Center for Computer Research in Music and Acoustics February 1975

Department of Music
Report No. STAN-M-2

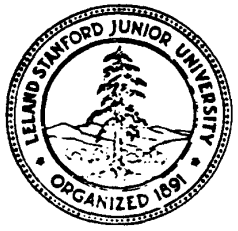
# AN EXPLORATION OF MUSICAL TIMBRE

by

John M. Grey

CCRMA

DEPARTMENT OF MUSIC
Stanford University



Center for Computer Research in Music and Acoustics February 1975

Department of Music
Report No. STAN-M-2

# AN EXPLORATION OF MUSICAL TIMBRE

by

John M. Grey

ABSTRACT

Due to its overwhelming complexity, timbre perception is a poorly understood subject in the field of auditory perception. Computer-based research tools have been developed that appear to be important for an investigation of timbre perception. In the work to be described, an exploratory approach was formulated for dealing with this highly multidimensional attribute of sound. This approach utilized a computer technique for the synthesis of musical timbres based on the analysis of natural instrument tones. This technique was useful for generating stimuli in timbre experiments because of its to effectiveness in allowing the investigator to specify and manipulate the physical properties of complex time-variant tones. An important discovery resulted suggesting that naturalistic tones can be synthesized from a vastly simplified set of physical properties. These simplified tones were useful as stimuli in further studies on timbre perception because of the great reduction in the number of physical factors to be considered in making psychophysical interpretations of perceptual data.

## ACKNOWLEDGEMENTS

It is somewhat difficult to detail this author's thanks to the many friends and colleagues who played important roles in the unfolding of this research effort, in that their contributions were so multiplicative. With that in mind, please allow this simplified attempt at acknowledgement. I would like to express my sincere gratitude to John Chowning, Max Mathews, Jean-Claude Risset and Loren Rush, whose various research projects in the synthesis of timbre initiated and continuously stimulated much of this exploratory investigation; to Jim Beauchamp and Dave Wessel, whose generously communicated work in the areas of computer analysis and perceptual scaling of timbre also inspired many of the directions of my own investigations; to Roger Shepard, not only for his vital role in the original development of the perceptual scaling techniques used in this work, but also for his substantive support of this multidisciplinary research project; to Earl Schubert for his invaluable guidance and engaging discussions in the areas of psychoacoustics and the perception of music; to Jay Dowling who sparked my interests while educating me in these fields as an undergraduate; to Herb Clark and Dorothy Huntington for their stimulating discussions on the perception of language and music, and for their useful advice on experimental techniques. I am very indebted to Andy Moorer for his implementation and development of computer techniques for the analysis of musical instrument tones with harmonic overtone structures, without which this research could not have been undertaken; to Phipps Arabie, who provided copies of the multidimensional scaling algorithms which he implemented on the PDP-10 computer; to Bruce Baumgart, whose advice and software aided in the construction of many of the graphical techniques so important in this research; to the Stanford Artificial Intelligence Laboratory, to the Stanford Institute for Mathematical Studies in the Social Sciences, to the Stanford Departments of Psychology and Hearing and Speech Sciences, and the Center for Music Experiment at the University of California at San Diego, all of whom generously provided resources for this work; and to the many friends and musicians who participated as performers in recording sessions or as listeners in the perceptual research. Finally, I express special gratitude to Elizabeth Dreisbach, who was largely responsible for running the perceptual experiments and whose continuing interest in and support of this work played a vital part in its eventual accomplishment.

of the attack). The third dimension was interpretable either in terms of a physical property (synchronicity in the attacks of higher harmonics) or as a *higher-level* distinction made between the tones on the basis of their musical instrument family. Another set of studies next initiated an exploration of timbre in terms of continuous versus categorical perception. An algorithm was designed to generate a set of tones interpolating between two naturalistic timbres. Identification, discrimination and perceptual similarity studies were performed using a set of stimuli generated by interpolations. The results of these studies suggested that interpolations were perceived to be continuous rather than categorical. Furthermore, the timbral similarities between a partial set of the naturalistic and interpolated tones revealed three perceptual dimensions that related directly to those found above for the total set of naturalistic stimuli. The first two physically-related dimensions were found, and the third dimension seemed to correspond to a higher-order distinction made between naturalistic tones and the interpolation-derived tones, this superseding the family distinction made for the total set of naturalistic tones. A notion of timbre is developed involving both a *higher-level* perceptual processing of tones that has access to stored information relating to the distinctive features of identifiable sources, and a *lower-level*, qualitative perceptual comparison of tones with respect to gross acoustical features lying outside of the domain of specific identification. Suggestions for future research are made.

## TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF ILLUSTRATIONS

# I. HISTORICAL REVIEW OF TIMBRE RESEARCH

*Introduction to State of Knowledge*

While there have been many significant discoveries in the field of audition during the last 100 years, there has been little advancement towards understanding the perception of sounds which have sets of attributes nearly approximating the full multidimensionality found in nature. This may be largely due to the overwhelming complexity of the auditory system which has necessitated a complementary simplicity in the stimuli used for its examination. Such sounds as pure tones, pulse trains, bands of noise, and more recently, small sets of sinusoids, have had the multiple advantages of being well-defined in their physical properties, relatively easy to produce, and simple enough to allow for interpretable responses in the auditory system. These are still considered by most investigators to be the best types of signals with which to study the psychophysical and physiological properties of the ear.

It is of no surprise, therefore, that the state of knowledge concerning the perception of sounds which have the complexity found in musical instrument tones is much less advanced on a scientific level than on the practical and intuitive basis obtained by experienced musicians. Indeed, it is the *musical* term 'timbre' which has come to denote these more complex qualities of tone. By timbre, the musician indicates the tonal qualities which characterize a particular musical sound. The meaning of the term timbre is not strictly defined, however. It may refer to the features of tone which serve to identify that a musical sound originates from some particular instrument or family of instruments, for example, that it is an oboe, or perhaps some sort of double-reed instrument, or maybe just some woodwind instrument. Timbre may also be used by the musician to denote some tonal quality of performance on a given instrumental source, as in a dark, or dull, or bright, or shrill oboe tone.

In the psychoacoustical literature there is also no firm agreement on the meaning of this term with respect to the nature of the various auditory phenomena which should be included in its definition. Even the most quoted definition, approved by the American Standards Association [1960], has given rise to many different interpretations: *Timbre is that attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar.* It is added that: *Timbre depends primarily upon the spectrum of the stimulus, but it also depends upon the waveform, the sound pressure, the frequency location of the spectrum, and the temporal characteristics of the stimulus.* It is easy to see why there exists no commonly accepted model for timbre, but rather there are large discrepancies in the specifications found in the literature for timbre. In essence, then, timbre may comprise any subset of acoustic phenomena other than pitch and loudness (and, we might add, duration and spatial location).

A presentation of substantive knowledge about the perception of timbre for complex tones is universally lacking in the classical treatises on hearing. In their very well-known book, Stevens and Davis [1938] follow chapters individually devoted to the perception of pitch and loudness

with a chapter on "other attributes of tones." In it they treat the dimensions of volume, density, and brightness for pure tones only. Similarly, in listing the "seven more important" attributes of tone, Boring [1942] lists only pitch, loudness, brightness, volume, locality, tonality, and density. Licklider [1951], in discussing the attributes of complex tones, concludes that "until careful scientific work has been done on the subject, it can hardly be possible to say more about timbre than that it is a 'multidimensional' dimension."

In this presentation of the literature on timbre perception, we will begin with that theory of timbre perception which proposes that the timbre of a tone can be characterized by its steady-state spectrum. Although this *classical* theory of timbre is still very much with us today, most contemporary investigators believe that the many and various temporal, or time-variant, factors in natural signals must also be taken into account in a full description of timbre perception. We will present empirical evidence which has been gathered in support of the notion that the *identification* of timbre is very dependent upon such temporal factors. The very important contributions of more recent computer-based research into the *analysis* and *synthesis* of timbre will also be covered. In this latter research, the investigator is able to control many of the physical factors for timbre in great detail, hence much information about the various features involved in the timbral evaluation of complex tones has come to light from such work. Finally, we will look at the most current psychoacoustical research on timbre perception, where other computer-based techniques for complex perceptual data analysis, *multidimensional scaling*, have been employed to explore this most 'multidimensional' dimension of tone.

