possibility is that the attack portions of the variable tones were considerably masked by the impulsive attack of the SLAP standard, making synchronization difficult regardless of the setting of the relative onset times.

In general, then, the results from Experiment III are undependable and have to be disregarded.

Response distributions, as seen in Figure 4.10, are too skewed, squashed, or spread out to yield confident measurements of RPAT, and subject variance is wider than the variance from Experiments I and II.

4.5 Discussion

Despite the disappointing results from Experiment III, we still have enough data from Experiments I and II to answer our three major questions. The most straightforward one to address is Q3, dealing with subject differences. We found close agreement among subjects in Experiment II, especially for standards EC and BN. In Experiment III, subject variance was wide; in Experiment I, it was moderate. Thus, a general conclusion is that subjects agree with regard to PAT in some instances and disagree in others.

By looking at the data in more detail, however, we find reasons for whatever disagreement exists. First of all, when tones were combined simultaneously, there was often a wide range of relative onset times between the pair of tones for which fusion occurred. Within this range, the attacks of the individual tones could not be resolved separately. In such a case, a subject's choice of onset asynchrony for which the stimuli's PATs sounded synchronous usually fell within the range for which fusion occurred—but *where* within that range depended on the particular strategy the subject had adopted for taking the experiment. For instance, some subjects always placed the standard ahead of the variable tone, regardless of the initial onset asynchrony determined by the computer, before moving the stimuli together to obtain perceptual synchrony. This was especially true in Experiment III (standard SLAP), in which it was clear which of a stimulus pair was the standard. Other subjects did the reverse, beginning each trial by placing the variable tone ahead of the standard. These two strategies tended to result in responses that were at either end of the range over which the stimuli fused into one tone. A third set of subjects used timbre cues when fusion occurred to help determine their choice of onset asynchrony; their responses tended to fall in between the responses for the other two sets of subjects.

A second reason for disagreement among subjects is that for certain standards (SLAP and V6), the synchronization task was more difficult than for other standards. The difficulty may have been due to masking or fusion; it may also have been due to the standard's attack being either very impulsive (SLAP) or very sluggish (V6). When two stimuli differed considerably in their attack characteristics, it was apparently difficult for subjects to synchronize the PATs of the stimuli. The difficulty of the task implies that some limits of auditory acuity were being approached or exceeded; if so, one would expect subjects' responses to scatter somewhat.

In Chapter III, it was concluded that differences among subjects for Experiment I were probably procedural rather than perceptual, and that subjects really don't vary in the way they hear PAT. Experiments II and III have not presented any evidence for altering that conclusion. It is possible, though, that perceptual differences among subjects could occur if subjects, in making their judgments of PAT, were to assign different weights to acoustical events that cue PAT (onset/tonguing noise, sudden increase in amplitude, or certain spectral changes),—but even then only if these events were not to take place simultaneously.

Now let us address Q1, concerning the accuracy of PAT measurements. We have seen that Experiments II and III did not yield more accurate measurements of RPAT than Experiment I. However, it was suggested at the end of Chapter III that the set of Δt values from Experiment I might need to be adjusted somewhat to obtain true RPAT representations for the 16 stimulus tones. Results from Experiments II and III may reinforce or contradict findings from Experiment I, and thus help us assess whether or not an adjustment of the original set of RPAT values is indeed necessary. If the set of Δt values from Experiments II and III are *significantly* different from the original set, however, a simple adjustment of that set is not appropriate; rather, it would be necessary to compile two sets of values, one for isochronous combinations of stimuli and one for simultaneous combinations. That is, a significant difference in the sets would imply that two separate auditory processes may be evident, one for judgments of isochronism and another for judgments of synchrony. Thus, it is useful to compile a comparison of all the sets of RPATs from the three experiments. Such a comparison will give us more than one sampling of the same RPAT value for each instrument, and will also indicate whether RPATs of stimuli presented isochronously are the same as or different

from RPATs of stimuli presented synchronously.

The data from Experiment III were found to be unreliable, and the RPATs compared to standard V6 in Experiment II were more variable than the RPATs compared to the other two standards (EC and BN). Hence, these two sets of Δt values will be disregarded in our overall comparison. This leaves three sets of RPATs, two from Experiment II (synchronous judgments with standards EC and BN) and one from Experiment I (isochronous judgments with standard EC). Since two of these three sets are RPATs relative to instrument EC, it will be convenient to use EC as the same reference instrument for all three sets. Thus, we will adjust each set of Δt's with a different additive constant such that the RPAT value for instrument EC in each set will be 0.

The three sets of RPATs adjusted in this manner are listed in Table IV.4. The first column of numbers (obtained from Experiment I) is the same as the list of mean Δt's from Table III.2 with 0.8 ms subtracted from each value (the Δt for EC in Table III.2 is 0.8). The second column in Table IV.4 is the same as the first column from Table IV.2 with 0.4 ms subtracted from each value; this column represents RPAT values from Experiment II that are relative to standard EC. The third column of numbers lists RPATs from Experiment II that are relative to standard BN; this column is the same as the second column from Table IV.2 with 15.6 ms subtracted from each Δt value.

The fourth column of numbers in Table IV.4 is a list of modal values obtained from the density graphs of the various response distributions. These modes are thus an added check on the consistency of RPAT values. Only one column is given instead of three, since for almost all instruments there was at most one of the three cases for which a mode distinctly differed from the mean Δt; the only exception was instrument BC, for which the same mode occurred twice. (The spaces for instruments TM and EC are blank because mode and mean coincided in all three cases.)

In general, the RPATs from Experiment II *do* seem to corroborate the values from Experiment I. Especially consistent are responses for instruments BN, FH, O9, O6, EC, SS, and X9; responses for instruments EH and X6 are not quite so consistent, but the three RPATs for each instrument are still within $\sim$6 ms of each other, which is less than 1 standard deviation. There are some definite inconsistencies, however: Instruments V3, FL, and V7 exhibit a clear dichotomy between

Table IV.4 Overall comparison of 3 sets of 16 mean Δt values. The sets are for standards EC and BN from Experiment II and for standard EC from Experiment I. The set of values for standard BN have been decreased by 16.8 ms; this adjustment yields Δt values that are comparable to those for the other two sets. Also given are modal values for the instruments that differ from some of the respective mean values. The modes occurred in one of the three experiments and were obtained from the density graphs. For cases in which no clear mode differed from the mean, no modal value is indicated. The three sets of Δt values have been averaged to yield an overall mean for each instrument. The final column lists values which the investigator feels best represent instrument RPATs relative to standard EC.

Instrument	Exp. I Mean Δt	Exp. II—EC Mean Δt	Exp. II—BN Mean Δt	Mode	Overall Mean (SD)	Chosen Mode
EH	−17.6	−24.2	−20.7	−17.8	−20.8 (8.6)	−19.0
BN	−17.9	−16.0	−16.0	−20.0	−16.6 (10.4)	−18.0
TP	−13.5	−14.8	−21.2	−18.0	−16.5 (10.7)	−16.5
FH	−12.1	−11.3	−14.2	−16.0	−12.5 (9.1)	−15.0
V3	−10.2	−27.6	−29.4	−15.3	−18.9 (10.5)*	−15.0
TM	−8.3	−13.3	−15.6		−12.4 (9.1)	−12.0
V6	−12.1	−13.9	−20.2	−13.9	−13.0 (10.7)*	−12.0
O9	−8.6	−8.8	−11.6	−11.8	−9.7 (9.4)	−11.0
V7	−5.9	−15.8	−17.1	−12.3	−10.8 (11.0)*	−10.0
O6	−5.4	−3.7	−6.4	−7.9	−5.2 (9.1)	−6.0
FL	5.8	−10.1	−13.1	−2.0	−5.8 (10.9)	−4.0
EC	0.0	0.0	0.0		0 (9.7)	0.0
SS	8.7	8.2	11.1	7.2	9.3 (8.5)	8.0
X9	15.0	15.3	14.5	13.4	14.9 (9.6)	14.0
BC	24.5	24.3	18.9	20.3†	22.6 (11.5)	21.0
X6	31.9	27.7	25.4	28.8	28.3 (11.9)	28.0

*Averages are over the EC cases only.

†This mode occurred twice.

the isochronous case and the simultaneous cases; instrument TM exhibits this dichotomy to a lesser extent. Instruments TP, V6, and BC, on the other hand, display some variance between the two standards from Experiment II.

Some of these inconsistencies are easily explained. For instance, some of the density graphs for instruments TP and BC are far from normal; mean Δt's are not necessarily representative of RPAT in these cases. Also, many of the modal values listed in column four of Table IV.4 tend to ameliorate certain inconsistencies. However, the dichotomy between synchronous and isochronous cases displayed by the string tones (V3, V6, V7), FL, and TM need to be discussed in more detail.

By reviewing the amplitude envelopes of the sixteen instrument tones (shown in Figure 3.4), we can see that their attack portions can be grouped into three categories, represented by the three graphs of Figure 4.11:

- Rapid increase in amplitude occurring essentially simultaneously with physical onset (similar to EH's attack),
- Rapid increase in amplitude, such as in the previous case, but preceded by at least 20–30 ms of low-amplitude but audible onset activity (similar to EC's attack), and
- Gradual but steady increase in amplitude beginning with physical onset (similar to V6's attack).

The five tones whose RPATs showed an isochronous-synchronous dichotomy all fall in the last category,—and the dichotomy can be accounted for by the gradual rise times these tones exhibit. Let us assume for the time being that the PATs for tones represented by the left two graphs of Figure 4.11, *for both the isochronous and synchronous paradigms,* are determined by the time of rapid increase in amplitude. This assumption is supported by the Δt's listed in Table IV.4. Now consider the attacks for the tones represented by the graph of V6's amplitude; these attacks are much less impulsive than for the other tones, and hence this strong cue for PAT is missing. Therefore, subjects

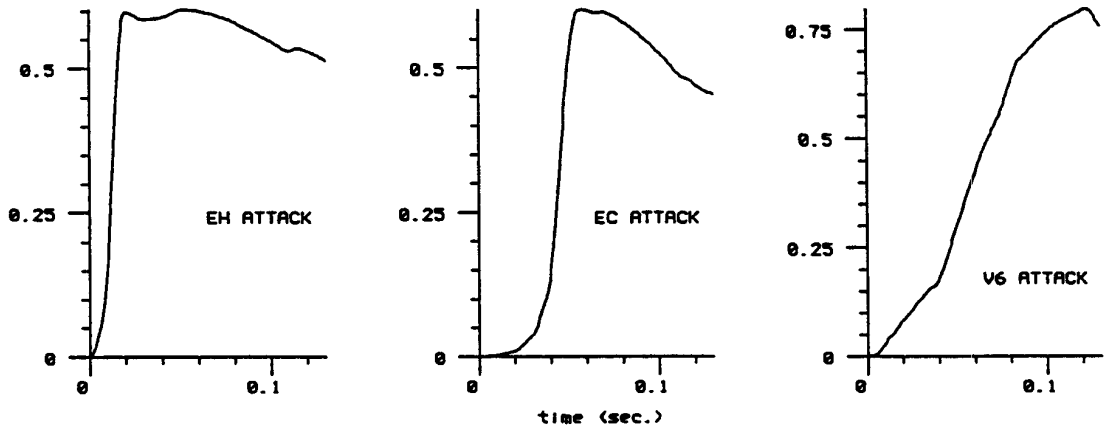

Figure 4.11 Attack portion (130 ms) of 3 of the 16 amplitude envelopes displayed in Figure 3.4. Units are normalized amplitude *vs.* time in seconds. The attacks are representative of 3 categories: (1) Rapid amplitude increase simultaneous with physical onset (EH), (2) Rapid amplitude increase delayed from onset (EC), and (3) gradual amplitude increase beginning at onset (V6).

must use acoustic cues other than rapid increase in amplitude in making their judgments of PAT for these tones. Furthermore, subjects seem to use different cues depending on whether tones are presented isochronously or synchronously.

When tones are presented isochronously, PAT judgments for tones without impulsive attacks seem to be primarily influenced by the long rise time these tones exhibit. Even though, as in the first category of attacks, considerable energy is present immediately following physical onset of these tones, the long rise time seems to cause subjects to hear PAT coming 10–20 ms after onset. Another cue that might be influencing PAT for instruments FL and TM is a noticeable spectral change coming tens of milliseconds after onset. The spectral change for instrument FL is the chiff effect resulting from the overblown octave; the fundamental rises in amplitude much later than the other harmonics. A similar spectral change occurs for the muted trombone (TM) tone, though the harmonic in question is the fourth rather than the fundamental. Since these spectral shifts mark what is essentially the beginning of the steady-state, they may be significant cues to PAT. If so, the subject would be forced to compromise the cue coming from onset energy with the spectral cue. That is, the subject might place PAT later than he would if the spectral cue were not present.

When a tone with non-impulsive attack is combined simultaneously with another tone, the long rise time tends to go unnoticed, being masked by the second tone's attack and steady-state. PAT for this tone is then essentially cued by perceptual onset, and will be judged by subjects to be much earlier than in the isochronous case. A comparison of the right two graphs of Figure 4.11 will make these points more clear; the onset amplitude for instrument V6 (representative of the five instrument tones we have been discussing) considerably exceeds that for instrument EC up to $\sim$40 ms, yet V6's maximum amplitude comes $\sim$70 ms after EC's maximum. Thus, for these five tones, when onset is more influential than rise time (simultaneous judgments of PAT), RPAT relative to EC will be earlier than when rise time is more influential (isochronous judgments of PAT); this is borne out by the Δt values listed in Table IV.4 for instruments V3, V7, V6, FL, and TM.

In short, PATs for tones with impulsive attacks seem to coincide with the physical time of attack regardless of whether presented isochronously or synchronously with other tones; however, PATs for tones with non-impulsive attacks seem to depend on the kind of presentation. Another

point worth mentioning here is that the duration of the attack for these latter tones is more perceptually apparent to the listener, and their PATs are not heard as single moments. (See the discussion on *Discrimination of Rise Time* in Section 2.3.) In other words, if one of these tones is combined with some other tone, either isochronously or synchronously, such that the Δt between onsets falls anywhere within an appropriate RPAT range, subjects will still hear the tones as being isochronous or synchronous. Thus, the dichotomy of responses for these tones under the two different presentation conditions is somewhat artificial; RPATs somewhere in between the two extremes should be applicable to both situations.

We see then that the data from Experiment I are for the most part duplicated by the results from Experiment II. Therefore, a simple adjustment of the original set of RPAT values, in preparation for further testing of PAT prediction models, seems to be appropriate. The accuracy of all RPAT measurements appears to be within 2–3 ms; but Q1 cannot truly be answered until these measurement values are independently verified. Such a verification will be discussed in Chapter V.

The adjustment of RPAT values was accomplished as follows: The three sets of RPATs shown in Table IV.4 were averaged together; the set of average values is listed in the penultimate column. (Listed in the same column are average standard deviations; these numbers carry no statistical significance and will be used only in certain graphs shown in the next section.) Because of the dichotomy of values for the string tones, only the RPATs relative to EC were used in calculating the averages for V3, V6, and V7. Since there were modal values in some of the response distributions (listed in column 4 of Table IV.4) that differed from the respective mean Δt values, the investigator felt the overall average set of RPATs needed some slight further adjustment; he therefore estimated, to the nearest millisecond, an RPAT value for each instrument based on all of the mean and modal Δt values obtained empirically. The set of estimated values are listed in the rightmost column of Table IV.4; Δt's in this set differ only slightly from the overall average set. The set of estimated RPATs will be correlated with APAT values predicted by the various models presented in the next section.

4.6 Prediction Models

In order to answer Q2 adequately, we need to be exhaustive in our selection of PAT prediction models. If more than one model is an accurate predictor of all 16 PATs, it will prove useful to determine which model is the best one according to the three criteria established in Chapter III. (The three criteria are successful prediction of PAT, consistency with auditory theory, and practicality in application.) On the other hand, if no accurate prediction model can be found, we need to make sure all possible models have been tried.