## A. Timbre as Tonal Quality in the Steady-State

### *The Classical Theory of Helmholtz on the Harmonic Amplitude Pattern*

As with most other phenomena in the domain of auditory and visual perception, Helmholtz [1954] became involved with the topic of the perception of complex tones. His now classical psychophysical theory holds that differences in the timbre of complex tones depend on the presence and strength of harmonics in the tones, but probably do not depend to any great extent on the differences in phase under which those partials unite. It is important to note that Helmholtz clearly *restricted* the domain of physical phenomena which is relevant to the definition of timbre. Only the spectrum of a *steady-state* periodic waveform is mentioned, thereby neglecting all the temporal phenomena which exist in natural tones. As we shall see, many investigators since Helmholtz have similarly limited their studies on timbre to the examination of this type of restricted signal, a steady-state tone. Far less attention has thereby been paid to that type of tone which dynamically changes with time, as is the case with all natural signals, which are characteristic of *normal* listening experiences.

After demonstrating that the complex periodic vibrations of musical and vocal sounds consist of sets of harmonics, Helmholtz showed that the ear can distinguish a number of these harmonics individually. He applied Ohm's acoustical law [1843], that the ear performs a Fourier analysis on a complex periodic waveform, yielding a series of harmonically related sinusoids of specific phases and amplitudes, and he proposed that the *amplitude pattern* and *bandwidth* of this series is what is perceived in judging the timbre of the complex wave. His use of the doctrine of unconscious inference allowed that the complex tone be analyzed into its harmonic series and also be fused in the normal mode of conscious perception.

Helmholtz found the following rules concerning the psychophysical relationships in the perception of musical tones [pp 118-119]: simple tones sound sweet and pleasant, without any roughness, but dull at low frequencies; complex tones with moderately loud lower harmonics, up to the 6th, sound more musical and rich than simple tones, but they are still sweet and pleasant if the higher harmonics are absent; complex tones consisting of only odd harmonics sound hollow and, if many harmonics are present, nasal; predomination of the fundamental gives a full tone, in the reverse the tone sounds empty; complex tones with strong harmonics beyond the 6th or 7th sound sharp, rough and penetrating.

The next matter which Helmholtz investigated was the dependence of timbre on the *phase* relationships between the harmonic components of a complex waveform [p 126]. On the basis of many observations of synthetic tones using a very ingenious apparatus which controlled the resonance of a series of tuning forks, he concluded that the phase pattern of the harmonic series contributed very little to the perception of timbre. He did, however, consider it to be possible that the phase relationships of the upper harmonics, which give rise to roughness in the sound, could influence the timbre of the complex wave.

*The Notion of Formant Regions and Critical Bandwidths*

A supplement to the classical theory of Helmholtz was the notion of *formant* regions in the harmonic series. A formant is a particular frequency range in which harmonics are much higher in amplitude than the harmonics in neighboring ranges. The usual graphical representation of a formant region displays successive harmonics of a complex tonal spectrum as a series of vertical lines, where the position of each successive line along the horizontal is related to its frequency and the height of each line is related to its relative amplitude. In such a display of the amplitude pattern of successive harmonics, one can connect the tops of the successive lines to one another, and this outlines the *spectral envelope* of the complex tone. A formant region appears to be a mountain in the spectral envelope.

The formant model was adapted to musical tones by Fletcher [1934], from theories regarding speech perception. Bartholomew [1945] similarly proposed that "the characteristic tone quality of an instrument is due to the relative strengthening of whatever partial lies within a fixed or relatively fixed region of the musical scale." The concept of formants does not actually change the domain of physical phenomena considered to be relevant to the perception of timbre beyond the steady-state amplitude pattern.

In view of more recent studies on the role of *critical bandwidths* in hearing, however, the formant model presents a more complete framework for understanding the perception of a spectral envelope. Critical bandwidths have been empirically defined as frequency regions in which the ear seems to integrate acoustical stimulation. These bandwidths are continuously spread along the frequency axis, and are roughly 1/3 octave in width. The exact size of a critical bandwidth varies with the absolute frequency region in which the particular phenomenon is being measured. Such actions of the ear as loudness summation, masking, threshold levels, and sensitivity to phase have been associated with critical bandwidths [Zwicker, Flottorp, and Stevens, 1957].

It is hypothesized that these phenomena may affect the perception of the amplitude pattern for a harmonic series, when sets of partials fall within the same critical band. Recent studies on the discrimination of individual harmonics have shown a critical bandwidth effect, where it was found that the number of partials which can be distinguished under the most favorable conditions is limited to the lower 5 to 8 harmonics, depending upon the frequency of the fundamental [Plomp, 1964; Plomp and Mimpen, 1968].

This has led many theorists to the conclusion that a formant model may be a better representation of the perceptual processing of complex tones since it is more compatible with the filtering action of the ear [Plomp, 1970; Roederer, 1972]. The relevant consequence of such a filtering network in the frequency domain of hearing to timbre perception is seen in the case that two or more harmonics of a complex tone fall within a critical bandwidth of one another. In that case, which is usually true for partials above the 6th to 8th harmonics, the ear, by definition, will not be able to discriminate the individual harmonics and their loudness will sum according to total energy present.

*Fixed versus Relative Formant Frequencies to the Fundamental*

After the introduction of the formant model to timbre, a question arose concerning the relationship between the locations of formant regions in tones having different pitches and their perceived timbres. The controversy was whether the frequency region of a formant for a particular timbre should change *relative* to the pitch, or remain at a constant *fixed* position to maximally preserve the timbre for different pitched notes. This controversy was an outgrowth of the same question in the perception of vowels, where the same vowel from male, female or child speaker will have very different pitches, but the formant regions were thought to have remained more fixed in frequency than relative to the pitch.

In that many musical instruments do not have physical formant regions fixed in frequency across a wide range of pitches, but rather exhibit transposable patterns of harmonic amplitudes relative to the frequency of the fundamental [e.g. Luce, 1963], there has been some question over the adequacy of the fixed formant model for the perception of timbre. Investigations of this issue still do limit the physical domain of tones to steady-state spectra.

Slawson [1968] elicited similarity judgments for pairs of vowel-like synthetic tones which had systematically and independently transposed fundamental frequencies, lower two formant frequencies, and higher formant frequencies. Half of the listeners were instructed to consider these tones to originate from musical instruments, and the other half to view them as speech sounds. The results in both cases were interpreted as favoring a modified fixed formant theory. On the average it was found that tonal quality was best preserved when the lower two formants were transposed by a small percentage, around 10 percent, of the amount by which the fundamental frequency was transposed. A related study on the multidimensional perception of pitch and timbre also supported the fixed formant theory [Plomp and Steeneken, 1971]. Pulse trains were presented at various rates through filters having different center frequencies, and tones filtered at fixed frequencies were judged to be more similar than tones filtered at frequencies relative to their pulse rates.

Webster, et. al. [1968a, 1968b, 1970] initiated another line of investigation concerning perceptual models for steady-state spectra. Rather than measure the perceived similarities between tones, as was the case in the above two studies, they examined the ease of identification for various complex tones with which listeners were trained to associate specific labels. The first studies employed sets of meaningless buzz-tones which had selected harmonics raised in amplitude. It was found that tones could be identified relatively well when the spacing between emphasized partials was wider than a critical bandwidth, or where two adjacent partials were both raised.

In addition to these meaningless tones, later studies also included steady-state portions of meaningful sounds such as vowels and musical instruments. The original sounds and transposed versions of them, up and down an octave, were used as the stimuli for identification

training. It was found that the non-transposed vowel sounds, which are presumedly dependent on fixed formant frequencies, were best identified. The musical instruments, however, were found to have no advantage at their original octaves. A fixed ratio, not a fixed frequency, hypothesis was thought to be applicable to the tones which showed no octave advantages, whereby the identification of such tones was considered to depend upon the transposable harmonic pattern itself rather than upon certain fixed formant frequencies. This would suggest that the perception of steady-state timbre may not be restricted to a single mode, and is highly dependent upon the strategy of the listener. However, it is not entirely unequivocal that the above set of studies present sufficient evidence in support of the fixed ratio hypothesis, in that the perceptual relationships between tones were not directly examined in terms of similarity.

*The Harmonic Phase Pattern*

A physical aspect of steady-state tones which Helmholtz dismissed as not being important to the perception of timbre is the phase pattern of a set of harmonics of a complex tone. An excellent history covering the controversy over the importance of phase since Helmholtz appears in the study we site here by Plomp and Steeneken [1969]. Using multidimensional scaling techniques [Kruskal 1964a,1964b], they were able to isolate the two phase patterns which are perceived to be maximally different for pairs of tones having identical harmonic amplitude patterns. It was found that the phase pattern did play a relatively *minor role* in the perception of timbre, and that the maximal difference due to phase was about equivalent to the changes in quality between very closely related vowels.

Measurements indicated that the qualitative effect of this maximal difference for the phase pattern decreased with increasing fundamental frequencies of tones. The perceptual effect of maximal phase pattern differences was compared to the variables of high-frequency attenuation rate and overall sound pressure level (SPL) of signals. When compared with the effect of changing the slope of high-frequency amplitude roll-off, the maximal phase difference was perceptually equal in magnitude to a change in the slope of high-frequency attenuation by about 2 dB/octave for fundamentals around 150 Hz and a change of .7 dB/octave for fundamentals in the neighborhood of 600 Hz. In the central range of fundamentals studied, the maximal difference based on the phase pattern was found to be quantitatively equal to an overall change in SPL by 2 dB. Results also indicated that the effect of phase pattern was independent of both the harmonic amplitude pattern and the SPL of the tone within the ranges studied.

**B. Verbal Measurements of Timbre**

From an assessment of the verbal descriptions of the differences in timbre given by subjects in the above study, Plomp and Steeneken [1969] found that the same labels, *brightness* and *sharpness*, were used to indicate differences based on two independent features: the pattern of

relative phase and the slope of the attenuation of the higher harmonics. They concluded that the measurement of timbre using semantic scales was not at all a simple matter, since such scales did not reveal factors which were *uniquely* related to independent properties of the stimuli. However, a number of studies on timbre have been based upon just this kind of verbal measurement.

As mentioned above, Helmholtz [1954] presented the psychophysical relationships for complex tones in terms of such verbal labels for perceptual attributes as *brightness, richness, sweetness, pleasantness, fullness,* and *roughness.* Many psychological studies of timbre perception have followed which have elicited judgments from subjects on a number of verbal scales. Lichte [1941] had listeners use semantic scales to evaluate sets of steady-state tones whose spectral envelopes were systematically altered in various ways. Three attributes for complex tones, apart from pitch and loudness, were found: *brightness, fullness,* and *roughness.*

*Brightness* was analyzed to be a function of the location on the frequency scale of the midpoint of the energy distribution. This sort of *brightness,* by the way, was not the same as that which is found for pure tones varying in frequency. *Fullness* was interpreted as a function of the relative presence of odd or even harmonics. *Roughness* was found to be present in tones consisting of consecutive higher harmonics above the 6th, and was also a function of the location of these harmonics in the whole sequence of higher partials. This *roughness* nicely corresponds to the phenomenon of dissonance, and would be expected to occur between pairs of harmonics which fall within a critical bandwidth of one another at certain frequency differences, as may occur for harmonics above the 6th. It was suggested that *roughness* and *fullness* may have been different functions of the same parameter of sound, the complexity of frequency ratios between partials. We may neatly reinterpret this finding also with respect to the existence of more than one successive harmonic within a critical bandwidth.

Several recent studies have also employed semantic scales to measure the qualities of complex tones. Typically, large numbers of verbal scales are offered to the listener on which to rate the stimuli. Factor analysis is then performed on the data in order to discover which subsets of the scales were independently applied to the stimuli, and then the possible physical correlates to these scales are inferred. Jost [1967] did an extensive study of the clarinet and found that *volume* and *density* seem to be the most prominent verbal dimensions of timbre. Solomon [1958] found 7 interpretable factors in ratings of 20 sonar sounds on 50 semantic scales, and similar results were obtained for speech sounds by Nordenstreng [1969], with *magnitude, colorfulness, clarity,* and *beauty* as the dominant factors. In a most recent set of semantic differential studies, von Bismarck [1974a] found that the scale *dull-sharp* is the most suitable verbal dimension for the description of timbre in general. He also attempted to show that sharpness was a psychophysical dimension distinct from loudness and pitch [von Bismarck, 1974b].

One of the major disadvantages in using verbal scales to investigate the properties of stimuli, of course, is that words may not exist to describe certain perceived differences, and even if such

words do exist, they may not have been included among the scales available to the subjects. Another problem has been mentioned above, that a single word may be associated with a number of independent stimulus dimensions, making it quite difficult to isolate the psychophysical relationship underlying the semantic judgment. It may be that semantic scales measure some of the complex aesthetic reactions to stimuli and don't provide direct information about many of the perceptual processes. It would therefore be possible to meet with only a limited success by using semantic scaling techniques to examine the perception of timbre

## C. Temporal Features and the Identification of Timbre

### *The Importance of the Attack*

A number of studies have measured the ability of listeners to *identify* the musical instrument which produced a tone which is played back on tape. Many of the tones are presented only in part, having had certain segments spliced away. This is done in order to study the relative importance of the various segments of the tones in their recognition. Instrument tones have classically been thought of to consist of three temporally adjacent segments: 1) the *attack*, which includes the initial transient, temporally changing events as the tone grows in amplitude; 2) the *steady-state*, a condition of stability reached, by definition, at the end of the attack; and 3) the *decay*, the temporally dynamic reduction in amplitude at the end of the tone until it has finished sounding. The temporal dynamics of the attack and decay transients, it was eventually discovered, may not only consist of an overall change in amplitude for all the harmonics of a tone, but also may include a change in the amplitude balance between the harmonics because of their different rates of growing or decaying.

It has generally been found that the attack segment is the *most* important portion of the tone for its identification [Luce, 1963; Saldanha and Corso, 1964; Berger, 1964; Wedin and Goude, 1972]. In one experiment Luce [1963] found better identification for just 60 milliseconds of the attack than for 150 milliseconds of the steady-state. Identification from the steady-state alone is better when tones have *vibrato* [Saldanha and Corso, 1964; Wedin and Goude, 1972], which, of course, is a temporally dynamic phenomenon in the frequency domain, and often causes temporally changing features in the amplitude domain. There seems to be no improvement by including the decay portion of the tone [Saldanha and Corso, 1964].

These experiments show that many of the important *cues* for the recognition of musical instruments are located in the attack segment of the tones. Certainly, the timbral *signature* of musical instruments which experienced musicians refer to must be largely dependent upon the dynamically changing features of tone, and especially those pronounced features in the attack segment of a note. Another possible component of the timbral signature of an instrument comes into play with the existence of more than one note, as in a melodic phrase, and has been referred to as the tracing of the *resonance structure* of an instrument by the sampling of several differently pitched spectra played upon that instrument. Of course, the existence of invariant temporal features in a set of differently pitched notes played upon the same instrument would

also comprise an important cue for identification. The various components of timbre recognition which exist between a set of notes have not been given attention in the literature of timbre perception, in that experiments have universally been done with single, contextless musical tones.

At any rate, if the importance of the attack is indeed general to the perceptual processing of complex tones, then there is much more to be studied than steady-state waveforms in the derivation of an adequate psychophysical model for timbre. This, of course, is an all too frequent observation for most researchers who have evaluated the timbral success of their synthesized sounds or who have examined the complex physical properties of real sounds.