In Chapter III, the ABS, PCT, and ENE models were tested using the original set of RPAT values obtained from Experiment I. Since the overall set of RPAT values presented in the previous section is different from the original set, it seems useful to test these three models using the new set of Δt's. Results are discussed below. Also discussed in this section are several algorithms for measuring rise time and envelope slope; these algorithms were used to develop other, more successful PAT prediction models.

ABS, PCT, and ENE Revisited

To test the ABS, PCT, and ENE models, precisely the same procedure was used as before: correlation and linearity measures were obtained over a range of threshold values, and the threshold with the highest such measures was chosen to determine the model's accuracy in predicting PAT.

Results for the ABS model are displayed in Figures 4.12 and 4.13. In Figure 4.12, the amplitude threshold parameter ranges between 0 and 0.3 (on a normalized amplitude scale), with the optimum value being $\sim$0.04; the loudness threshold parameter ranges between 0 and 120 (units corresponding roughly to sones), with the optimum value approximately equal to 48 (though in this latter case there again seems to be a range of threshold values that do equally well). As before, APAT values were predicted by the ABS model with the parameter fixed to the desired threshold; the APATs were then correlated with the new set of RPATs. Graphs of these correlations are shown in Figure 4.13.

By using the adjusted set of Δt values, correlations for both the amplitude and loudness

functions have been improved (compare Figures 3.9 and 3.10). However, the improvement is not enough to generate complete confidence in the accuracy of the ABS model. There are still some predictions more than 5 ms away from the desired line of slope 1, and in general there is too much deviation away from this line. The fit of the model's predictions to the data is very good; but this should not deter us from seeking a model that does even better.

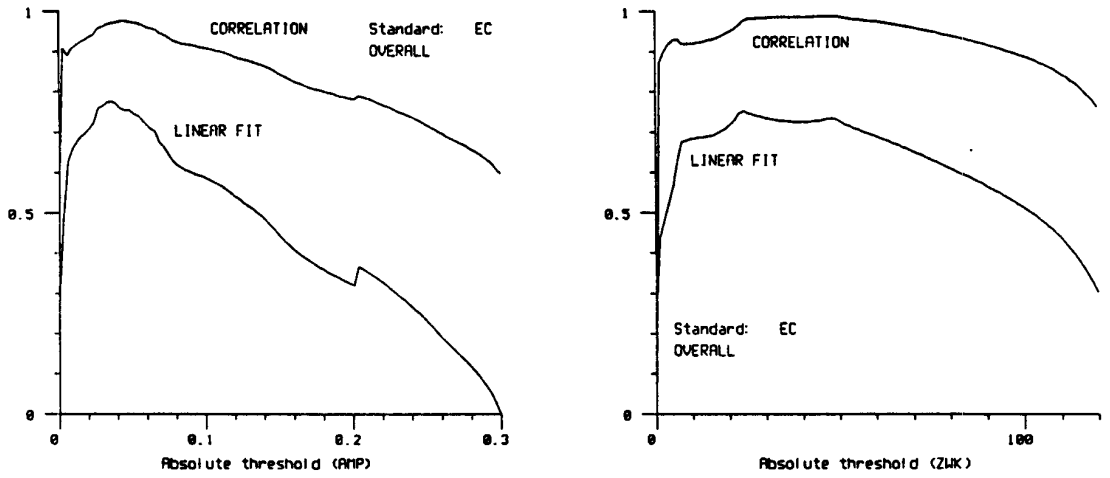

Figure 4.12 Correlation and linearity measures as absolute threshold varies over a range of values. Abscissa scales correspond to the ordinate scales of figures 3.4 (amplitude envelopes) and 3.6 (loudness envelopes).

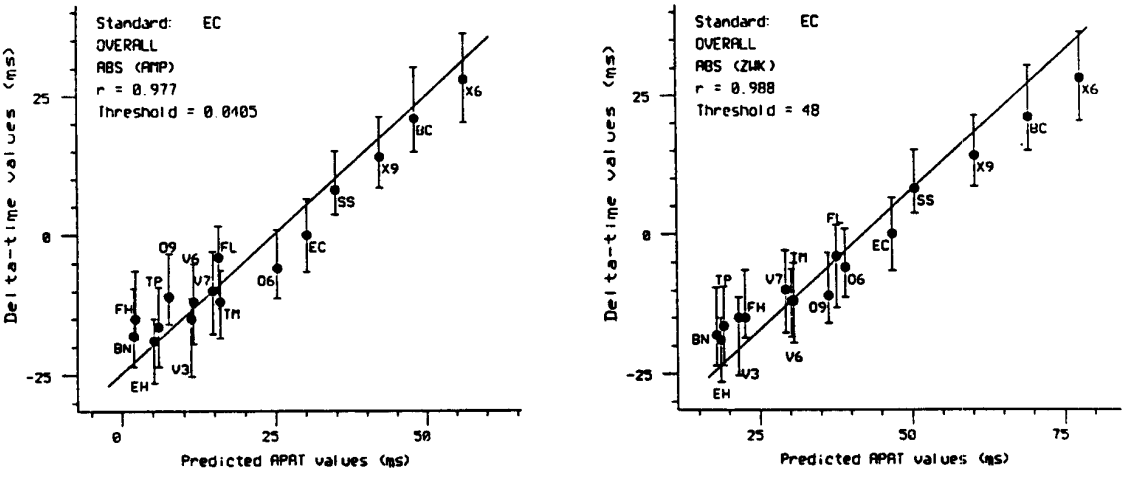

Figure 4.13 The overall set of RPAT (Δt) values correlated with predicted APAT values according to the ABS model. Amplitude (left graph) and loudness (right graph) envelopes were used as input sound representations to the model. Threshold values and product-moment coefficients are indicated.

Let us now examine the performance of the PCT model. The correlation and linearity functions indicated that the best correlation for the amplitude functions would be obtained were the relative threshold fixed at 5.82% of maximum amplitude, and that the best correlation for the loudness functions would be obtained were the relative threshold fixed at 36% of maximum loudness. The correlation graphs are shown in Figure 4.14. Again, correlation has been improved (see Figure 3.11), but the same conclusions just drawn regarding the ABS model can be applied to the PCT model as well.

The ENE model was tested with the new set of RPAT values, but as in the previous case, the model was found to do more poorly than both the ABS and PCT models. (No figures for the ENE model are given.)

Models Based on Slope and Rise Time

In Chapter III was discussed a prediction model based on the slope of the various envelope functions. The slope was approximated by a first-order difference equation, but this approximation was found to exhibit problems that prevented it from being incorporated into a usable prediction model. Nonetheless, it seems appropriate to develop one or more PAT models based in some way on envelope slope, because slope is an important concept in regard to the perception of attack transients.

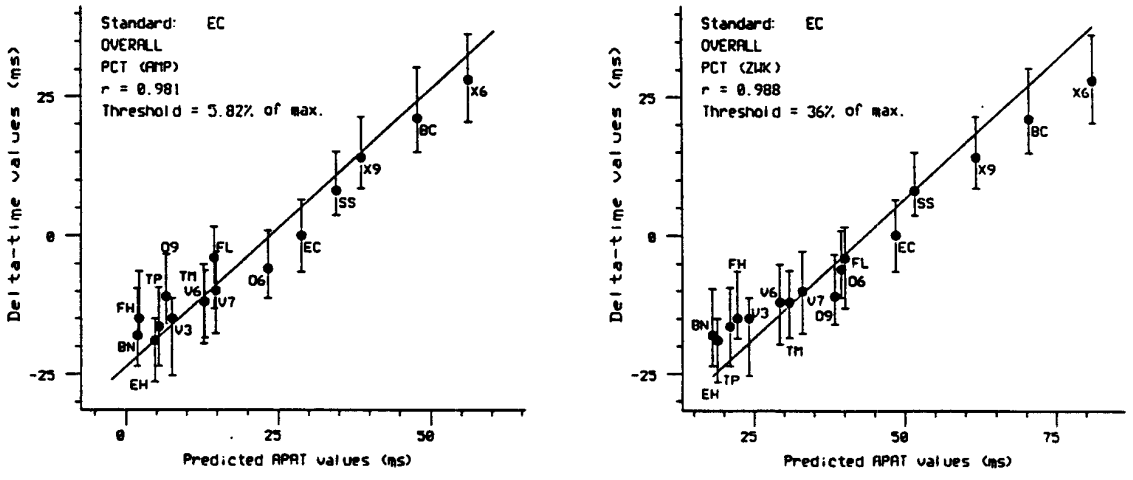

Figure 4.14 RPAT values correlated with APAT values predicted by the PCT model. Examples are given for both amplitude (left graph) and loudness (right graph) envelopes.

To see this, consider the theory of temporal integration (reviewed in Section 2.5). Audible detection of sound depends on the integration of energy by the ear over time. If the sound's amplitude (and therefore power also) increases rapidly,—which means amplitude *slope* is relatively large,—the total integration over a set time interval will be greater than if the amplitude increases slowly or not at all. Furthermore, temporal integration of energy also occurs at suprathreshold levels as a means of determining the sensation of loudness. Thus, a sudden increase in integration will occur whenever there is a sudden increase in amplitude or power slope.

The slope of a function is calculated by taking its derivative. A gross approximation to slope can be had by measuring envelope rise time. That is, there is an inverse relationship such that a short rise time implies a large slope, and a long rise time implies a small slope. The envelope derivative (dy/dt), then, can be approximated by $\Delta y/\Delta t$, where Δy is total amplitude change and Δt is rise time, or the time in which the amplitude change takes place.

Critical to this slope approximation is how rise and rise time are measured. Traditionally, rise time has been measured as the time it takes amplitude to go from one relative threshold to another (relative threshold carrying the same meaning here as it did in the PCT model, namely a percentage of maximum amplitude); the two standard pairs of relative thresholds have been 10–90% of maximum and 0–100%. This method of measuring rise time is reasonable and will therefore be adopted for our purposes of approximating envelope slope—at least for the time being. However, the standard relative threshold pairs used to measure rise time are somewhat arbitrary, and there is no reason to rule out the use of other pairs. For instance, 1% and 70% (40 dB and 3 dB below maximum amplitude, respectively) are both plausible threshold values.

- *The Slope Model (SLP)*

We now have the means to develop a PAT prediction model based on slope, which will be referred to as the SLP model. We could hypothesize that PAT occurs at the time the rise begins (the time amplitude crosses the first relative threshold of the pair); however, this hypothesis would correspond exactly to the PCT model. Let us rather hypothesize that PAT occurs at some time *during* the rise; PAT can thus be expressed as a sum resulting from adding a portion (percentage)

of physical rise time, referred to below as durational threshold, to the time the rise begins. This method of determining PAT was the one used by the SLP model.

There are three parameters inherent to the SLP model: the upper and lower amplitude bounds used to measure rise time and the durational threshold expressed as a percentage of rise time. To test the validity of the SLP model, the two bounds determining rise time were fixed (though several combinations of relative thresholds were tried) and the durational threshold parameter was allowed to vary. As for other models, correlation and linearity measures were then obtained over the range of this parameter.

The best results were obtained when the two bounds defining rise time were set to 6% and 70%. Other combinations tried were 10–90%, 1–90%, 10–70%, 1–70%, and 6–90%. The durational threshold parameter was varied over a range of 1–99% of physical rise time. (Using the two extremes of 0 and 100% would have been equivalent to testing the PCT model with two different relative threshold parameters.) Correlation/linearity functions thus obtained for the 6–70% combination are shown in Figure 4.15. Setting the durational threshold to 7.5% of rise time for the amplitude functions and to 36% of rise time for the loudness functions resulted in the correlations graphed in

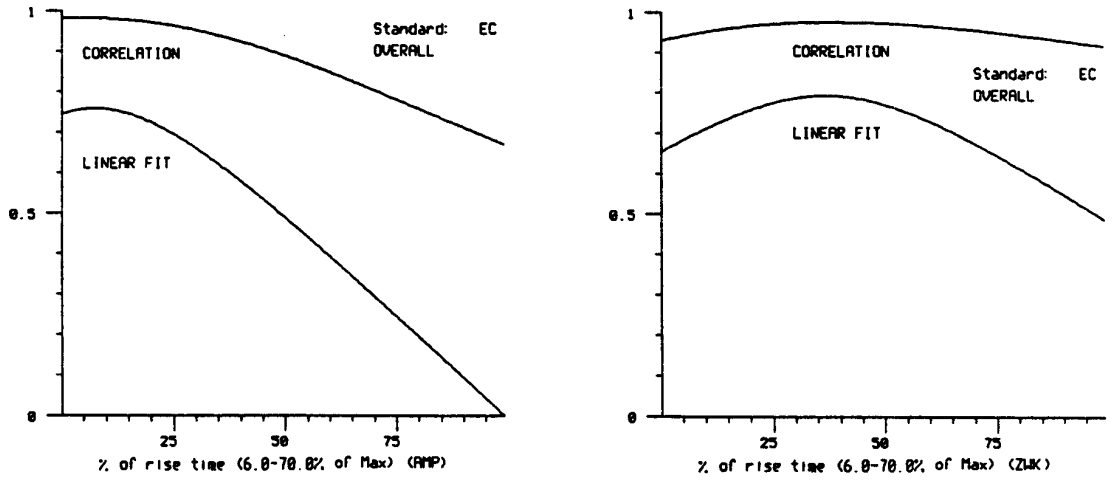

Figure 4.15 Correlation and linearity measures for the SLP model as the durational threshold parameter varies over a range of values. This parameter is defined as a percentage of rise time; rise time is defined as the time it takes the envelope to go from one relative threshold (6% of maximum) to another (70% of maximum). The model predicts APAT as the time the first relative threshold is crossed plus the duration corresponding to the setting of the parameter.

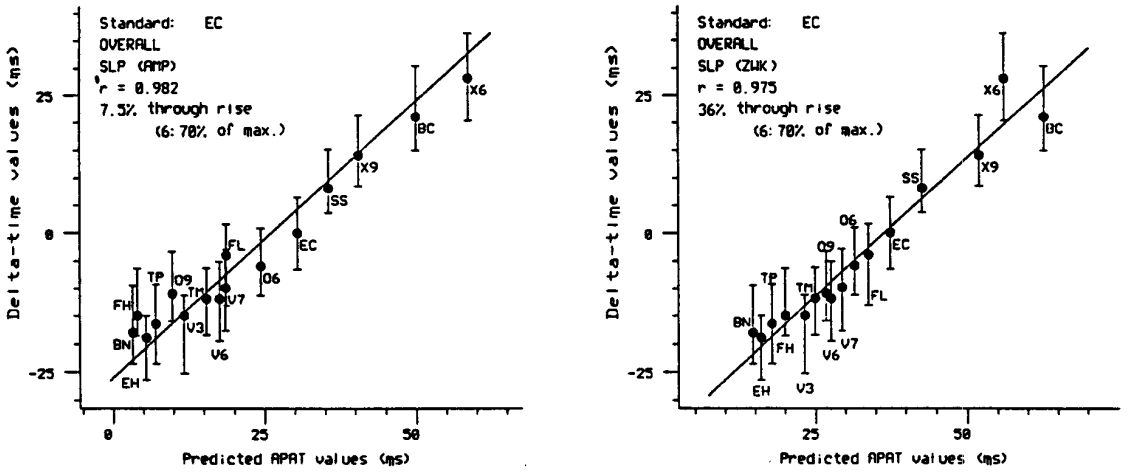

Figure 4.16 RPAT (Δt) values correlated with APAT values predicted by the SLP model. Examples are given for both amplitude (left graph) and loudness (right graph) envelopes.

Figure 4.16.

The results of applying the SLP model to the amplitude envelopes (left graph of Figure 4.16) are very similar to those from the PCT model (see Figure 4.14). Results for the loudness envelopes are almost perfect; the only significant deviations are PATs for instruments X6 and BC. The SLP model is thus seen as not being an improvement over previously tested models. However, the method used to measure rise time is purely operational, and a better way of defining rise time should be developed. Also needing development are more accurate methods for calculating envelope slope (a better approximation to dy/dt than $\Delta y/\Delta t$). With more accurate slope measurements, we could represent the beginning of amplitude rise as the time some slope threshold is crossed. This time could then be used directly as a prediction of PAT; or, as in the case of the SLP model, a portion of rise time could be added to it, with the sum being used to predict PAT.

- *Slope and Rise Time Calculation*

For the slope threshold model presented in Chapter III, slope was approximated by a first-order difference equation. The main problem with this approximation was its sensitivity to noise. What was desired, therefore, was an approximation to slope that represented a more general trend less sensitive to noise. Such an approximation can be achieved by means of linear regression, calculating the best

linear approximation to some small number of points. (It should be recalled that the amplitude, power, and loudness functions are discrete sample-data sequences; hence, linear regression can be applied directly to some subset of sample values.) The linear approximation can be thought of as a tangent to the function envelope; the instantaneous slope of the function (dy/dt) is then equivalent to the slope of the tangent line.

Linear regression analysis was thus used to generate slope functions for all three (amplitude, power, and loudness) sets of function envelopes. These are shown in Figures 4.17, 4.18, and 4.19. The number of points used in calculating each tangent was determined empirically; values between 2 and 25 were tried. The best choice seemed to be 19 samples, corresponding to 19 ms. (It is also roughly equivalent to the 20-ms time constant often cited in the literature as being the limit of resolution for certain temporal events such as succession of musical tones or echoic images in a reverberant environment.) The 19-ms window smoothed out noisy irregularities, but still allowed for enough detail. A new tangent was calculated for each millisecond; thus, there was an 18-ms overlap of input samples from one tangent calculation to the next. The tangent was calculated only for the first 160 ms of each envelope, since the attack portions of all functions were contained within this range.

In the attempt to avoid confusion, we will make an explicit distinction between the three sets of functions just presented and the three sets presented in Chapter III. The sets presented in Chapter III will hereafter be referred to as envelope functions (amplitude, power, and loudness envelopes); the ones displayed in Figures 4.17, 4.18, and 4.19 will be referred to as slope functions (amplitude, power, and loudness slopes).

Some comments should be made regarding the three sets of slope functions. The attack portion (time of sudden increase in amplitude, power, or loudness) for most instruments is evident by a large single peak in the respective slope function. However, as we would expect, the slope functions for the string tones (V3, V6, V7) and FL (and to a lesser extent TM) are flatter, more spread out, and even multi-peaked in some cases. The maximum slope value for these instruments is also less than the maximum slope for other instruments, implying that the amplitude, power, or loudness for these instruments increased less rapidly or suddenly than for other instruments. (The set of 16 slopes

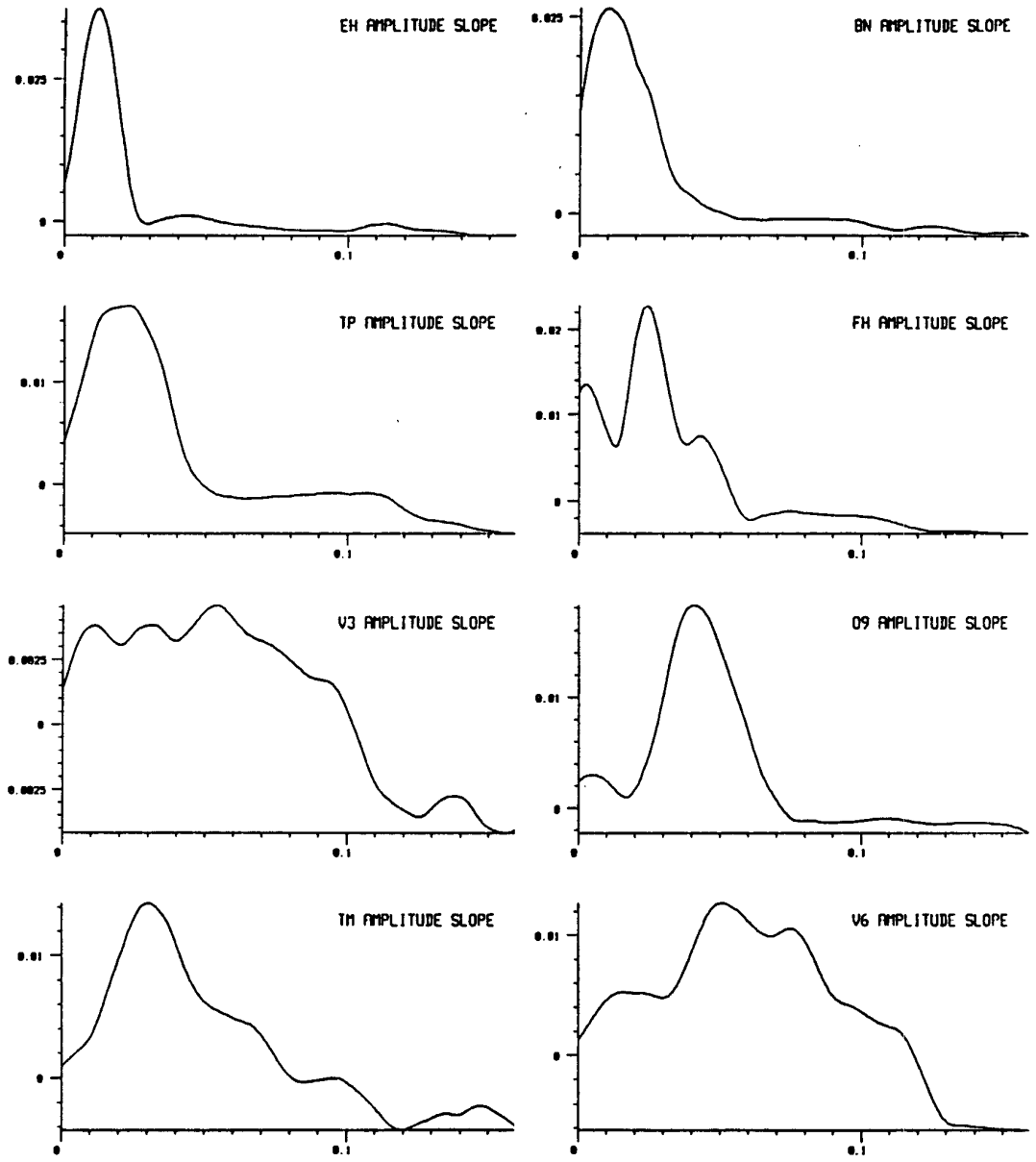

Figure 4.17 *Part 1.* Slope functions for the 16 stmulus tones, obtained from the amplitude envelope functions by means of linear regression. Units are amplitude change per millisecond (slope) *vs.* time in seconds. 8 of the functions are shown above.

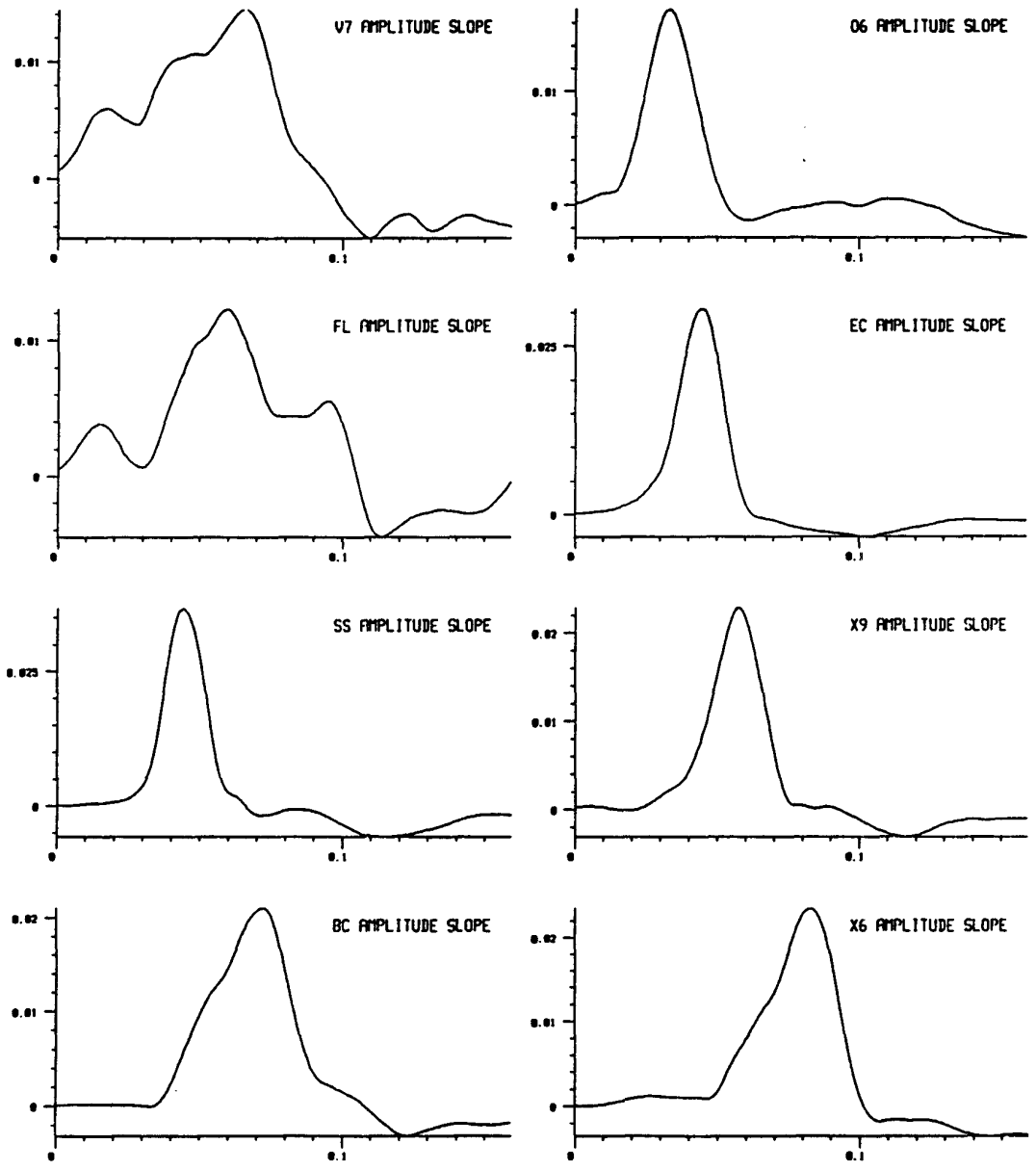

Figure 4.17 *Part 2.* **The remaining 8 out of 16 amplitude slope functions are shown above.**

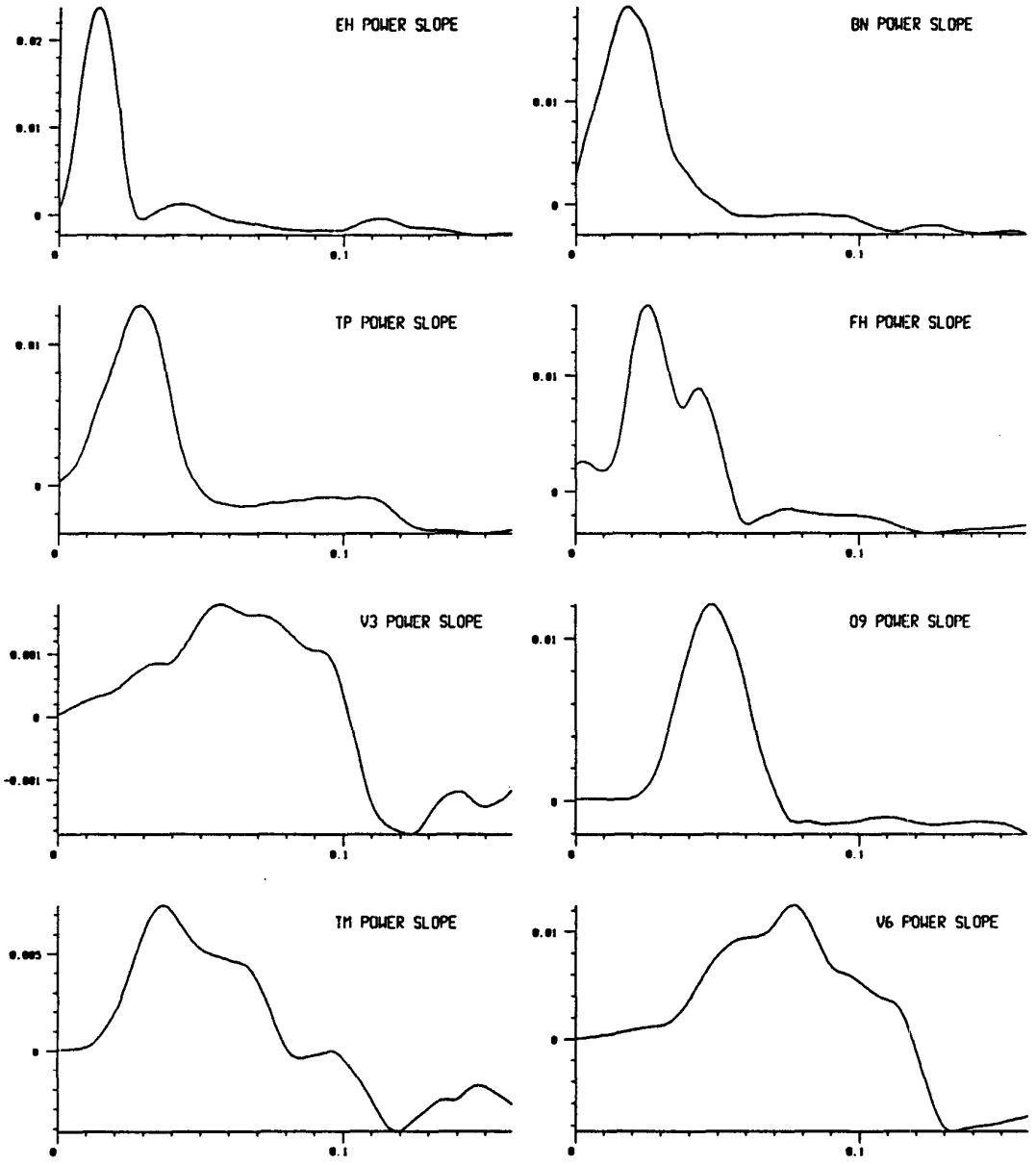

Figure 4.18 *Part 1.* Slope functions obtained from the 16 power envelopes by means of linear regression. Units are change in amplitude squared per millisecond (slope) *vs.* time in seconds. 8 of the functions are shown above.