*The Analysis of Natural Tones*

With the growing sophistication of electronic technology, it became possible to analyze the complex microstructure in natural sounds. The most prominent observation was that the amplitudes of the harmonics were always changing relative to one another. It was further found that the various sounds from speech and music each showed very different types of harmonic relationships both during the attack and during the so-called steady-state [Backhaus, 1932]. It has been persistently claimed by acousticians involved with the analysis of natural sounds that the attack portion is a very fertile location for important perceptual cues [Young, 1960; Luce, 1963; Taylor, 1965; Winckel, 1967; Roederer, 1972].

Luce [1963] performed a computer analysis of the acoustical properties of 14 of the nonpercussive instruments of the orchestra, using a Fourier analysis technique which produced *time-variant* information about the amplitudes and phases of the individual harmonics of the tones. A number of invariant or quasi-invariant physical features for various instruments and families of instruments were discovered both in the attack and steady-state portions of the tones. These physical invariants each could be considered to be possible perceptually distinctive features in the auditory processing of the instrumental timbres, and suggested a number of necessary psychological studies. Various papers followed this work which focussed on particular aspects of the instrumental tones.

One paper presented evidence for the existence of perceptual families of instruments [Clark, Robertson and Luce, 1964]. In that the instruments in a family have similar physical characteristics, such as their methods of excitation, their resonant or tone-modifying structures, and their means for radiating acoustical vibrations, an affinity of timbres created within a family was expected. Perceptual *confusions* in the identification of tones played back at differing speeds was tested within families of instruments, such as strings, brass, and double-reeds. The analysis of these confusions revealed tight associations between most instruments within these families. In that attack times were altered with the speed of playback, it was felt that the timbre was not dependent on the mere *duration* of the attack.

It was hypothesized that the timbre of the nonpercussive instruments of the orchestra is characterized by certain gross features of the attack, characteristic modulations during the steady-state, somewhat by one or more formants, and probably by some other physical features not well understood.

The durations of the attack transients were then measured statistically and it was found that they varied with both the particular pitch and player for any single musical instrument, as well as with the type of instrument being measured [Luce and Clark, 1965]. The duration of the attack was not found to be systematically dependent upon either the amplitude level of a tone, the overall length of a tone, or the existence of vibrato in the tone. Since it was not invariant with pitch or player for any one instrument, it was concluded that the duration of the attack is not a sufficient cue for the identification of the timbre of musical instruments. This, in view of the evidence of the extreme importance of the attack segment for the recognition of such tones, suggested that the waveform and/or the changes in the waveform were the cues which characterized musical timbre.

*Timbre and the Loudness of a Played Tone*

Another study by Clark and his colleagues examined the dependencies of timbre on the tonal loudness produced by musical instruments [Clark and Milner, 1964]. Musically competent listeners were asked to guess the *original* dynamic level, either pianissimo, mezzo-forte, or fortissimo, at which various tones were taped. These tones were played back to the listeners at *altered* amplitude levels, so as to be equal with each other in loudness. The lack of success in being able to infer the original dynamic level from the remaining timbre cues alone led to the conclusion that timbre is a very weak function of the amplitude of a tone. This was in spite of the physical observation that the harmonic richness of a tone varies with its overall amplitude level. It was further concluded that the attack transient cannot be relied on as a cue for the amplitude of an instrument.

Conflicting results to these were obtained by Jost [1967] in an investigation which used semantic scales to examine the perceptual attributes of clarinet tones. Notes were recorded at three dynamic levels and were altered in playback to be at any one of those levels, hence a tone recorded at piano could be played back at forte. A factor analysis of the semantic evaluations tended to show some influence from the original amplitude level of the tones, indicating that this original level was in some way recovered in spite of the alterations. It is not clear, however, that either of the two conflicting studies employed methods suitable for a direct examination of the effects of signal amplitude on the perception of timbre.

## D. Synthesis of Timbre and Perceptually Important Features

Strong and Clark [1967a, 1967b] performed a set of experiments on the synthesis and perturbation of wind-instrument tones. A simplified model for synthesis was used, in which the evolution of the partials was controlled by a single *spectral* envelope and three *temporal* amplitude envelopes for three subsets of partials. The shapes of the envelopes and the grouping of the partials into three subsets were both based on the analyses of the physical properties of these instruments [Luce, 1963]. Musically literate subjects identified the synthesized tones with 66 percent accuracy, as compared to 85 percent accuracy for the original tones, and if intrafamily confusions were allowed for, the accuracies raised to 77 percent and 94 percent, respectively.

The relative significance of the spectral and temporal envelopes for the various tones was then evaluated by exchanging these envelopes among the instruments and by producing artificial or perturbed spectral envelopes for the synthesis of new tones. It was found that for instruments which had unique spectral envelopes with respect to the maximum formant frequency and/or the range of the instrument, such as the oboe, clarinet, bassoon, tuba and trumpet, the spectral envelope is indeed dominant in their identification. Where the spectral envelope was not unique, as in the flute, trombone and horn, the temporal envelopes become dominant in significance. Interfamily confusions were fewer when the spectral envelope was of greater importance for the identification of the musical instruments. Although adequacy of the stimuli in these experiments to represent natural sounds can be questioned, the studies themselves mark a large step towards dealing with the multidimensional aspects of timbre perception.

Important computer analyses and syntheses of the time-variant properties of trumpet tones were done by Risset [1966]. By synthesizing tones based on various models for *simplifying* the complex, analyzed parameters of the sounds, Risset concluded that three particular features were aurally important: 1) the *relationships* of the attack times of the harmonics, whereby successively higher harmonics take longer to appear and grow more slowly; 2) the fluctuation of the frequency, which is of small amplitude, fast, and quasi-random; and 3) the harmonic content of the tone, which becomes richer in high-frequencies when the overall intensity increases.

The dependence of the timbre of computer synthesized tones on the evolutions of harmonics has been studied for many classes of sounds by Chowning [1973]. Using a system of synthesis which allows a specification of the amplitude function, the spectral evolution, and the frequency ratios between the partials of a waveform, sounds were synthesized which closely resembled natural musical instruments. Brass-like tones were successfully synthesized on the basis of the relationships of the attack times and evolution of the harmonics, as observed by Risset above. The utility of this method of synthesis suggested that an adequate model for the perception of timbre must include considerations of the nature of the evolution of amplitudes and the frequency relationships for the partials of a tone.

The qualities of violin tones have been studied by Mathews and Kohout [1973] by simulating the *resonance* properties of the instrument's body with a set of analogue filters. The source function of the tones was produced in the normal manner, using bowed violin strings which were amplified and passed through the filters, hence it was the resonance properties of the violin which were being investigated. It was found that the spacing of the peak resonance frequencies with respect to the frequencies of the harmonics, the steepness of the resonance curves, and the closeness and number of resonances were all important determinants of the tonal quality of the simulated instruments. The adequacy of the simulation was demonstrated by the extreme difficulty in distinguishing the natural from the electronically processed tones.

Related studies by Mathews on the qualities of tones in the bass range indicate that the arrangement of the peaks for the upper harmonics determines the resulting consonance of the tone. When several harmonics which lie within a critical bandwidth of each other are emphasized, the tone is rough and harsh in quality. Similar relationships were found in earlier studies cited above on steady-state tones [Helmholtz, 1954; Lichte, 1941; Webster, 1968a]. Indeed, it has become clear that the psychoacoustical concepts of *consonance and dissonance* relate directly to the perception of certain spectral aspects of timbre [see Plomp and Levelt, 1965].

### E. Recent Perceptual Studies using Multidimensional Scaling

*Plomp*

Recent investigations by psychologists on the perception of timbre have shown the use of more sophisticated data analysis techniques, primarily the computer-based multidimensional scaling algorithms (see Appendix C). Plomp [1970] was the first such researcher to investigate the *multidimensional* attributes of tones derived from musical instruments, and he used nonmetric multidimensional scaling techniques, hereafter referred to as MDS, to handle perceptual similarity data [Shepard, 1962a, 1962b; Kruskal 1964a, 1964b]. The study here cited was entirely concerned with that limited definition of timbre which includes only the *steady-state* spectrum of a tone. A model for analysis depended heavily upon the notions of formants and critical bandwidths.