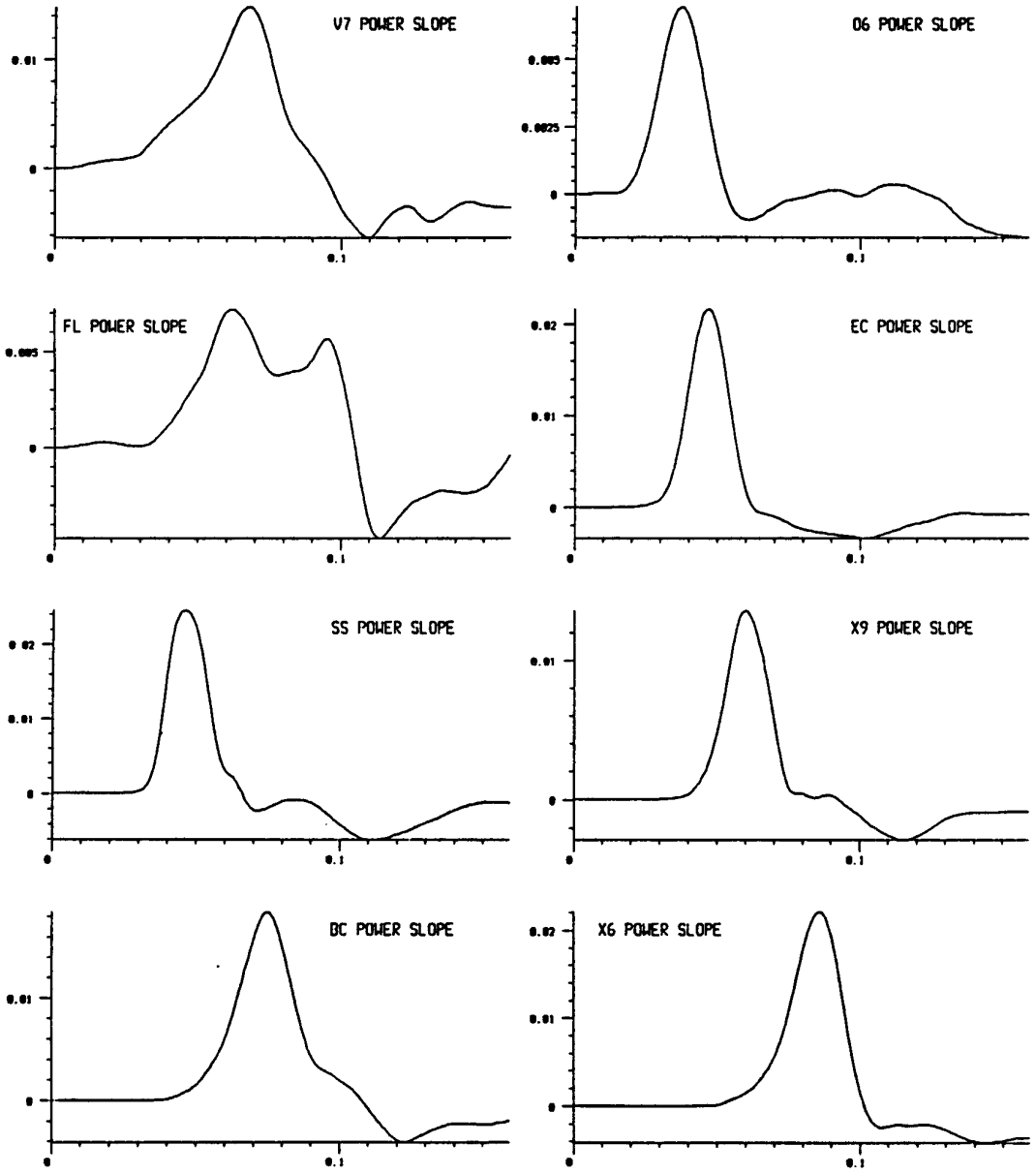

Figure 4.18 *Part 2.* Shown above are the remaining 8 out of 16 power slope functions.

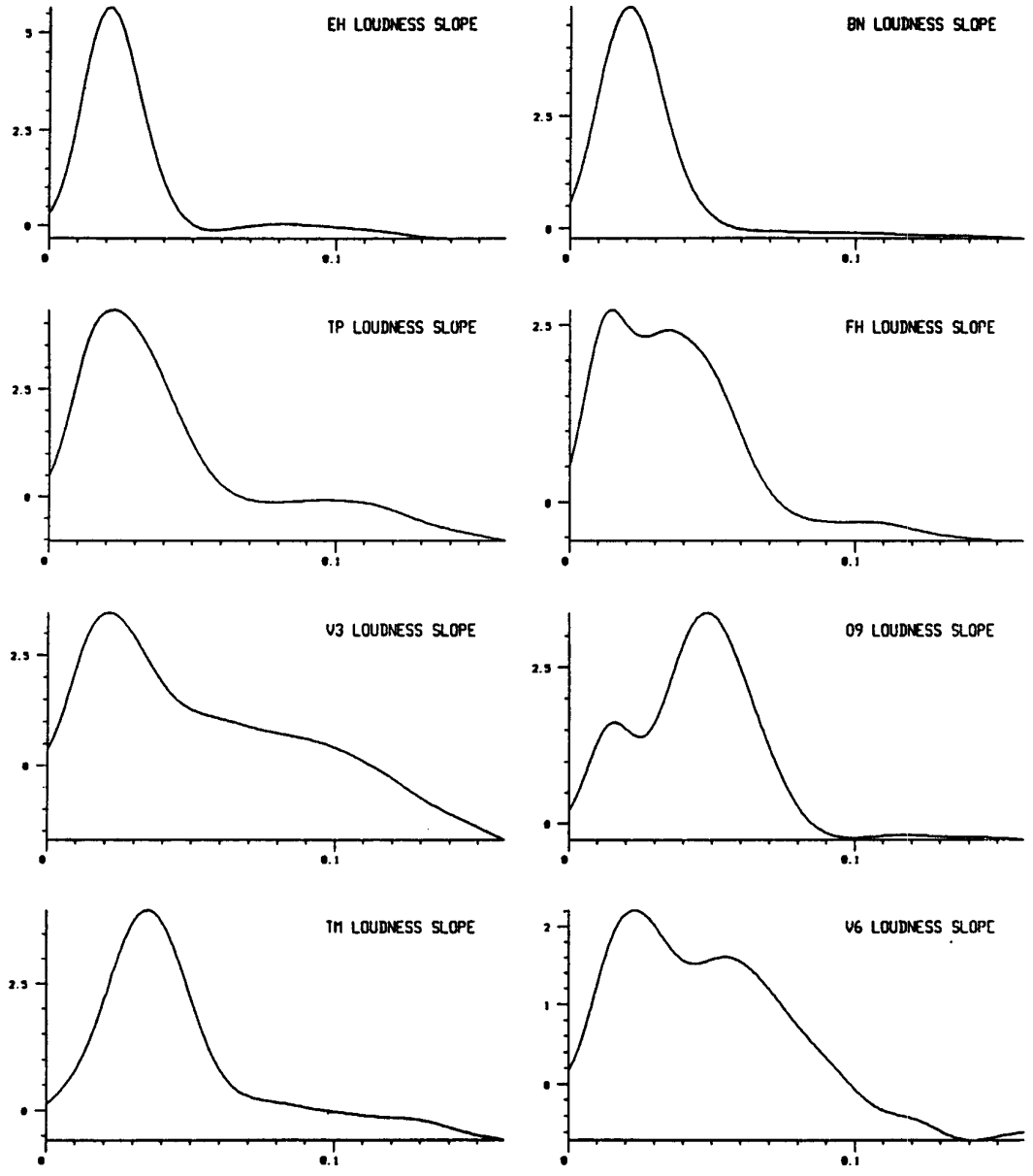

Figure 4.19 *Part 1.* Slope functions of the Zwicker-transform loudness envelopes. The slopes were generated by means of linear regression. 8 of the 16 functions are shown above. The units for the abscissa are *seconds*; those for the ordinate correspond approximately to change in sones per millisecond.

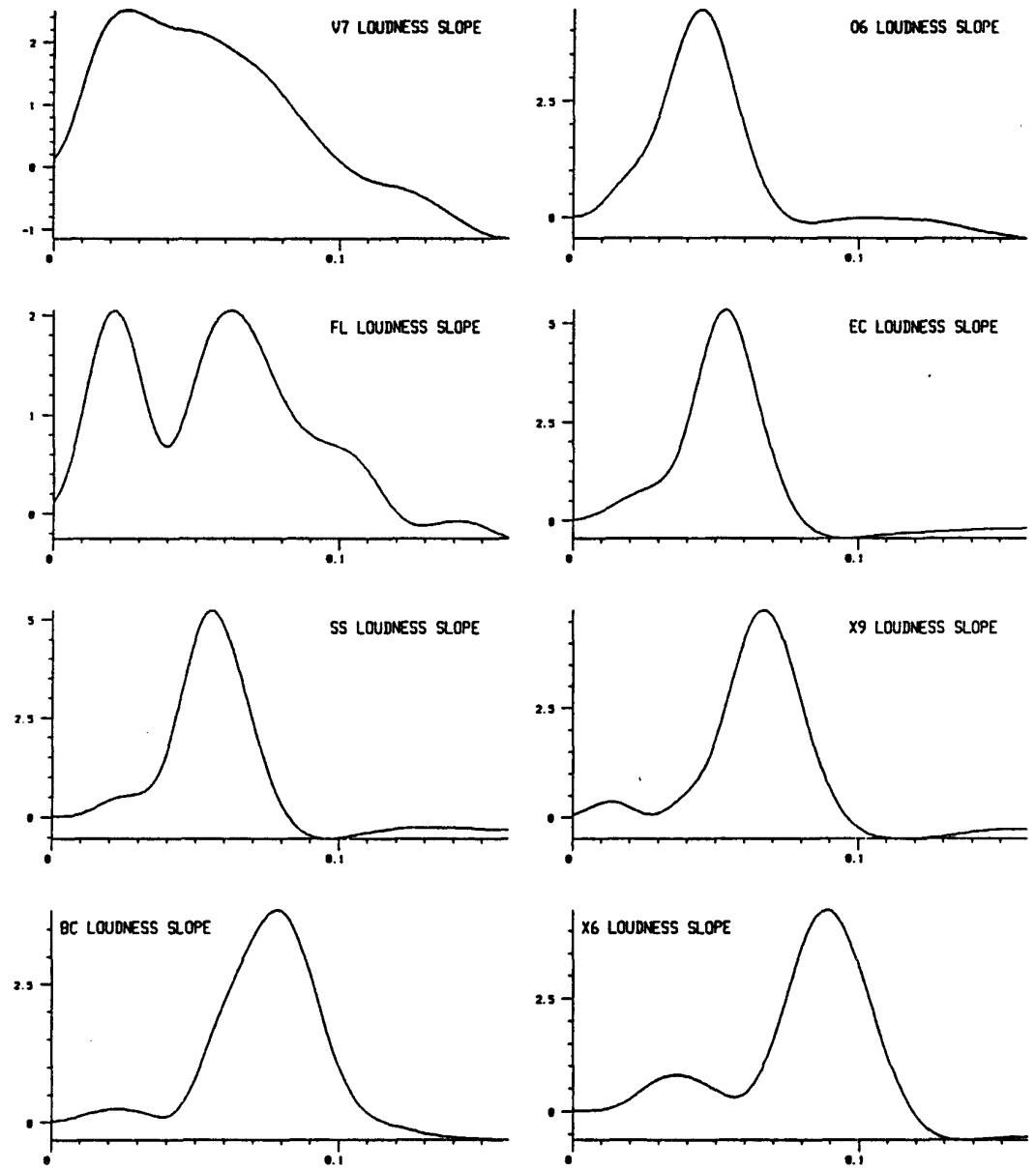

Figure 4.19 *Part 2.* **The remaining 8 out of 16 loudness slope functions are shown above.**

displayed in Figure 4.19 are more uniform in shape, though the two peaks for instrument FL are distinctive.) Also of interest is the high initial value of the slope for instrument FH (Figure 4.17); this is due to the "blip" in the attack of the French horn tone (see Figure 3.4).

Before developing a PAT prediction model based on slope threshold, we should consider the range of values exhibited by any set of slope functions. From Figure 4.18, for instance, we see that the maximum power slope for V3 is less than one-twelfth the maximum for EH or SS. A difference this large implies that the slope functions need to be normalized prior to model testing (however, the range of values for the loudness slope functions is not as extreme as those for the other two sets). There are two different ways to normalize: One way is to calculate the slopes from the envelope functions, and then normalize the maxima of the *slope* functions to the same value; the second way is to normalize the maxima of all the envelope functions to the same value prior to calculating the slope (this would not result in a significant difference for the set of loudness slope functions, since maximum loudnesses for the 16 stimuli were already essentially equal). Method one thus differentiates and then normalizes; method two normalizes and then differentiates.

Let us compare the results of these two normalization procedures. Consider the unnormalized set of (amplitude) envelope functions and the respective set of slope functions. Call the maximum of some instrument's envelope function M, and the maximum of that instrument's slope function X. If x (or $x(t)$) represents the envelope function, then the normalized envelope function would be x/M (normalizing the maximum of x/M to 1.0), and the slope of this function would become dx/M; this then would be the result of the second normalization procedure. The first procedure would take the slope of x, or dx, and then normalize *that*,—so that *its* maximum is 1.0,—resulting in dx/X. Since M and X are uncorrelated, the two procedures are clearly not equivalent. Therefore, both should be tried. Notice that both normalizations can be accomplished using the set of slope functions (corresponding to dx functions) displayed in Figures 4.17, 4.18, and 4.19; in the first case, the functions are simply scaled by $1/X$, whereas in the second case they are scaled by $1/M$ (X and M being different for each instrument).

As was done for the SLP model, we may wish to incorporate the influence of rise time into our new PAT models; but this means we must develop an accurate way to measure rise time. It is

convenient to define the *beginning* of an instrument's rise as the time slope threshold is exceeded, which would be the same time as predicted PAT for models that do not allow for rise time influence. We can then define the *end* of an instrument's rise as the time slope goes back below the slope threshold. Thus, rise time is defined as the duration between rise beginning and rise ending.*

For each of the following models, there will be two basic methods for predicting PAT. The first will simply predict an instrument's PAT as the time that instrument's slope function (either dx/M or dx/X) crosses some threshold (or equivalently as the time the rise begins). The second method will add a percentage of rise time to the time the rise begins and use the sum as predicted PAT. All three sets of slope functions (amplitude, power, and loudness) were used as inputs to each model tested; however, since the power slope functions always yielded better fits than the amplitude slope functions, only results for power slope and loudness slope functions will be presented below.

- *Normalization Type 1: Normalized Slope (RIS, RIT)*

As mentioned above, the first normalization method differentiates the envelope functions and then normalizes the results by scaling the slope functions by $1/X$, where X is each slope function's maximum value. PAT is then predicted as the time the normalized slope function crosses some absolute slope threshold. Applying an absolute threshold to normalized slope functions, however, is equivalent to applying a relative threshold (relative to the function's maximum value) to unnormalized slope functions. Therefore, the threshold for the RIS and RIT models will be expressed as a percentage of maximum slope.

The RIS model predicts PAT according to method 1; that is, it is a simple threshold model that does not include rise time as a factor. Correlation and linearity tests indicated that optimum values for the threshold were 1.04% of maximum slope for the power slope functions and 17.85% of maximum slope for the loudness slope functions. RPAT/APAT correlations with the threshold parameters fixed to these values are plotted in Figure 4.20. Both graphs exhibit a clustering of

*We are assuming that slope functions are monotonically increasing between 0 and maximum. Actually, this assumption is incorrect, as is evident from the "humps" preceding the large peak in some of the slope functions graphed in Figures 4.17, 4.18, and 4.19. However, though it will prove necessary to modify the models later to account for these "humps," there are relatively few of them, and they won't greatly affect the results of the PAT models now being developed.

predicted PATs near 0 and at least one PAT several milliseconds away from the desired line. The power slope functions do better than the loudness slope functions.

The RIT model predicts PAT according to the second method, which includes rise time as an additional factor. This added dimension resulted in an improvement over the RIS model,

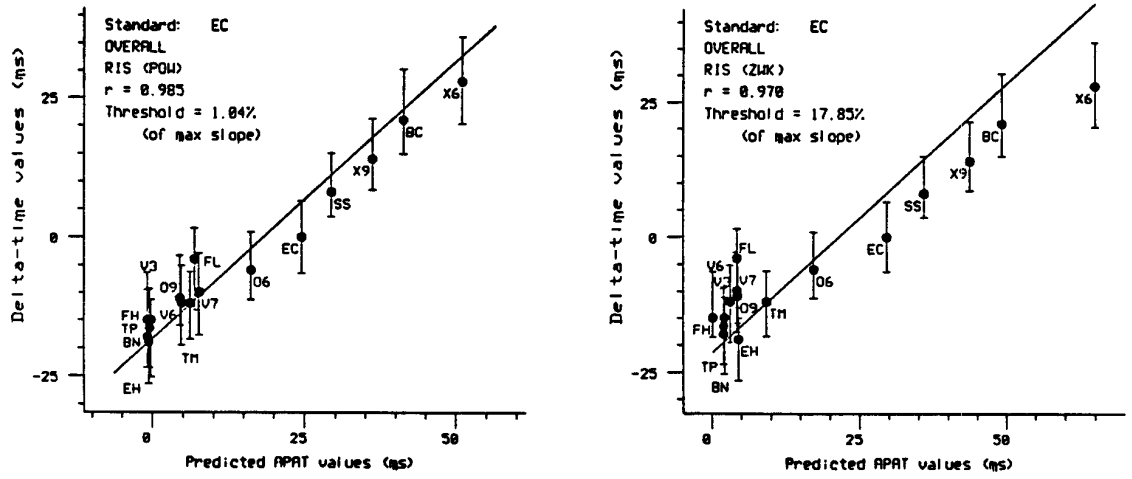

Figure 4.20 Overall RPAT values correlated with APAT values predicted by the RIS model. The threshold parameter is expressed as a percentage of maximum slope. Shown are correlations for both power slope (left graph) and loudness slope (right graph) functions.

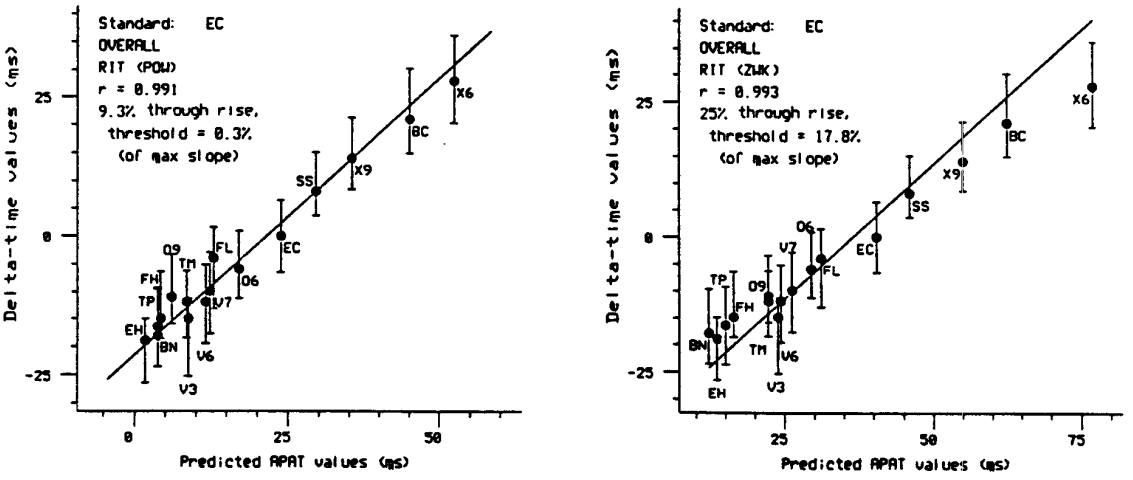

Figure 4.21 RPATs correlated with APATs predicted by the RIT model. Two parameters are inherent to the model, one for relative threshold (as in the RIS model) and one for rise time influence (percentage of rise time). Examples are given for both power slope (left graph) and loudness slope (right graph) functions.

especially for the loudness slope functions; correlations are shown in Figure 4.21. These correlations are exceptionally good; of concern, however, are certain deviations from the desired line, such as instrument FL in the left graph and instrument X6 in the right graph.

- *Normalization Type 2: Slope of Normalized Envelope (NAS, NAT)*

The NAS and NAT models correspond to the RIS and RIT models, respectively, except that in the NAS and NAT cases the second method for normalization is used (normalization of envelope function followed by differentiation). As mentioned above, this method of normalization is equivalent to scaling the original slope functions by $1/M$. Thus, the NAS and NAT models set rise beginning as the time the dx/M function crosses some absolute threshold τ—or as the time dx crosses $\tau \times M$.

Results for these two models were virtually identical to those for the RIS and RIT models. As in previous cases, results were improved with the added dimension of rise-time percentage; that is, the NAT model was an improvement over the NAS model. The best values of the two parameters for the NAT model applied to the power slope functions were 0.1245×10^{-3} (τ) and 6% of rise time. The resulting RPAT/APAT correlation is shown in the left graph of Figure 4.22. The similarity of this graph to the left graph of Figure 4.21 (including the correlation coefficient of .991) is striking.

- *Normalized First Moment Model (MNS, MNT)*

It was hypothesized that PAT might be influenced by a combination of envelope threshold and slope threshold. Thus, a model was developed based on the product of envelope and slope ($x\,dx$), or first moment. The MNS version did not include a rise time parameter (analogous to RIS and NAS), whereas the MNT version did. Results were no better than the RIT and NAT models, and will not be shown.

- *Refinements of the NAT Model*

APAT predictions according the NAT, RIT, and MNT models are all excellent, though there are some small deviations from the ideal APAT set. For instance, as seen from the left graph of Figure 4.22, instruments FL and O9 are predicted ~5 ms too early. The reasons for these deviations will be discussed below; of even more concern than the APATs for instruments FL and O9, though,

are the APATs predicted so early that the values contradict auditory theory.

The contradiction becomes apparent by examining the set of APAT-rise time pairs listed in the first column of Table IV.5. The formula used to calculate each APAT in this column was 6% of rise time added to the time the power slope function crossed the threshold of $0.1245 \times 10^{-3} \times M$ (M being the maximum value of the power envelope function). Notice that for instruments EH, BN, TP, FH, V3, and O9, 6% of rise time exceeds the corresponding predicted APAT. That is, the NAT formula determined that the slope function for these six instruments crossed the threshold prior to physical onset!

The reason for this is that the slope functions for these instruments were greater than the slope threshold at physical onset; the formula uses linear interpolation to estimate the true time of threshold crossing, and hence for these six instruments the formula interpolated between the times -1 and 0 ms (since the threshold was already exceeded at time 0). The NAT formula thus displays an artifact when applied to instruments whose amplitudes (powers) rise simultaneously with, or shortly after, physical onset. This implies some modification to the formula is necessary to account for this special case. (Such a modification will be discussed below.)

Further examination of the first column of Table IV.5 reveals that all sixteen APAT values are

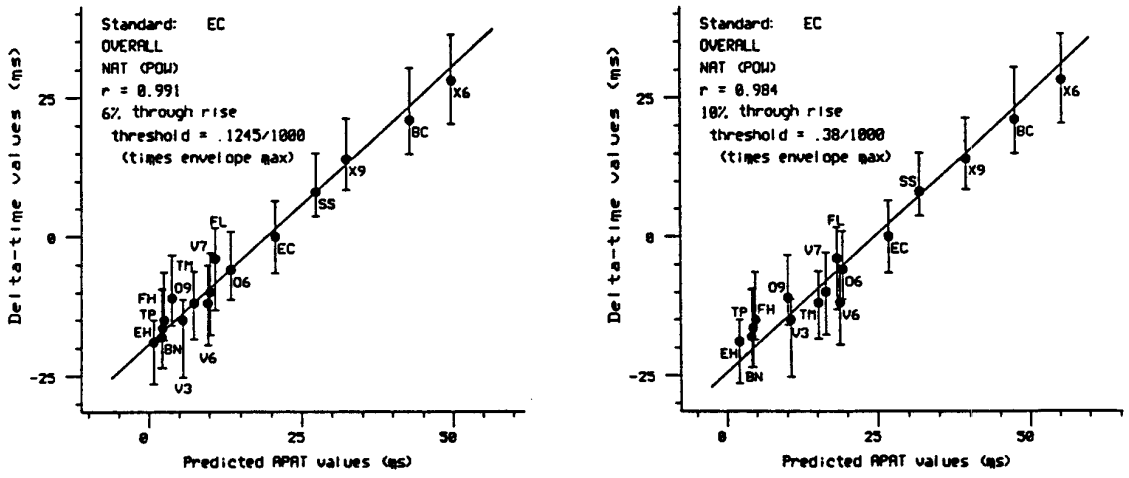

Figure 4.22 RPATs correlated with APATs predicted by the NAT model. Two parameters are inherent to the model, one for relative threshold and one for rise time influence. Both graphs are for power slope functions, but with two different pairs of settings for the two parameters.

Table IV.5 Three sets of APAT values—and corresponding rise times—predicted by the NAT model. In Case 1, the rise threshold parameter was set to .1245/1000 and the rise time percentage parameter was set to 6% of rise time. In Case 2, the respective parameter settings were .38/1000 and 10%. In Case 3, the respective settings were .36/1000 and 8%, but two additional modifications were made to the NAT model; see text for further details.

Instrument	Case 1 (r = .991) APAT (Rise Time)	Case 2 (r = .984) APAT (Rise Time)	Case 3 (r = .995) APAT (Rise Time)
EH	0.7 (27.7)	1.9 (27.5)	5.0 (23.6)
BN	2.0 (50.3)	4.0 (49.5)	6.7 (45.6)
TP	2.1 (49.4)	4.3 (48.8)	7.0 (44.8)
FH	2.4 (56.1)	4.7 (55.9)	7.2 (51.9)
V3	5.6 (102.4)	10.7 (101.2)	10.4 (99.3)
TM	7.4 (77.5)	15.2 (71.9)	13.5 (72.2)
V6	9.7 (115.7)	18.8 (110.5)	16.3 (110.8)
O9	3.8 (72.2)	10.2 (68.0)	11.3 (65.4)
V7	10.0 (87.7)	16.4 (84.2)	14.5 (84.4)
O6	13.5 (41.9)	19.2 (37.3)	18.4 (37.4)
FL	10.9 (100.0)	18.2 (96.2)	21.3 (90.8)
EC	20.7 (44.3)	26.7 (39.3)	25.8 (39.6)
SS	27.4 (41.0)	31.8 (37.8)	30.9 (38.0)
X9	32.4 (63.4)	39.5 (57.8)	38.0 (58.2)
BC	42.7 (69.1)	47.6 (66.4)	46.1 (66.6)
X6	49.7 (55.0)	55.4 (50.9)	54.3 (51.0)

less than might be expected according to the theory behind the NAT model. The time of real increase in power slope, determined visually from Figure 4.18, seems to be later than the corresponding NAT-predicted APAT in almost every circumstance. The high correlation (.991) between this set of APATs and the overall set of RPATs thus appears to be almost coincidental. In other words, correlation/linearity tests, similar to those done for models discussed earlier, indicated that the particular pair of parameter settings used with the NAT model was optimum; however, the theory behind the NAT model indicates that the setting of the slope threshold parameter should be greater.

The right graph of Figure 4.22 illustrates the RPAT-APAT correlation obtained by changing the two NAT-model thresholds to $0.38 \times 10^{-3} \times M$ and 10% of rise time. APAT values thus predicted, along with corresponding rise times, are listed in the second column of Table IV.5. The new APATs for instruments V3, TM, V7, O6, EC, SS, X9, BC, and X6 are much more in line

with theory; V6's APAT appears to be too late, but this is probably not critical since, as was mentioned earlier in this chapter, the string tones seem to be accommodated by a range of PAT values. The remaining six instruments have APATs predicted too early. The higher slope threshold is still exceeded at onset by four of these, EH, BN, TP, and FH, implying a modification to the NAT formula for this special case is still necessary. (It can be seen from Table IV.5 that 10% of rise time for these four instruments is greater than the corresponding APAT in each case.) The other two remaining APATs, however, those for instruments FL and O9, were predicted too early by the NAT model for different reasons.

Neither of the power slope functions for these two instruments increases strictly monotonically from the initial value at onset to maximum slope. Because of this, there is a local maximum in each function whose value is in the vicinity of the slope threshold used in the NAT model. These local maxima can be seen in Figure 4.18, though the one for instrument FL is more pronounced than the one for O9. (Instrument FH also exhibits a local maximum, but its value is much greater than the threshold.) Were the local maxima for FL and O9 removed, the threshold crossings would occur much later, resulting in larger predicted APATs for these two instruments. Similarly, if the slope threshold were gradually raised, there would be a sudden increase in APAT at the point threshold just exceeded the local maximum. It seems appropriate therefore to modify the model slightly to account for this phenomenon; such modification is presented in detail in Appendix A.

Now let us consider the predicted APATs for instruments EH, BN, TP, and FH, whose slopes exceed slope threshold at time of physical onset. As mentioned above, the negative APATs for these instruments (prior to the addition of a percentage of rise time) are due to the near-impulsive nature of their attacks. This raises a question about the applicability of a slope-threshold model to such tones. Without audible onset activity at least 5–10 ms prior to attack, there is nothing to signal the ear of the impending attack; yet such activity is implicitly inherent in the theory behind the NAT model. In short, for cases in which slope threshold is exceeded at the time of physical onset, the model breaks down. It is not difficult however to extend the model to account for instruments with impulsive attacks; essentially what is involved is a simple delay of the predicted APAT value. More detail can be found in Appendix B.

The overall set of RPATs was again correlated with the APATs predicted by the NAT model after making modifications according to the formulas of Appendices A and B. The weight parameter (Appendix A) was set to 0.75; the maximum shift amount (Appendix B) was set to 4 ms. Results are shown in Figure 4.23; APAT and rise time values are listed in the third column of Table IV.5.

The fit of predicted values to the data is remarkably good ($r = .995$). The NAT model is also consistent with auditory theory and is practical to apply to recorded instrument tones. Our three criteria for judging PAT prediction models have therefore been met, and it seems justified therefore to answer Q2 with the statement that PAT can indeed be accurately predicted.

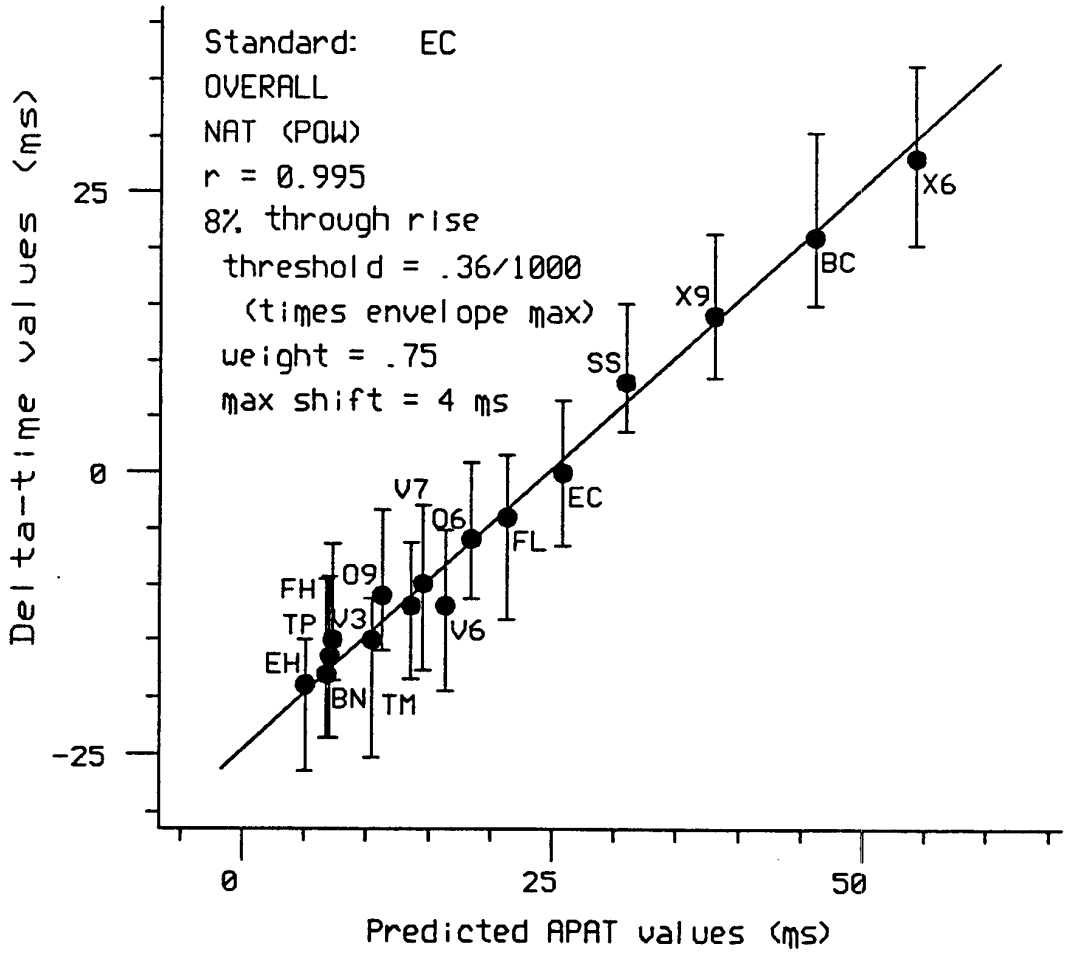

Figure 4.23 RPAT (Δt) values correlated with APATs predicted by the NAT model after all modifications to the model have been made. The settings for the original two parameters are 8% of rise time and a slope threshold of $0.36 \times 10^{-3} \times M$. The weight and shift parameters (0.75 and 4 ms, respectively) are due to modifications discussed in detail in Appendices A and B.