Single periods were cut from the steady-state portions of 9 different instrumental tones and fundamental frequency-normalized periodic tones were then synthesized by repeating any one of those periods for a given length of time. The concept of timbre was thereby restricted to exclude temporal properties of the original tones. The experiment essentially attempted to quantify the perceptual and the physical relationships between all of the tones, in terms of a distance function. This would then allow a comparison of the psychological and physical features of these tones using MDS for the perceptual data and factor analysis for the physical properties, leading to inferences regarding the psychophysical relationships in timbre perception.

The psychological measurement of distance was the relative perceptual similarity between each pair of tones. The physical measurement of distance was made in terms of the differences in amplitude pattern between pairs of tones, and the model for the measurement was based on assumed properties of the ear. The amplitude pattern of a tone was quantified by taking the intensity levels for 18 consecutive 1/3 octave filters, in a simulation of critical bandwidths in the ear. Differences between pairs of instruments were then computed with a physical distance equation which summed the differences in intensity levels obtained for the 18 consecutive filters, hence the results of these calculations placed the sounds as points in an 18-dimensional space. The perceptual similarity matrices and the physical distance matrix were reduced to three-dimensional representations by multidimensional scaling [Kruskal, 1964a, 1964b] and factor analysis, respectively. A good correlation was found between the configuration of perceptual similarities and the differences in the physical spectra measured for the tones. It was noted that the frequencies of the two lower *formants* and the level of the second formant was highly correlated with the three physical factors analyzed.

The validity of the model used by Plomp [1970] in quantifying the physical differences between tones was tested by Wessel [1973]. He examined the perceptual independence of adjacent critical bandwidth amplitude levels, this being implicitly assumed in the Plomp's model which placed tones as points in a space having 18 independent dimensions. Wessel synthesized tones having four harmonics and had subjects rate their perceptual similarities. Upon close inspection, the similarity configurations obtained from MDS gave evidence that the harmonic amplitude levels could not be considered as independent dimensions. Their interactivity, which may be due to the dependence of masking patterns on amplitude patterns, posed a problem for the data reduction model used by Plomp.

*Wedin and Goude*

Wedin and Goude [1972] took a different approach in studying the multidimensional psychophysical relationships for music instrument tones. Both the original full recorded tones from 9 instruments and their segmented-out steady-state portions were used as stimuli for similarity ratings. Factor analysis was applied to the perceptual similarity data according to a vector model of similarity [Ekman, 1965]. The results for the tones with and without the attack and decay transients were found to be extremely similar, and it was concluded that the most important correlates to the factors in timbre perception lie in the *amplitude pattern* of the steady-state spectra of the tones. An analogue spectrum analyzer produced the spectra of the tones which were then factor analyzed with respect to the levels of the first 9 harmonics. Three factors were found, corresponding to: 1) decreasing intensity of the upper harmonics; 2) low fundamental intensity and increasing intensity around the 4th and 7th harmonics; and 3) high fundamental intensity and maximal intensity in the middle frequency range. These factors were fairly correlated with the factors found for the perceptual similarities of the tones.

One quite interesting aspect of the study by Wedin and Goude [1972] was their attempt to investigate the relationship between "perceptual structure", defined to be the factors which describe the listener's experience of the auditory stimulation of the instruments, and "cognitive structure", thought of as the classification scheme based upon intellectual information about the instruments and their functioning. A similarity matrix was obtained with the presentation of just the *names* of the 9 instruments to their 9 most musically-sophisticated subjects. Results from a factor analysis revealed factors based on the classical familial relationships of instruments, and these had a low correlation with the factors found in the perceptual similarity data. They concluded that the cognitive structure was indeed different from the perceptual structure.

The same problems with the model used to quantify the differences between the physical spectra of musical sounds by Plomp [1970] are also very applicable to the model formulated by Wedin and Goude [1972]. In addition, there may be a serious problem in attempting to represent tones, which vary in their number of partials, from 10 to 20, all by their lower 9 harmonics. The use of this metric scaling technique to handle perceptual similarity data is seriously questionable as to its appropriateness for the treatment of *psychological* judgments. The findings of highly similar perceptual matrices for the two types of tones, those with and without their attack and decay transients, might be explained in part by the lengthy durations for all of the tones, the existence of characteristic vibrato in some of the tones in both conditions, and the approximate nature of the equalization of pitches and loudnesses for the tones.

*Wessel*

Wessel [1974] also did a study on the multidimensional perception of musical timbre. His stimuli consisted of the full tones from 9 instruments of the orchestra, each instrument playing a series of tones at approximately the same pitch and loudness. Listeners could freely rate pairs of tones for their similarity by switching between two ongoing channels of a tape loop. The recently developed individual differences multidimensional scaling techniques, hereafter referred to as INDSCAL [Carroll and Chang, 1970; also see Appendix C], were applied to the data and the resulting two-dimensional configurations were interpreted with the aid of computer analysis of the tones [Beauchamp, 1969]. The INDSCAL technique allows the investigator to evaluate the behavior of individual subjects in the relative importances that they appeared to assign to the various perceptual dimensions uncovered in the analysis.

One of the dimensions related to the amplitude pattern of the harmonics, that is, the manner in which energy was distributed through the frequency regions of the tone. Tones with energy concentrated in their lower harmonics appeared at one end of the scale, while tones with more high-frequency energy appeared at the other end. The other dimension seemed to relate to the temporal properties of the tones, in that tones were grouped by family: trumpet-trombone-

French horn, oboe-bassoon-clarinet, and violin-viola-cello. Several physical factors might come into play in this dimension of tone.

Multidimensional scaling was also performed on the similarity ratings of the aural *images* which subjects generated, corresponding to pairs of instrument names presented [Wessel, 1974]. The resulting configuration demonstrated a difference in the perceptual and cognitive structures for the instruments, the latter seeming to be more weighted by the common temporal properties across families of instruments, suggesting that the steady-state, spectral information of musical tones is not used to any large extent in the imagined condition.

A composite scaling of the two conditions in this experiment, the heard and imagined, along with a re-scaling of the three conditions in the Wedin and Goude study, the heard full-tone, the heard steady-state, and the imagined, using the INDSCAL technique, confirmed the above observations [Wessel, 1974]. Two dimensions were interpreted, the amplitude-pattern, energy distribution dimension, and the temporally related family axis. In addition, in both sets of data there was an orderly decrement in the perceptual weighting given to the energy-distribution dimension from the heard steady-state, to the heard full-tone, to the imagined conditions. This confirmed the interpretation of the dimensional axes, and it re-affirmed the notion that in imagining tones, the steady-state properties play a minor role compared to the temporally-related cues for single tones.

# II. PRESENT CONCERNS IN TIMBRE RESEARCH

## A. Basic Goals for Current Research in Timbre Perception

*Analysis and Synthesis of Natural Timbre for Distinctive Features*

A first order goal for current research in timbre perception is the development of a methodology for the analysis and synthesis of natural tones which is able to artificially generate tones that are effectively indistinguishable from original tones. The most productive work in detailing the salient physical factors in tones cited in the literature of timbre perception has been that which was based in some sort of analysis-synthesis strategy [Luce, 1963; Luce and Clark, 1965; Risset, 1966; Strong and Clark, 1967a, 1967b]. Until such strategies were evolved, the researcher could make only the most general sorts of statements, such as, *timbre depends primarily upon the spectrum of the stimulus, but it also depends upon the waveform, the sound pressure, the frequency location of the spectrum, and the temporal characteristics of the stimulus.*

Many contemporary investigators desire somewhat more than a speculative enumeration of the possible physical dimensions of tone upon which timbre is dependent, but rather desire to make some detailed statement about those specific factors which are important and how they might interact, on a case by case basis. In essence, the next step in timbre research might be the development of a *distinctive features* model analogous to that in speech perception, where the perceptually salient acoustical factors of tone are systematically uncovered and related to a corpus of sounds.