Chapter V

Model Verification

We have seen that the sets of RPAT measurements obtained from Experiments I, II, and III all differed from each other. For many of the sixteen instruments, there were only slight differences (2–3 ms) in Δt values among all the various RPAT sets. For other instruments, however, RPATs varied by as much as 20 ms (after values were adjusted to be relative to the same standard). The assumption was made that the large variations in RPAT for this class of instruments were due to side effects (e.g., masking and fusion) from certain experimental paradigms, and that PAT is independent of presentation conditions. Based on that assumption, a single overall set of Δt's was compiled (listed in the rightmost column of Table IV.4) for the purpose of testing the accuracy of various PAT prediction models. Each Δt in this set was derived from statistical evidence, and was presumed to best represent RPAT for the corresponding instrument tone.

The set of APAT values predicted by the NAT model was seen to correlate almost perfectly with the compiled set of RPAT values. However, this high correlation alone does not confirm the model's validity, since the accuracy of each Δt value in the compiled RPAT set has not been established. Therefore, we need to verify whether or not the assumptions behind the derivation of the RPAT values are correct before drawing final conclusions regarding the NAT model. An alternative approach, however, is to verify that the *predicted* APAT values correspond to *actual* APAT values.

Two separate verification procedures were carried out, both of which will be discussed in this chapter. The first procedure involved the synthesis of simple musical examples, the rhythms of which were constructed according to the APAT values predicted by the NAT model; subjects were asked informally whether or not the rhythms sounded regular. More detail and results will be presented in the following section. A separate section will be devoted to the second verification procedure, which involved applying the NAT model to the stimuli used by Vos and Rasch; the APAT values thus predicted by the NAT model were correlated with the Δt values measured by the investigators. As can be seen from the following sections, both verification procedures confirmed the validity of the NAT model and its applicability to certain practical situations.

5.1 Synthesis of Regular Rhythms

In determining the validity of any acoustical theory, the discriminating listener is the best judge—auditory confirmation is the best confirmation. The theory behind the NAT model presented in Chapter IV is that the perceptual time of attack for a recorded instrument tone can be accurately predicted by applying a formula to the time-varying slope of that tone's power envelope. The theory thus implies that a rhythmically regular, or isochronous, musical phrase can be constructed from individually recorded tones with different attack characteristics, provided that the predicted PATs for the tones are used to determine their rhythmic arrangement. Auditory confirmation or contradiction of the theory can be obtained by asking musically trained listeners to judge whether or not the musical phrases thus constructed sound rhythmically regular.

To test the validity of the NAT model, an informal study was conducted using the method just described. However, in order to ensure a thorough test, it was necessary to take into consideration some design factors,—even for an informal study. These factors can be categorized as follows:

- *Paradigm.* There are many experimental paradigms from which to choose, but only two will be considered here. One possibility is to present a melodic sequence of two or more instrument tones, asking the subject to indicate whether or not he perceives the rhythm as being

regular. This would test only isochronous relationships among stimulus tones, but synchronous relationships could be tested with a variational paradigm that presented two sequences (or even single tones) simultaneously. A problem with either variation of this paradigm is that the subject has no comparison sequence upon which to base his judgment; thus, rhythms with slight but audible irregularities (or tones with slight asynchronies) could conceivably be judged as regular (synchronous).

Another paradigm to be considered is the sequential (as opposed to simultaneous) presentation of two or three melodic sequences, with the subject forced to choose the most rhythmically regular sequence among them. Again, this would test isochronous relationships among stimulus tones, but tests of synchrony could be obtained by using two-voice sequences rather than monophonic melodies. This paradigm allows the subject to make a comparative judgment, but a forced-choice response does not indicate whether the most regular sequence is indeed perceived as regular or whether it is simply perceived as less irregular than the other sequences.

- *Auditory Temporal Acuity.* Perception of regularity and perception of synchrony are subject to certain auditory limits. Lunney found that discrimination of regularity is limited to $\sim$2% of the beat period [*Lunney* (1974)], whereas the findings of Experiment I have indicated that the limit is closer to 1% of beat period for musically trained subjects. Thus, the rate at which successive instrument tones of a melodic sequence are presented should give a rough indication of how large a deviation from perceptual regularity we can expect a subject to perceive.

It is useful at this point to introduce *physical onset asynchrony* (POA), to be defined operationally as follows: Consider a set of instrument tones arranged rhythmically in a physically isochronous sequence; depending on the relative PATs between adjacent tones, the sequence may or may not be perceptually isochronous. To ensure perceptual isochronism, the physical onset of each tone in the original sequence has to be adjusted by an amount corresponding to the RPAT between it and the first tone in the sequence; the amount of this adjustment is defined as POA. Thus, the POA for each tone in a physically isochronous sequence will be 0; the POA for a tone in a perceptually isochronous sequence will be 0 only if that tone's APAT is identical to the APAT of the first tone in the sequence.

If each POA in a perceptually isochronous sequence is smaller than the limit on the discrimina-

tion of regularity, then the same sequence with each POA set to 0 (i.e., physically isochronous) should also be perceived as isochronous. In other words, we would not expect a subject to be able to perceive a difference in the rhythmical regularity of these two sequences. On the other hand, if a perceptually isochronous sequence comprising tones with relatively large POAs (implying large RPATs between tones) was compared against its physically isochronous counterpart, we would expect the subject to perceive a difference in regularity. Thus, the particular selection and order of stimulus tones that make up a sequence should not be made arbitrarily; rather, they should be carefully determined according to the specific hypothesis being tested, taking into consideration either the measured RPATs or the predicted APATs among the tones.

- *Precision of Testing.* Although understanding the limits in the auditory acuity of subjects will help to a large extent in determining the stimulus makeup of sequences used for testing, there will still be innumerable permutations and combinations of instrument tones possible. There are essentially two approaches to take in selecting which instrument tones to use in constructing the stimulus sequences: (1) a general aural evaluation of all the predicted APAT values, accomplished by combining many different instrument tones into each sequence; (2) specific individual examination of each APAT, accomplished by limiting the elements of each sequence to one or more copies of each of two different tones. Thus, the form of the first kind of sequence would be similar to **A-B-C-D-E**, whereas the second type of sequence would have the form **A-B-A-B-A**.

The second approach would of course require many more trials than the first approach. However, it would allow for more precise testing of certain predicted APATs whose values may be in question. For instance, it was hypothesized earlier that for the bowed string tones and perhaps the flute tone, any APAT within a range of values would be satisfactory to subjects; this hypothesis could be tested directly by using the second kind of sequence and controlling the particular APAT value used.

Now let us review the actual choices made in designing the informal study. The second type of paradigm, in which the subject compares two or more sequences for each trial, was used exclusively. The subject was asked to indicate not only which sequence appeared to be closer (or closest) to strict rhythmic regularity, but also if both (or all) sequences seemed to be regular or both (all) seemed to

be irregular. Sequences were short monophonic melodic phrases, each phrase being either a 5-note ascending or descending scale, a 7-note major triad arpeggio (ascending then descending), or a 9-note scale-like melody. Thus judgments were of isochronous, not synchronous, rhythms. Since all original stimulus tones were pitched at the E-flat above middle-C, it was necessary to transpose copies of some tones to other pitches in order to realize the appropriate melodies. A tone was transposed by multiplying each harmonic's frequency function in the analyzed form of the tone by the same constant scale factor, and then realizing the sample-data form by means of the digital synthesizer. The transpositions were always within a fifth of the original E-flat pitch; i.e., they spanned an interval of a ninth, ranging from the A-flat below E-flat to the B-flat above.

The beat period, or time between adjacent (perceptual) attacks within each sequence, was set to 200 ms. It was assumed therefore that subjects should be able to perceive any deviations from perceptual regularity that were larger than 2–4 ms (see discussion above under *Auditory Temporal Acuity*). This theoretical limit on discriminability was a useful guide in determining which instrument tones to group into a single sequence. For instance, the predicted APATs for the first four instruments listed in Table IV.5 (EH, BN, TP, and FH) are all within $\sim$2 ms of each other; therefore, one would expect a subject to hear a sequence comprising these tones as isochronous regardless of whether the POA for each tone is set to zero or the appropriate APAT value. This hypothesis was tested by constructing comparative sequences of tones with POA values that differed by 6 ms or less. In addition, several other sequence pairs were constructed that used larger POA differences; subjects were expected to perceive differences in regularity for these pairs.

It was desired to obtain both a general evaluation of the success of the NAT model and a more specific verification of particular APAT predictions. Therefore, both approaches (see discussion above under *Precision of Testing*) were taken in selecting the particular instrument tones to use in constructing the various stimulus sequences. That is, some of the sequences were composed of 5–9 different instrument tones, and others were combinations of specific instrument pairs.

The particular sequence arrangements are listed in Table V.1. The order of trials was randomized for each subject; the numerical order indicated in the table is only for purposes of discussion later. The table is segmented according to sequence type (general groupings, pairs, or triads). The

Table V.1 Sequences of instrument tones used to test the APAT values predicted by the NAT model. Physical onset times of each tone within a sequence were adjusted according to values given in the POA column. Each RPAT listed in the rightmost column is the RPAT of the second instrument relative to the first instrument, or the predicted APAT of the second tone minus the predicted APAT of the first tone.

GENERAL GROUPINGS

Trial	Instrument Sequence	POA	RPAT
1	TP-BN-V3-EH-FH	0 *vs.* $0 < t \leq$ 5.4 ms	
2	TM-V6-V3-V7-O9	0 *vs.* $0 < t \leq$ 5.9 ms	
3	SS-O6-BC-EC-X6-FL-X9	0 *vs.* $12.5 \leq t \leq$ 33.0 ms	
4	O9-X9-TM-BC-V6-X6-V7	0 *vs.* $24.5 \leq t \leq$ 39.8 ms	
5	V3-SS-FH-EC-TP-FL-BN-O6-EH	0 *vs.* $11.7 \leq t \leq$ 23.7 ms	

SEQUENCE PAIRS

Trial	Instrument Sequence	POA	RPAT
6	SS-X6 [7-note]	23.4 *vs.* 14.0 ms	23.4 ms
7	BN-TM [7-note]	10.8 *vs.* 4.7 ms	6.8 ms
8	EH-FL [9-note]	15.0 *vs.* 24.0 ms	16.3 ms
9	V3-O6 [9-note]	8.0 *vs.* −2.0 ms	8.0 ms

SEQUENCE TRIADS

Trial	Instrument Sequence	POA	RPAT
10	V7-EC [7-note]	11.3 *vs.* 5.3 *vs.* −0.7 ms	11.3 ms
11	V6-EC [9-note]	12.5 *vs.* 6.5 *vs.* 0.5 ms	9.5 ms
12	O9-EC [7-note]	14.5 *vs.* 5.5 *vs.* −3.5 ms	14.5 ms
13	O9-EC [9-note]	14.5 *vs.* 5.5 *vs.* −3.5 ms	14.5 ms

particular POA settings for each sequence in a trial are listed, along with RPAT differences (for specific instrument pairs only) predicted by the NAT model. These RPATs are simply differences between the respective APATs listed in Table IV.5.

In addition to the sequences listed in Table V.1, a two-voice sequence was constructed that consisted of a Mozart melody (familiarly known as "Twinkle, Twinkle, Little Star") and accompanying counterpoint. Thus both isochronous and synchronous judgments were made in evaluating the sequence's rhythmic regularity. The beat rate was set to 325 ms, and all 16 instrument tones were represented five times each, for a total of 80 notes. Both melodic instrument pairs (horizontally

adjacent tones) and harmonic pairs (simultaneous tones) were determined so that the relative RPAT between any pair was usually at least 10–15 ms. Both a physically isochronous (and synchronous) version (all POAs equal to 0) and a (theoretically) perceptually isochronous version (POAs set according to appropriate RPATs) were presented to the subject.

- *Results*

Eight subjects participated in the informal study. Results can be summarized as follows:

- Subjects could not distinguish between sequences whose respective POAs were less than 6 ms from each other (trials 1, 2, and 7). In other words, responses were at a chance level. In all cases, each sequence was perceived as isochronous.

- When respective POAs between sequences were greater than 9–10 ms, subjects in general showed a clear preference for POA settings determined according to predicted APAT (response was $> 75\%$). This finding applies to trials 3, 4, 5, 6, and 9. An exception to this finding, however, was trial 8, involving instruments FL and EH; responses in this case were at a chance level. Since both sequences in the pair were perceived as isochronous, we can assume that subjects are satisfied with APATs for the FL tone that fall between 21.3 and 30.3 ms (see Table IV.5).

- For triadic comparisons (trials 10–13), there was a slight tendency for subjects to prefer POA settings corresponding to APAT values predicted by the NAT model (and corresponding to the rightmost column of Table V.1); for the most part, however, subjects could not really distinguish among the three sequences. This implies that we can draw the same conclusion regarding the string tones that was just drawn for the FL tone, as was hypothesized earlier. That is, a range of times, rather than a single moment, seems to be appropriate for predicting APAT for these tones. This also seems to be true for the O9 tone (trials 12 and 13), which implies that subjects are confused by the local maximum phenomenon this tone exhibits (see the end of Chapter IV or Appendix A). Also, confusion from the local maximum in the flute tone perhaps explains why responses to trial 8 were at a chance level.

- Some sequence combinations produced 93–100% response in favor of APAT values predicted by the NAT model. These were trials 3, 4, 5, and 6, and the trial involving the Mozart melody.

In general, then, it appears that the APAT values predicted by the NAT model for all 16

instrument tones are indeed correct. Melodies constructed by concatenating these tones can be made perceptually isochronous by adjusting physical onset times (POAs) by appropriate APAT values.

5.2 NAT Model Applied to Vos-Rasch Data

The five stimulus envelopes used in the Vos and Rasch experiments were deterministic [*Vos and Rasch* (1981)], the formula used to generate them being

$$R(t) = \frac{1}{2} + \frac{1}{2}\sin[(\frac{t}{\rho} - \frac{1}{2})\pi],$$

where $R(t)$ is the function describing the rise portion of the amplitude envelope, t is time, and ρ is a parameter representing rise time as measured from physical onset to maximum amplitude. In describing the rise functions, Vos and Rasch used a different variable, ρ', to represent rise time as measured from 10% of maximum amplitude to 90%; the two variables were related then by a multiplicative constant: $\rho = 1.69\rho'$. The authors chose, for their particular experiments, five settings of the rise time parameter: $\rho' = 5$, 20, 40, 60, and 80 ms. Thus, the five values of ρ were 8.45, 33.8, 67.6, 101.4, and 135.2 ms.

For our purposes, it was necessary to generate sample-data sequences that corresponded to these five functions; this was done by sampling function values every T = 1 ms. These sample-data sequences, then, being normalized in amplitude on a scale of 0 to 1, were completely analogous to the set of 16 amplitude functions displayed in Figure 3.4. Power slope functions for the five envelope sequences were obtained by precisely the same method described in Chapter IV, and these were input to the NAT model to obtain predicted APAT values; these values were 4.8, 7.7, 15.9, 23.9, and 32.0 ms, respectively.

Vos and Rasch obtained RPAT measurements for all possible stimulus pairs. That is, there was an RPAT value for each envelope relative to every other envelope. Unfortunately, these values were not entirely consistent, and it was therefore difficult to correlate them with the predicted APAT

Table V.2 RPAT relationships among the five stimuli used in Vos and Rasch's first experiment. The upper right grid is a modification of the upper left grid; this modification was necessary to obtain consistent transitive properties. The lower two grids represent predicted RPATs according to the NAT model with two different settings for the slope threshold. All values are in milliseconds.

Measured RPATs

Rise Time	80	60	40	20
5	35	22	17	6
20	30	20	8	
40	20	10		
60	9			

Adjusted RPATs

Rise Time	80	60	40	20
5	35	25	15	6
20	29	19	9	
40	20	10		
60	10			

Predicted APATs
Threshold = $.36 \times 10^{-3}$

Rise Time	80	60	40	20
5	27.2	19.1	11.1	2.9
20	24.3	16.2	8.2	
40	16.1	8.0		
60	8.1			

Predicted APATs
Threshold = 1.1×10^{-3}

Rise Time	80	60	40	20
5	35.7	25.3	15.4	5.6
20	30.1	19.7	9.8	
40	20.3	9.9		
60	10.4			

values. The inconsistency can be more easily seen by referring to Table V.2. In the upper left grid are displayed all the measured RPATs from Vos and Rasch's first experiment, rounded to the nearest millisecond. It is assumed that these values should obey a transitive rule; for instance, the RPAT for the 5-60 envelope pair (22 ms) should equal the sum of the RPATs for the 5-20 pair (6 ms) and the 20-60 pair (20 ms), which is 26 ms, or the sum of the RPATs for the 5-40 pair (17 ms) and the 40-60 pair (10 ms), which is 27 ms.

In order to make a meaningful comparison between the predicted and measured values, the inconsistency in RPAT measurements had to be resolved. It was decided to accomplish this by adjusting some of the entries in the RPAT grid; however, it was difficult to ascertain which values to adjust. By careful analysis of all possible combinations (similar to the reasoning illustrated at the end of the previous paragraph), it was determined that values should be adjusted to those listed in the upper right grid of Table V.2; that is, the 5-60 RPAT was increased by 3 ms, the 5-40 RPAT was decreased by 2 ms, the 20-40 and 60-80 RPATs were each increased by 1 ms, and the 20-60 and 20-80 RPATs were each decreased by 1 ms. For this set of adjusted values, the transitive rule holds

in every circumstance.

The lower left grid of Table V.2 shows RPAT relationships that were derived from the five predicted APAT values. By comparing this grid with the upper right grid, it can be seen that most entries differ by several milliseconds, though some are nearly equal. However, the Vos-Rasch stimuli were presented at an intensity level of 82 dB_A, whereas the level used for Experiments I, II, and III was 90 dB_A. Therefore, it was reasoned that the NAT model parameters might be dependent on presentation intensity level, and that other APAT predictions should be obtained by allowing these parameters to vary. It seemed that the slope threshold parameter would be more affected by intensity level than the percent-of-rise-time parameter (see the theoretical discussion below); hence, the rise time parameter was kept at 8%, but the slope threshold was allowed to increase. Also, the maximum shift for impulsive attacks was reduced from 4 ms to 3 ms.

With the slope threshold (τ) set to 1.1×10^{-3}, the five predicted APATs were 3.9, 9.5, 19.3, 29.2, and 39.6 ms. These values resulted in the RPAT relationships shown in the lower right grid of Table V.2. The similarity of this grid with the upper right grid is striking; all predicted relationships are within 1.1 ms of the corresponding measured RPATs.

Vos and Rasch repeated their experiment three times, each time with a different intensity level; that is, they measured changes in RPAT relationships as intensity level varied. The levels used were 77 dB_A, 57 dB_A, and 37 dB_A. Only three different rise times were used in these experiments ($\rho' = 5$, 40, and 80 ms), but RPATs were measured for both possible orders of each combination pair. Thus, there were two RPATs for each combination. The upper grids of Table V.3 show the measured RPAT values for each of the three intensity levels. The lower grids are corresponding predicted RPATs and the respective slope thresholds used to obtain these values from the NAT model. Again, the correlation of values in the upper three grids with corresponding values in the lower grids is excellent.

Discussion

The NAT model has been shown to yield accurate APAT predictions not only for the 16 stimuli used in Experiments I, II, and III, but also for the standardized envelopes employed by Vos and

Table V.3 RPAT relationships among the three stimuli used in Vos and Rasch's second experiment. The upper grids represent measured values, in milliseconds, for three different intensity levels. The lower grids represent corresponding RPATs predicted according to the NAT model with three different settings for the slope threshold.

Measured RPATs
Level = 77 dB_A

Rise Time	80	40
5	37–41	17–21
40	19–24	

Measured RPATs
Level = 57 dB_A

Rise Time	80	40
5	42–46	19–21
40	21–23	

Measured RPATs
Level = 37 dB_A

Rise Time	80	40
5	53–58	24–28
40	27–31	

Predicted APATs
Threshold = 1.55×10^{-3}

Rise Time	80	40
5	39.2	17.2
40	22.0	

Predicted APATs
Threshold = 2.15×10^{-3}

Rise Time	80	40
5	43.6	19.6
40	24.0	

Predicted APATs
Threshold = 5.2×10^{-3}

Rise Time	80	40
5	55.2	23.9
40	31.3	

Rasch. However, in order to attain success at varying intensity levels, the slope threshold parameter had to be allowed to vary; i.e., this parameter could not be set to a constant.

A slope threshold dependent on intensity level is not necessarily inconsistent with auditory theory, especially if the variation is relatively small compared to the change in intensity.* Two lines of reasoning support this statement. First, as the intensity level of a tone is lowered, more and more of its onset activity preceding perceived attack becomes inaudible. If slope threshold were to remain constant under such circumstances, we might find hearing threshold being crossed after slope threshold is crossed! In short, then, as the time of perceptual onset is delayed more and more from physical onset due to lowered intensity levels, we would expect PAT to increase accordingly. (See page 3 to review definitions of perceptual and physical onset.)

The second reasoning is as follows: A 10-dB drop in intensity corresponds to dividing the intensity by 10. (A 10-dB decrease also corresponds approximately to a halving of the loudness—see

*It might seem remarkable that the thresholds for the data in the Vos-Rasch study are as close as they are to the threshold value used to predict APATs for the stimuli used in Experiments I, II, and III. After all, the tones used in the experiments we have been studying were presented through a single speaker, whereas the Vos-Rasch tones were presented diotically through earphones; also, their stimuli were synthetic, homogeneous, and extremely brief (~150 ms), whereas the stimuli used in Experiments I, II, and III were recorded from real instruments, diverse in character, and ~350 ms in duration. These differences point out the robustness of the NAT model.

[*Stevens, S.S.* (1955)].) However, if the physical slope is expressed in terms of dB/ms, it remains constant regardless of the intensity level. But an increase from 30 dB_A to 90 dB_A in 10 ms (6 dB/ms) should have more perceptual impact than an increase from 0 dB_A to 60 dB_A in 10 ms (also 6 dB/ms), if only because loudness is not linearly related to the dB measurement near threshold level. Thus, it seems reasonable to expect slope threshold to increase as intensity is lowered, but by an amount proportionately less than the decrease in intensity.

To develop a formula expressing slope threshold as a function of intensity is beyond the scope of this dissertation, and should not be attempted without further empirical research. However, it is evident that such a dependency exists, and that the dependency does not detract from the applicability of, or theoretical support for, the NAT prediction model.

Chapter VI

Conclusions

This dissertation has been focused on the concept of PAT, or the time an instrumental tone's attack is perceived relative to the tone's physical onset. We have been concerned with how accurately PAT can be measured (Q1), how well it can be predicted (Q2), and how much variability in PAT there is among subjects (Q3). In order to better address these three issues, three separate experiments were run using 16 recorded instrument tones as stimuli. Based on results from these experiments, a PAT prediction model was developed, and the accuracy of this model was verified by two independent methods. This chapter will be devoted to a final review of Q1, Q2, and Q3, and will present some suggestions for future research.

Discussion of Measurement Accuracy (Q1)

When the RPAT-measuring paradigm involves isochronous judgments, as was the case in Experiment I, measurement accuracy will be directly related to the discriminability of regularity. This discrimination limit is $\sim$1–2% of the beat period, or time between successive tones. Thus, a faster beat rate (shorter beat period) should yield less absolute variation in RPAT measurements.

When synchronous judgments are involved, as was the case in Experiments II and III, RPAT measurement accuracy will depend to a large extent on whether or not paradigm side effects (e.g., fusion or masking) are avoided. If side effects are avoided, accuracy will be limited by the discrimination of temporal order, which appears to have a lower limit of 2 ms [*Patterson and Green*

(1970), *Green* (1973), *Wier and Green* (1975)]. Also, judging from the results of Experiment III, it is apparently difficult for subjects to synchronize the PATs of stimuli with very different attack characteristics, such as a drum slap and a bowed string tone. Perceived duration of attack is probably the critical feature; PATs of tones with equal perceptual durations are relatively easy to synchronize, whereas PATs of tones with grossly unequal perceptual durations are extremely difficult for subjects to synchronize.

There are several RPAT-measurement paradigms that have not been used, but would probably yield more accurate results. These should be considered for future research, and are listed as follows:

- *Isochronous stimuli,* **A-*n*·B**. The standard **A-B** paradigm exhibits a duple rhythm, but it may prove more advantageous to involve judgments of isochronism for triple, quadruple, or sextuple rhythms. A physically isochronous sequence of the same tone (in our case, **B**) repeated several times in succession will of course be perceptually isochronous, and would establish a standard for the subject. Therefore, an aberration from this standard, caused by an improperly adjusted **A**, should be more noticeable, or more obviously irregular, than an irregular duple rhythm.

- *Isochronous stimuli,* **A-B** *pattern randomized.* It has been mentioned that sometimes a subject can be somewhat hypnotized by a slightly irregular duple rhythm, in that he becomes convinced the rhythm is regular. One possible way to break the mesmeric pattern is to use the previously mentioned paradigm; another way is to randomize the presentation pattern. An example of a random pattern might be **A-B-B-B-A-B-A-A-B**.... The subject would not be able to predict which of the two tones he might hear at any particular moment, and therefore would be forced to concentrate more on the isochronism of the rhythm.

- *Isochronous stimuli, all possible* **A-B** *pairs.* RPAT measurements are obtained by comparing tones against an established standard. As was seen from the results of Experiment II,—even though simultaneous stimuli were used rather than isochronous ones,—the choice of standard can greatly affect the set of RPAT values. A more thorough paradigm would not single out one stimulus as the standard, but would obtain RPAT measurements for all possible stimulus pairs. This would yield a matrix of values that could be checked for inner consistency, as was done for the Vos-Rasch data in Chapter V. Ideal as this paradigm is, however, it becomes more impractical as the number

of stimuli increases. The total number of combinations will be $n(n-1)/2$; the total number of trials will thus be $kn(n-1)/2$, where k is the number of replications. If five replications of all possible pairs of 16 stimuli were run, the total number of trials would be $5 \cdot 16 \cdot 15/2 = 600$ trials!

- *Simultaneous stimuli,* **A** *and* **B** *different in frequency or location.* As we saw in Chapter IV, fusion of stimuli confounded RPAT measurements of the simultaneously presented stimuli in Experiment II. The drum slap was used as the standard in Experiment III in an attempt to avoid the fusion problem. Another possible way to avoid fusion, however, is to use tone pairs that differ in frequency. Of course, the interval between tones should be chosen carefully; otherwise, fusion might not be eliminated. An octave or a fifth, for example, might not be any more satisfactory than a unison. A similar approach would be to present each tone from a separate loudspeaker; resulting phase differences might be enough to prevent fusion from occurring.

- *Isochronous stimuli, differing in amplitude.* A topic for future research is how PAT varies with amplitude. This issue was discussed in some detail in Chapter V, but no quantitative formula was derived. A paradigm that might be used in examining this question is one in which the stimuli differ in amplitude (**A** more intense than **B**, for example). However, it is likely that this paradigm would introduce side effects into the measurement procedure; a sense of rhythmic accent would undoubtedly influence judgments of isochronism [*Fraisse* (1978)].

Aside from paradigm, there are other factors that can limit the accuracy of measuring RPAT. One has already been mentioned, namely the difficulty subjects have in reconciling attacks whose rise times differ considerably. Another factor is that certain tones exhibit contrasting PAT cues. Examples in the experiments we have been studying are FL and O9. Subjects simply have difficulty in precisely determing PAT for these tones, regardless of whether they are presented isochronously or simultaneously with another tone. It is difficult to assess how subjects weight amplitude cues versus spectral or onset cues, but it is important at least to realize that PAT for some stimuli is not as clearly defined as it is for other tones.

It should also be mentioned that the tones used in Experiments I, II, and III are not necessarily representative of their producing instruments. All instrumentalists, especially string players, are able to produce any of the three standard attacks illustrated in Figure 4.11. Thus, one cannot

generalize that particular instruments will always display PATs corresponding to values presented in this dissertation.

Discussion of Prediction Accuracy (Q2)

To test the predictability of PAT, many different prediction models have been developed and analyzed. Almost all of the models predicted sets of APATs that yielded excellent correlation with measured RPATs, though some models resulted in higher correlations than others. This should not seem surprising, since for most of the tones used as stimuli in the three experiments, the rise portion of the attack was quick, implying that large changes in amplitude, integration, and slope occurred in a relatively brief period of time. That is, absolute amplitude, relative amplitude, integration, and slope thresholds were all crossed by the same instrument at nearly the same moment. Thus, all of the models' predicted APATs for each of these tones were within only a few milliseconds of each other. This is why so many correlation plots were similar in shape.

We see then why it was important not to be satisfied with the very good results that were obtained from many of the models, and to be exhaustive in our testing. The evidence for each model is really good enough to support the theory behind it, making it easy for one to draw incomplete or erroneous conclusions regarding any one model.

The best results (.995 correlation) were obtained from the modified NAT model. Furthermore, this model satisifies the three criteria we established for testing its validity, namely (1) prediction accuracy, (2) consistency with auditory theory, and (3) practicality of application. The model was extended to account for variations in intensity, though a more precise relationship between slope threshold and intensity is left to be worked out in the context of future research. Thus, although many models can be applied to the prediction of PAT with moderate to excellent success, the NAT model seems to be the one most consistently reliable over a broad range of instruments.

We have also seen that the accuracy of PAT prediction for tones with long rise times is not as critical as it is for tones with impulsive or near-impulsive attacks; a range of PAT values for these tones seems to be sufficient. However, if specific PAT values are desired for certain applications, the PATs predicted by the NAT model for tones with long rise times seem to be good choices.

Discussion of Subject Variance (Q3)

Though subject differences were significant in all three experiments, these differences were seen to be procedural rather than perceptual. In most cases, subject variance was within the expected limits of discrimination. One possible source of perceptual differences among subjects is the different weightings subjects may give to contrasting PAT cues, such as spectral shifts *vs.* amplitude shifts. However, there is no clear indication of this in the data from Experiments I, II, and III. Also, there seems to be close agreement among listeners regarding isochronism or synchrony in musical contexts.

Perhaps the most interesting and unexpected results in terms of subject performance are those from Experiment III, in which the drum slap was used as the standard. Though fusion was avoided in making judgments of simultaneity, subjects nonetheless found it very difficult to synchronize an instrumental tone with the drum sound. The reason for this difficulty was the inability of subjects to resolve attacks with very different rise characteristics, especially rise time. This might imply that the drum is a poor choice for its traditional rôle as "timekeeper" in ensemble performance. However, in rhythmic passages, the non-percussive instruments are also producing tones with relatively sharp attacks; thus, there should not be the same kind of difficulty in synchronizing with the drum that we found in Experiment III.

Appendix A

Modification of the NAT Model Due to Local Maxima

The favored PAT model developed in Chapter IV predicts PAT as the time the slope of an instrument's power envelope crosses some threshold. This model is based on the assumption that the slope is monotonically increasing throughout most of the attack portion of the tone, and that once the slope crosses the threshold, it remains higher than the threshold until it reaches a point corresponding to the steady-state portion of the tone.

When this assumption of monotonicity fails, the slope exhibits a "local maximum"; the threshold is exceeded temporarily, but the true increase in slope comes later. In such a case, PAT will be artificially predicted too early.

Figure A.1 illustrates two possibilities that need to be taken into consideration. In the left graph is pictured a segment of a slope function that crosses the slope threshold at time t_1. After a local maximum is reached, the function attains a local minimum at time t_2. This minimum is still above the threshold, but t_1 is in all probability too early a prediction for PAT. In the right graph, the local minimum dips below the threshold; in this case, the slope crosses above the threshold at points t_1 and t_2.