It is clear from a review of the literature that among the closest approaches to such distinctive feature models for timbre, most have been heavily dependent upon computer analysis and/or synthesis of natural tones. First, the analysis process reveals to the investigator the physical properties of any specific tones which may be used in perceptual research, in levels of detail not previously obtainable. This is absolutely necessary in attempting to determine those physical features of tone which are important in perception, for until one can specify the physical features in sufficient detail, no psychophysical correlation is possible.

Secondly, the analysis process should in the ideal case lend itself to a synthesis of the natural tones. In other words, a necessary test of the analysis technique is that it could provide information for a re-synthesis of an analyzed tone such that the synthesized tone would be *indistinguishable* from the original tone. Given this level of perceptually measurable success for the analysis procedure, the researcher has at hand a very powerful and rich tool for the further investigation of critical relationships between the physical and perceptual properties of timbre. This is especially important since the perceptual research may then be based upon the synthesized tonal stimuli rather than the original tones, and therefore the stimuli may be manipulated in any of a number of fashions by the experimenter for the purposes of pinning down the critical physical dimensions in detail.

The above dependencies for extended timbre research upon the computer analysis and synthesis of musical tones make the development of perceptually validated analysis-synthesis strategies a primary goal for research. This means that the *perceptual measure* of indistinguishableness between the product of analysis-synthesis and the original tone upon which that product was based must be applied to any prospective strategy, a measure which has not been successfully applied with any rigor in the past.

The strategy which we desire to test here is analysis-based additive synthesis. The analysis is made by way of the heterodyne filter which, when centered on the fundamental of a complex tone, produces time-variant amplitude and frequency/phase functions for each harmonic. The heterodyne filter is briefly explained in Appendix A. For a full description of the heterodyne filter see Moorer [1973], and for a discussion of how this method of analysis relates to previous similar techniques see Moorer [1974].

*Simplification of the Complexity of Physical Information in Timbre*

Once a perceptually validated analysis-synthesis method has been obtained for timbre, there remains the problem of the vast complexity of physical parameters given by the analysis. The next important goal for research, therefore, is the development of a strategy for manipulation of the very complex physical factors of tones such that they might be *systematically* simplified and related to perception. The potential wealth of information to be obtained with this approach has been indicated by earlier studies based upon this principle [Risset, 1966], where the salient cues for a certain type of timbre were specified in detail by an elimination of those factors which were not perceptually potent. Note that the end goal for such a data reduction is a simplified representation of timbre which is capable of generating a tone that is virtually indistinguishable from the original tone upon which the analysis was applied.

A second advantage to obtaining a simplified representation of tonal material is that it becomes much easier to manipulate timbres in various other ways. Work cited in the literature [Strong and Clark, 1967a, 1967b] has indicated the promise of this approach to the investigation of the interactivity of critical cues for tones. This has been attempted by the alteration of physical parameters controlling synthesis and by the direct exchange of cues from one sort of instrumental tone with those of a different instrumental source.

Further types of manipulation will be suggested below, where the goal is to produce a continuum of tones which are interpolations between two given tones. The production of such an interpolation of tones only becomes feasible given a simplified, manageable physical representation of timbre which is effectively perceptually indistinguishable from the original tones.

Following the successful data reduction of Risset, where small numbers of line segments were substituted for the complex time-variant functions of analyzed trumpet tones, we have attempted in our current research to extend this type of reduction to a wide range of timbres.

*General Explorations of Timbre Perception using Multidimensional Scaling*

A major aim of research in timbre perception is the development of a general model for the salient dimensions or features of classes of sounds. Given that timbre is clearly a *multidimensional* attribute of sound, the computer-based techniques of perceptual data analysis which fall under the category of multidimensional scaling seem especially well-suited for examining the complex aspects of timbre perception [see Appendix C for a brief description of the analysis techniques which will be utilized in our research]. Among the more recent psychological investigations of timbre perception, those which have been oriented to the use of multidimensional scaling have been the most productive in reducing the complexities of timbre to a small set of dimensions which were interpretable in terms of the physical factors composing them [Plomp, 1970; Wessel, 1974].

The spatial analogy implicit in multidimensional scaling techniques, whereby a *psychological distance* function is mapped into a geometric space of low dimensionality, seems appropriate for modelling musical timbre. The concept of the relative distances of musical tones, in terms of perceptual contrast or dissimilarity, is not a foreign one in the listening experience of the musician.

In addition, the ability to generate a data structure on the sole basis of perceptual judgments, which then may be interpreted with respect to the physical parameters of the sounds, is especially appealing with stimuli so complex as natural timbres. It has become quite clear that attempts to deal with timbre on the basis of *a priori* assumptions of the importance of one or another physical component of sound will generally not succeed. Thus, to start with a subjective data structure is most advantageous.

That the perceptual judgments are generalized subjective distance reports, rather than some more specific verbal labeling, as with semantic differential measurements, is also to great advantage in dealing with timbre. As mentioned above, the use of verbal rating scales to classify timbres is a very precarious if not a dubious tactic.

We therefore will pursue the type of perceptual experimentation based on similarity judgments which are analyzable with multidimensional scaling algorithms with the hope that it will provide the most powerful and direct method for exploring the many factors involved in timbre perception. It will also provide a means for evaluating the possible effects of musical training on perception, in that individual differences scaling techniques provide information regarding the relative behavior of the individual listeners on the analyzed dimensions of perception [Carroll and Chang, 1970].

*Examination of the Continuous versus Categorical Nature of Timbre Perception*

A final goal of current research is an examination of the nature of the timbre space in terms of its *continuity*. The existence of categorical perception with timbre is a controversial issue at present, and remains to be tested for tonal material as complex as natural timbre. Even in the speech perception literature, categorical perception is still a somewhat unsettled issue [Lane, 1965; Studdert-Kennedy, 1970; Pisoni, 1974].

If perception were *continuous* in the timbre space, then a set of acoustical stimuli which were spread along some physical dimension or set of dimensions at *equal* intervals between the extremes would be heard as a continuous succession of sounds, gradually changing from one extreme to the other. Conversely, if the stimuli were perceived *categorically*, then, rather than hearing a gradual interpolation between the two extreme stimuli, the listener would hear the first extreme suddenly switch to the second, the tones in between taking the identification of one of the two extremes with little or no ambiguity even in the center of the transition.

In addition, the discriminability of the successively placed signals at *acoustically equal* intervals would be dramatically affected by the existence of a categorical mode of perception. If the interpolation between extremes were continuous, then all intervals should be equally discriminable. However, if there existed a sharp boundary where one extreme seemed to jump to the other, categorically, then a different discrimination function could be expected: a pair of signals that fell within the same category would be much *less* discriminable than a pair of signals, at the same interval, which came from two different categories.

The most extensive work, and indeed, much of the original work, on categorical perception has been done in the field of speech research [e.g. see Liberman, et. al.,1967]. A general finding seems to be that categorical perception is strong with consonants, but weak with vowels. This has lead to speculations that speech perception involves a decoding of information relating to speech production, in that consonants are formed in a more discrete manner, with no possible physical interpolation between two given values, while vowels are produced on much more of a continuum which does allow for transitions between any two values. In that there exists some similarity between the transients of consonants and steady-state of vowels in speech, and the attack and decay of a musical note surrounding some steady-state condition, it seems appropriate to examine the timbre space in terms of categorical perception. Cutting and Rosner [1974] found categorical perception for adult listeners with sawtooth-wave stimuli, differing in rise time, heard as plucked or bowed notes. Interestingly enough, they later found categorical perception in two-month-old infants as well, and they interpreted the results as undermining the basic notion of categorical perception [Jusczyk, et. al, 1974].

The development of algorithms for the interpolation of a set of tones between two given

naturalistic end-points will be necessary in order to examine the existence of categorical perception with timbre. The speech research at Haskins Lab, referred to above, has employed a simple model for analysis and synthesis in which stylized time-variant formant regions and bandwidths of noise compose the synthesis data. Due in part to the simplicity of this synthesis model and, in the main, to several different sets of phonemes that were found to be identical in all but a single, easily-represented parameter for synthesis, many interpolations in the speech domain were readily discovered.

In the domain of musical tones, however, the use of analysis-based additive synthesis has not revealed such simple relationships between different instruments. A general interpolation algorithm has been applied to the individual time-variant amplitude functions for synthesis, after they have been approximated by small numbers of line segments, and the individual frequency functions, after they have been approximated by constant, non-varying values. This algorithm attempts to: 1) interpolate the maximum of an amplitude function temporally as well as in amplitude between the two end-point functions; 2) perform a weighted average on the activity before the maxima of the two end-point functions to generate the attack segment of the interpolated tone. If a delineated segment of activity occurs in the initial part of the attack, it will be handled as precedent activity to the constructed function, and will be scaled in magnitude and duration with respect to the weighting - this provides for the tones which have segmented precedent activity in many harmonics; and 3) perform a similar weighted average on the activity following the maxima to generate the decay segment (see Figure 14).

It is worth noting that it is only possible to test the effects of a given interpolation algorithm, and that there is certainly more than one way to generate a transition between two end-point tones. Therefore, an *ultimate* proof of the existence of categorical perception with musical timbre is impossible, in that any specific measurement relates to just that algorithm which generates the transition. This fact has been ignored in speech research, although it remains just as applicable. Any conclusion resulting from the use of transitions in multidimensional physical space, as is the case proposed and in speech work, must be tempered by the fact that there will be no *unique* manner of transition.

## B. Problems which Occur in Timbre Research

### *Specification of the Physical Properties of Timbral Stimuli*

A primary problem in most psychological studies of timbre perception to date has been the lack of an adequate specification of the physical properties of the stimuli. Typically information gleaned from speech spectrographs for steady-states has been utilized. This has placed severe restrictions on the interpretability of results in terms of psychophysical correlations, and has undoubtedly contributed to the great reluctance of most investigators for dealing with anything other than steady-state signals. As a result, timbre research has made few advances beyond the classical theory of Helmholtz, and the most widespread definition of timbre has it that it

essentially consists of the steady-state spectrum of a tone. This ignores the temporal information which has been so strongly implicated in the identification of timbre.

It is clear that, with the computer-based analysis-synthesis strategies which have been developed in the last decade, timbre research will take a great step forward. Now it becomes possible to obtain detailed information about the physical characteristics of tones, as modelled by any specific analysis technique, which will include the time-variant nature of the spectral information - how the harmonics change with time in both amplitude and frequency. In addition, if there is any question of the adequacy of analysis to give the complete picture of timbre, it remains possible to *re-synthesize* a tone on the basis of the analysis, and therefore to know precisely the totality of physical factors which compose that tone. If the synthetic tone is indiscriminable from the original, then the most stringent test of the value of any proposed analysis-synthesis technique has been met and no perceptual information has been lost.

A potential problem facing the researcher armed with such analysis-synthesis techniques is the usually overwhelming complexity of the analysis. As mentioned above, any data simplification which does not violate the perceived qualities of the synthesized tones would be of enormous benefit to the researcher. Simplified representations of tones eliminate non-essential physical factors and help reduce the number of possible acoustical cues operating in a given perceptual judgement.

*Control through Equalization of Pitch, Loudness and Duration*

While finding no psychophysical investigations which have directly assessed the influences of pitch, loudness, or duration on the perception of timbre, there have been studies which indirectly indicate that these parameters do interact with timbre. In addition, informal studies at our laboratory have shown that slight inequalities in these three physical factors will be heard as inequalities in timbre by most listeners, even trained musicians.

Let us first consider pitch. It is well known that the tonal quality of a pure tone varies with pitch [Stevens and Davis, 1938]. Investigations on the timbral similarities between complex tones having different pitches have invariably found significant perceptual changes in timbre with good-sized steps in pitch [Slawson, 1968; Plomp and Steeneken, 1971]. Both the saliency of the pitch dimension and its exclusion in the definition of timbre strongly advises that some attempt be made for pitch equalization in a study of the perceptual structure of timbre - one which is not specifically interested in examining the interaction with pitch, that is.

Observations which we have made when comparing the timbre of two tones, that were either approximately or actually equal in all other dimensions, have revealed that very slight mistunings of the tones will be heard as a difference in timbre, for example, in *brightness* versus *dullness*.

Let us next consider the effects of loudness on timbre. Measurements of the physical intensities of tones which were played by musicians who were striving for equal loudness have shown wide discrepancies in the levels of adjacent notes [Winckel, 1967]. This phenomenon has been found for all of the musicians which we recorded, and seemed to relate to the resonance properties of the instruments played. If there are no wide variations in the perceived loudness for such sets of notes, then we might conclude that there is a certain range of acceptable and generally unnoticeable variation in the physical intensities of musical tones, both for the performer and the listener. However, this type of observation does not provide a sufficient rationale to ignore the dimension of loudness in a controlled experiment on timbre perception.

In fact, we have made some informal observations that suggest that the effects of loudness on timbre perception are very pronounced. In a test to discriminate between two tones having equivalent spectra and approximately equal temporal patterns, it was found that if the tones were first carefully balanced using a VU meter they became much less discriminable. The point is, in their unbalanced state, the differences were heard to be definite differences in *timbre*, and not differences in *loudness*. The observation on the effects of pre-balancing the tones was made after speculating that the different readings of the tones on the meter might have something to do with their unexpected differences in tone quality.

Experiments have been done on the perception of loudness for tones which vary in spectral and onset characteristics [Gjaevenes and Rimstad, 1972]. It was found that the loudness of a tone was dependent on its spectral composition as well as its attack time. Again, this provides no direct evidence for the interaction of timbre perception with the loudness of a tone, but it does show that the loudness of a tone can be influenced by certain parameters which are also involved in the perception of timbre. While we have certain proof that some degree of variation will not significantly affect the perception of timbre in musical situations, we do not have sufficient data to state what the tolerable range of such variation might be in an experimental context. In view of this lack of explicit information, and also in view of the most accepted definition of timbre, it seems wise to equalize stimuli for loudness as well as for pitch.

Finally, let us briefly consider the possible effects of the duration of tones on their perceived timbre. Interactions have been discovered between duration and various other dimensions in auditory perception. The loudness of tones which have the same physical energy has been found to be a function of duration for pure tones under one second in length [Plomp and Bouman, 1959]. There may be some effect on the perceived timbral relationships between two tones if they vary in duration. Studies on the perception of speech have indicated that duration serves as a cue in the recognition of synthetic vowels [Ainsworth, 1972]. This finding gives evidence of an interaction between duration and timbre. We must conclude that it seems appropriate to control the duration of stimuli which are used to study timbre perception.

Considering the range of durations that we plan to use in these studies, on the order of one-half a second, it would seem especially critical to equalize for perceived duration. Most of the

interactions between other dimensions and duration are quite strong in that range. Our own informal listening experiences with tones that naturally varied in duration in that range dictated an attempt to equalize their duration before utilizing them in experiments where similarity is to be measured.

It becomes clear that a major advantage in using computer-synthesized tones for stimuli is the ability to *independently* manipulate the frequency, intensity and physical duration of the signals. However, no unequivocal technique now exists for performing an *a priori* equalization of either the pitch, the loudness, or the perceived duration of complex, time-variant, natural tones. We therefore must employ an empirical method to perform such equalization. It remains possible to actually compute a set of tones which vary incrementally and independently along the three physical continua of frequency, intensity and duration. Having such a set of tones, listeners could attempt to match the three respective perceptual qualities of pitch, loudness and duration of some given tone to a standard comparison tone. Not only would a set of normalized stimuli be obtained, but valuable data would be collected for mathematically modelling the influences of spectral and temporal factors on the perception of pitch, loudness and duration.

*The use of Isolated Tones to Investigate Timbre Perception*

While the ultimate test of timbre perception would be in some musical context, the use of isolated tonal stimuli may be sufficient to deal with certain, limited domains of perception. This sufficiency, of course, is yet to be tested. However, the existence of analytic tools for dealing with melodic phrases of notes is still in the future. We have no tools to precisely determine the totality of acoustical events which exist between connected notes for different instruments. Perhaps empirical rather than mathematical research will uncover methods for the adequate simulation of connected musical lines. This should be a primary goal for future research but it remains an insurmountable problem for current research.