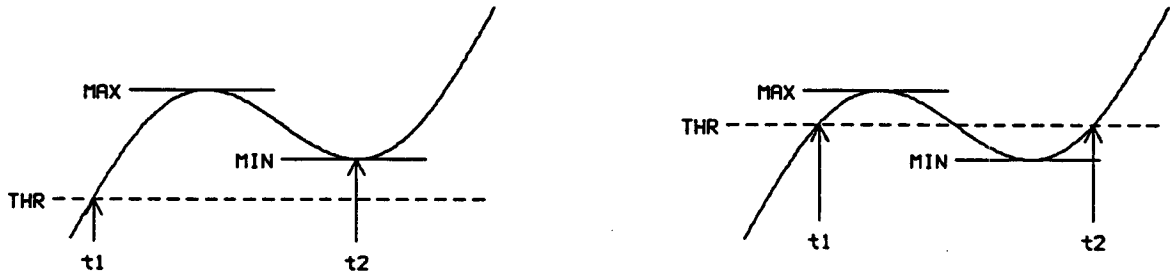

Figure A.1 Two possible relationships between a local maximum and the slope threshold. The left graph shows a minimum at time t_2 that does not dip as low as the threshold, which is crossed at time t_1. In the right graph, the minimum *does* dip below the threshold; the slope thus crosses above the threshold at times t_1 and t_2. In the right graph, the computed "α" (see text) will be 0.

A.1 Derivation of Modification Formula

In Figure A.1, t_1 and t_2 are both logical choices for predicting PAT, regardless of whether or not the minimum actually dips below the threshold. It seems reasonable then to modify the formula so that an average of the two times (t_1 and t_2) is used to predict PAT. In fact, a *weighted* average seems appropriate, with the depth of the minimum determining the value of the weight. If the minimum dips below the threshold (as in the right graph of Figure A.1), we might expect the ear to ignore the local maximum entirely—t_2 alone would thus determine PAT. If there were no minimum (or only a "flattening out" spot, such that the "minimum" was the same height as the "maximum"), t_1 alone would be used to determine PAT. As the minimum gets closer to the threshold, more weight should be given to t_2.

The weight, then, called α, can be determined as:

$$\alpha = \begin{cases} \dfrac{\text{minimum - threshold}}{\text{maximum - threshold}}, & \text{if minimum} > \text{threshold;} \\ 0, & \text{if minimum} \leq \text{threshold.} \end{cases}$$

PAT then becomes a weighted average of t_1 and t_2:

$$\text{PAT} = \alpha t_1 + (1 - \alpha) t_2.$$

It can be seen that when the minimum dips below the threshold, α becomes 0, and all weight

will be given to t_2. When the minimum and maximum are equal, α becomes 1, and all weight is given to t_1.

Further Weighting

Unfortunately, this formula is too simplistic. In the case illustrated by the right graph of Figure A.1, t_1 needs to be given more influence, if only because it signals the *first* crossing of the slope above threshold (i.e., the ear does *not* totally ignore the local maximum). We can thus assign a second weight, ω, to t_1 regardless of α, and the remaining weight $(1-\omega)$ to the above formula:

$$\begin{aligned} \text{PAT} &= \omega t_1 + (1-\omega)[\alpha t_1 + (1-\alpha)t_2] \\ &= [\omega + (1-\omega)\alpha]t_1 + (1-\omega)(1-\alpha)t_2. \end{aligned}$$

Setting $\omega = 1$ places full weight on t_1; setting $\omega = 0$ places full weight on the α-weighted average. In practice, $\omega = 0.6$–0.8 seemed to be an appropriate compromise value.

Appendix B

Shift Formula for PATs of Impulsive Tones

The slope-threshold model for PAT developed in Chapter IV is based on two assumptions that do not always hold. The first, dealing with monotonicity of slope, was discussed in Appendix A. The assumption regarding impulsive attacks will be discussed here.

When low-amplitude onset activity is detected by the ear prior to the true impact of the attack, it serves as an "advanced warning" for that attack; the slope threshold that signals the time of attack can thus be a relatively low one. When no such onset activity is present and the attack comes simultaneously with onset, the ear is given no advanced warning. In this case, slope threshold is crossed at the time of physical onset, and will result in a predicted PAT that is too early. (It can be safely assumed that the ear needs at least 2–3 ms to respond to impulse-like attacks.)

To render the slope-threshold model applicable to both kinds of attacks, we need to be able to distinguish between impulsive attacks and ones which are preceded by onset "warning" activity. This distinction is easily accomplished by applying the PAT model to the slope of the power envelope as before, but also checking to see if the PAT thus predicted is negative or very near zero. If so, the attack will be impulsive and the predicted PAT will have to be shifted (delayed) slightly; if not, no such shift is necessary.

B.1 Derivation of Shift Formula

A negative value for PAT can occur due to interpolation. The slope function is discrete, with samples occurring every msec. In general, the threshold is crossed in between sample periods, and linear interpolation is used to approximate the actual "time of crossing." When the first sample (at time 0) is larger than the threshold, interpolation will thus yield a PAT between -1 and 0. The more impulsive the attack, the closer PAT will be to -1.

We need to determine a "critical PAT," such that PATs less than this will be shifted, and ones greater than the critical value will not be shifted. Zero seems an appropriate value for the critical PAT; however, slopes that cross the threshold in the first 1–2 ms are sufficiently impulsive that resulting PATs should also be shifted. The critical PAT should thus be in the vicinity of 1–2 ms.

The amount of shift should depend on how impulsive the attack is. Thus, PATs near -1 should be shifted the maximal amount, and ones near the critical PAT (but still below it) should hardly be shifted at all. We therefore desire a monotonically decreasing curve between -1 and the critical PAT (t_c), which ranges in value from the maximum shift, s, to zero. The maximal shift should be 3–4 ms; this should be lessened somewhat for low intensity levels, however, since lowering the intensity renders attacks less impulsive.

The curve can be approximated with a parabola: $y = at^2 + bt + c$, where t is the independent variable, PAT. We already have two constraints: (1) $y = s$ when PAT $= -1$, or $y(-1) = s$, and (2) $y = 0$ when PAT is the critical value, t_c: $y(t_c) = 0$. To make the formula unique, we can add a third constraint of making the derivative zero at $t = -1$:

$$\begin{aligned} dy &= 2at + b, \\ dy(-1) &= 2a(-1) + b = 0, \\ b &= 2a. \end{aligned}$$

The first constraint can then be used to determine c:

$$\begin{aligned} y(-1) = s &= a(-1)^2 + b(-1) + c, \\ &= a - b + c, \\ &= a - 2a + c = c - a, \end{aligned}$$

or

$$c = a + s.$$

Finally, the second constraint determines a:

$$\begin{aligned} y(t_c) = 0 &= at_c^2 + bt_c + c, \\ &= at_c^2 + 2at_c + a + s, \\ &= a(t_c + 1)^2 + s. \end{aligned}$$

Thus,

$$a = -\frac{s}{(t_c + 1)^2}.$$

In Chapter IV, since the listening level for Experiments I, II, and III was 90 dB_A, s was set to 4. t_c was set to 1. In this case, $a = -4/(1+1)^2 = -1$, $b = 2a = -2$, and $c = a + s = 4 - 1 = 3$. Then, $y = -t^2 - 2t + 3$, $y(-1) = 4$, $y(t_c = 1) = 0$, and $y(0) = 3$. Thus, all PATs initially less than 0 are shifted between 3 and 4 ms; ones between 0 and 1 are shifted proportionately much less.

In Chapter V, s was set to values less than 4, due to lower intensity levels used in the Vos-Rasch experiments.

References

Allan, Lorraine G. "Is there a constant minimum perceptual duration," *Quarterly Journal of Experimental Psychology 28* (1976) 71–76.

Allan, Lorraine G. "The perception of time," *Perception and Psychophysics* (1979) 340–354.

Bengtsson, I. and A. Gabrielsson. "Methods for analyzing performance of musical rhythm," *Scandinavian Journal of Psychology 21* (1980) 257–268.

Broadbent, D. E. & P. Ladefoged. "Auditory perception of temporal order," *Journal of the Acoustical Society of America 31* (1959) 1539.

Chistovich, L. A. & V. A. Ivanova. "Mutual masking of short sound pulses," *Biophysics 4, no.2* (1959) 46–57.

Cutting, J. E. & B. S. Rosner. "Categories and boundaries in speech and music," *Perception and Psychophysics 16* (1974) 564–570.

Cutting, J. E., B. S. Rosner, & C. F. Foard. "Perceptual categories for musiclike sounds: Implications for theories of speech perception," *Quarterly Journal of Experimental Psychology 28* (1976) 361–378.

Deatherage, B. H. & T. R. Evans. "Binaural masking: Backward, forward, and simultaneous effects," *Journal of the Acoustical Society of America 46* (1969) 362–371.

Efron, Robert. "The relationship between the duration of a stimulus and the duration of a perception," *Neuropsychologia 8* (1970*a*) 37–55.

Efron, Robert. "The minimum duration of a perception," *Neuropsychologia 8* (1970*b*) 57–63.

Efron, Robert. "Conservation of temporal information by perceptual systems," *Perception and Psychophysics 14* (1973) 518–530.

Fraisse, Paul. "Time and rhythm perception," *in Carterette, E. C. and M. P. Friedman, eds., Handbook of perception*; New York: Academic Press, v. 8 (1978) 203–254.

Fraisse, Paul. "Rhythm and Tempo," *in Deutsch, D., ed., The psychology of music*; New York: Academic Press (1982) 149–180.

Gabrielsson, Alf. "Performance of rhythm patterns," *Scandinavian Journal of Psychology 15* (1974) 63–72.

Getty, D. J. "Discrimination of short temporal intervals: A comparison of two models," *Perception and Psychophysics 18* (1975) 1–8.

Gol'dburt, S. N. "Persistence of auditory processes within micro-intervals of time (new data on retroactive masking)," *Biophysics 6, no.6* (1961) 76–81.

Green, David M. "Temporal acuity as a function of frequency," *Journal of the Acoustical Society of America 54* (1973) 373–379.

Grey, John M. *An exploration of musical timbre*; Stanford, CA: Stanford University, Department of Music Report No. STAN-M-2, 1975.

Grey, John M. & John W. Gordon. "Perceptual effects of spectral modifications on musical timbres," *Journal of the Acoustical Society of America 63* (1978) 1493–1500.

Hirsh, Ira J. "Auditory perception of temporal order," *Journal of the Acoustical Society of America 31* (1959) 759–767.

Homick, J. L., L. F. Elfner, & G. G. Bothe. "Auditory temporal masking and the perception of order," *Journal of the Acoustical Society of America 45* (1969) 712–718.

Liss, Phillip. "Does backward masking by visual noise stop stimulus processing," *Perception and Psychophysics 4* (1968) 328–330.

Lunney, H. W. M. "Time as heard in speech and music," *Nature 249* (1974) 592.

Marcus, Stephen M. "Acoustic determinants of perceptual center (P-center) location," *Perception and Psychophysics 30* (1981) 247–256.

Massaro, Dominic W. "Preperceptual auditory images," *Journal of Experimental Psychology 85* (1970) 411–417.

Massaro, Dominic W. "Effect of masking tone duration on preperceptual auditory images," *Journal of Experimental Psychology 87* (1971) 146–148.

Massaro, Dominic W. "Preperceptual images, processing time, and perceptual units in auditory perception," *Psychological Review 79* (1972*a*) 124–145.

Massaro, Dominic W. "Stimulus information versus processing time in auditory pattern recognition," *Perception and Psychophysics 12* (1972*b*) 50–56.

McGill, William J. "Loudness and reaction time: a guided tour of the listener's private world," *Acta Psychologica 19* (1961) 193–199.

Michon, J. A. "Studies on subjective duration: I. Differential sensitivity in the perception of repeated temporal intervals," *Acta Psychologica 22* (1964) 441–450.

Michon, John A. "Le traitement de l'information temporelle," *in Fraisse, Halberg, Lejeune, et al, eds., du temps biologique au temps psychologique,* Symposium de l'Association de

psychologie scientifique de langue française, Poitiers, 1977; (1979) 255–287.

Morton, J., S. Marcus, & C. Frankish. "Perceptual centers (P-centers)," *Psychological Review 83* (1976) 405–408.

Pastore, R. E., L. B. Harris, & J. K. Kaplan. "Temporal order identification: Some parameter dependencies," *Journal of the Acoustical Society of America 71* (1982) 430–436.

Patterson, James H. & David M. Green. "Discrimination of transient signals having identical energy spectra," *Journal of the Acoustical Society of America 48* (1970) 894–905.

Penner, M. J. "The perception of offset: A problem of decision criteria," *Perception and Psychophysics 17* (1975) 587–590.

Penner, M. J. "Variability in offset judgments," *Bulletin of the Psychonomic Society 12* (1978) 32–34.

Penner, M. J. "Evidence for two temporal processes in forward masking," *Journal of the Acoustical Society of America 68* (1980) 455–457.

Pollack, Irwin. "Onset discrimination for white noise," *Journal of the Acoustical Society of America 35* (1963) 607–609.

Rasch, R. A. "The perception of simultaneous notes such as in polyphonic music," *Acustica 40* (1978) 21–33.

Rasch, R. A. "Synchronization in performed ensemble music," *Acustica 43* (1979) 121–131.

Roederer, Juan G. *Introduction to the physics and psychophysics of music*; New York: Springer-Verlag, 1975.

Rosen, Stuart M. and Peter Howell. "Plucks and bows are not categorically perceived," *Perception and Psychophysics 30* (1981) 156–168.

Samoilova, I. K. "Masking of short tone signals as a function of the time interval between masked and masking sounds," *Biophysics 4, no.5* (1959) 44–52.

Schafer, Ronald W. and Lawrence R. Rabiner. "A digital signal processing approach to interpolation," *Proc. IEEE 61* (1973) 692–702.

Stevens, J. C. and J. W. Hall. "Brightness and loudness as functions of stimulus duration," *Perception and Psychophysics 1* (1966) 319–327.

Stevens, S. S. "The measurement of loudness," *Journal of the Acoustical Society of America 27* (1955) 815–829.

Tenney, James C. "Discriminability of differences in the rise time of a tone," *Journal of the Acoustical Society of America 34* (1962) 739 (A).

van Heuven, V. J. J. P. & M. P. R. van den Broecke. "Auditory discrimination of rise and decay times in tone and noise bursts," *Journal of the Acoustical Society of America 66* (1979) 1308–1315.

Vos, Joos & Rudolf Rasch. "The perceptual onset of musical tones," *Perception and Psychophysics 29* (1981) 323–335.

Whitfield, I. C. "The neural code," *in Carterette, E. C. and M. P. Friedman, eds., Handbook of perception*; New York: Academic Press, v. 4 (1978) 163–183.

Wier, Craig C. & David M. Green. "Temporal acuity as a function of frequency difference," *Journal of the Acoustical Society of America 57* (1975) 1512–1515.

Winer, B. J. *Statistical principles in experimental design, 2nd edition*; New York: McGraw-Hill, 1971.

Wright, H. N. "Audibility of switching transients," *Journal of the Acoustical Society of America 32* (1960) 138.

Wright, H. N. "Temporal summation and backward masking," *Journal of the Acoustical Society of America 36* (1964) 927–932.

Zwicker, Eberhard. "Procedure for calculating loudness of temporally variable sounds," *Journal of the Acoustical Society of America 62* (1977) 675–682.

Zwicker, E. & B. Scharf. "A model of loudness summation," *Psychological Review 72* (1965) 3–26.

Zwislocki, J. "Theory of temporal auditory summation," *Journal of the Acoustical Society of America 32* (1960) 1046–1060.

Zwislocki, J. J. "Temporal summation of loudness: an analysis," *Journal of the Acoustical Society of America 46* (1969) 431–441.

Zwislocki, J. J. "Masking," *in Carterette, E. C. and M. P. Friedman, eds., Handbook of perception*; New York: Academic Press, v. 4 (1978) 283–336.