At the present, the researcher is still involved with the development and perceptual measurement of analysis-synthesis techniques which work with single isolated tones. This concern also includes possible data reductions which can be imposed upon analyzed tones in the hopes that such reductions may eventually also have some bearing on the ability to connect notes in musical phrases.

# III. EMPIRICAL RESEARCH ON TIMBRE

*Outline of Current Research*

The work which follows employed an analysis-based additive synthesis technique for the computer generation of musical timbres. The analysis of natural tones is performed by the heterodyne filter, which derives time-variant functions detailing the amplitudes and frequencies of each harmonic of a complex tone. These functions are then used to control the respective amplitudes and frequencies of sinusoids, which are summed together to replicate the original tone by additive synthesis. The reader is advised at this time to see Appendix A for a brief explanation of the analysis techniques and Appendix B for a description of the graphical displays which will frequently be used in presenting the results of these analyses.

Four basic phases of research are described:

A) a measurement of both the *discriminability* and the *perceptual distance* between tones produced by the analysis-synthesis and data reduction techniques and the original tones upon which they were based. This constitutes a measurement of the degree to which the analysis-synthesis technique is able to replicate perceptually 16 original tones from various musical instruments, and a measurement of the perceptual adequacy of a number of data reductions of the complex results of analysis, when such simplified functions are used for synthesis.

B) a *perceptual equalization* of the 16 data reduced synthesized tones which will be utilized in further experimentation. They are equalized empirically in the perceptual dimensions of pitch, loudness and duration by a matching procedure.

C) a collection of *perceptual similarity* and *identification confusion* matrices, which relate the subjective distances of all 16 tones to one another. Multidimensional scaling techniques are employed to present the similarity-based distance structures in low dimensional Euclidean space, and interpretation is attempted in terms of the known physical attributes of the synthesized tones. Confusions in learning to identify the 16 tones by given names are examined with reference to the space.

D) an exploration of the *perceptual continuity* of the timbre space by the utilization of interpolated series of tones between given end-points chosen from the 16 data reduced tones. Four types of measurements were taken:

*1) perceptual cross-over points* in identification for sequential interpolations between end-points, testing for the extent of directional differences in the points at which one tone perceptually turns into the other, and hence examining the width of categorical boundaries in the sequential condition of presentation;

*2) identification* and *goodness ratings* for isolated interpolated tones, where the identification of a tone is multiple-choice from a list of 10 alternatives, and the goodness measure for a tone is its perceptual closeness to the non-interpolated tone of the type so identified - examining the cross-over points and sharpness of categorical boundaries for the isolated mode of presentation;

*3) discrimination* and *perceptual distance* functions for pairs of interpolated tones, to study the effects of categorical boundaries on discrimination, and to correlate with the distances of the end-points in the derived timbre space for the 16 synthesized natural tones;

*4) perceptual similarity* measurements to derive the subjective distances for a subset of 12 out of 16 of the natural tones, plus 6 interpolated midpoints, again for multidimensional scaling, to assess the contextual effects of the interpolated material in a larger set of natural tones.

## A. Discrimination and Subjective Distance Estimation between Original, Re-Synthesized and Data Reduced Tones

### *Introduction*

The use of computer analysis and synthesis of natural timbres is an important step in the direction of being able to investigate the perception of complex musical tones. Among the several recent attempts to analyze and re-synthesize musical tones [Luce, 1963; Freedman, 1967, 1968; Beauchamp, 1969], though success was reported in beginning to be able to faithfully replicate the original tones, no rigorous perceptual measurements were taken. In this study we have obtained discrimination functions and subjective distance estimations between the original and re-synthesized tones derived from analysis-based additive synthesis. The method of analysis is the heterodyne filter [Moorer, 1973; Appendix A]. Time-variant amplitude and frequency functions are obtained for each harmonic of a complex tone, and the waveform is re-synthesized by summing a set of harmonic sinusoids which are controlled in time by the analyzed functions. Processed in this way for perceptual measurements were 16 notes of short duration, taken from the brass, string, and woodwind families of instruments.

In addition, any significant data reduction in the complex results of analysis is an invaluable aid to the researcher attempting to pinpoint relevant physical cues to specific perceptual judgments. In fact, the process of simplification itself, in eliminating unessential physical components derived from the analysis, narrows in on those factors relevant to perception. The value of data reduction has been shown in the work of Risset [1966] for trumpet tones, and we have pursued a similar line of data simplification for the other timbres listed below. These data reduced tones were added to the set of original and re-synthesized tones for pairwise discrimination and subjective distance estimation. Various levels of data reduction were be studied.

*Stimuli*

The stimuli were derived from 16 instrumental notes played near the pitch of Eb above middle-C, approximately 311 Hz, whose durations ranged between 280 to 400 milliseconds. The tones were performed and recorded in an IAC acoustical isolation chamber, which is a very dry environment, but passes some small amount of reverberation. The recording was done on a Revox tape recorder, half-track monaural at 7.5 ips, using Scotch 206 1/4-inch low-noise tape. They were then digitized by an Analogic 14-bit analogue-to-digital converter at a 25.6 KHz sampling rate and the high-order 12 bits were stored in digital form for computer analysis or playback. The 16 instrumental tones consist of:

> oboes *(2 different instruments and players)*, English horn, bassoon, Eb clarinet, bass clarinet, flute, alto saxophones *(2 tones from one instrument, played at p and mf)*, soprano saxophone, trumpet, French horn, muted trombone *and* celli *(3 tones from one instrument, played normally, muted sul tasto and sul ponticello)*.

Each of the 16 instrument notes appeared in at least 4 of the 5 following *tonal conditions*:

> 1) the original tone in the state of digitization. We shall call this the *original* tone.
>
> 2) a tone which was the product of analysis-based additive synthesis, using the complete results of the analysis process; that is, the complex amplitude and frequency functions derived from the heterodyne filter for each harmonic were used to control the parameters of a set of sinusoids which were then added together to re-synthesize the tone. We shall refer to this as the *re-synthesis* or the *complex synthesis*, and it is illustrated for a particular tone in Figure 1a in a perspective plot of Time x Amplitude x Frequency and in Figure 1b in a spectrographic display [see Appendix B for an explanation of these graphical techniques].
>
> 3) a tone which was produced by using approximations to the complex time-variant amplitude and frequency functions with small numbers of line segments; between 4 to 8 segments were used per function, and an attempt was made to model any activity in the attack that was analyzed, including initial, low-amplitude inharmonicity. We shall call this the *line segment approximation* to the original, and it is illustrated for the above tone in Figures 2a and 2b.
>
> 4) a tone which excluded any clearly delineated initial segments of low-amplitude inharmonicity in the attack; the resulting tone was similar in all other respects to the condition of data reduction immediately above. This condition was applicable only to 9 of the 16 tones: 2 oboes, Eb clarinet, bass clarinet, 3 saxophones, trumpet and French horn. We shall label this the *cut-attack approximation*, and it is shown for the above tone in Figures 3a and 3b.

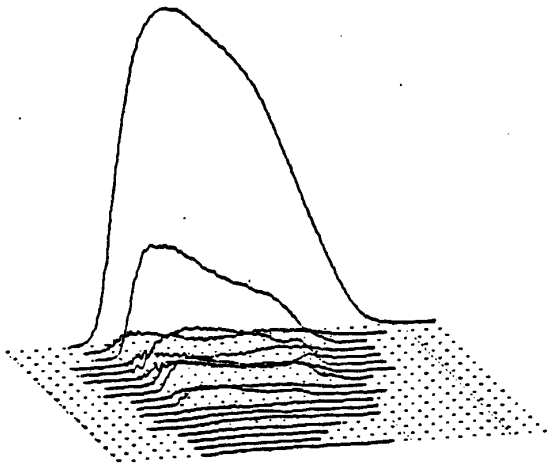


FIGURE 1a: Complex functions for re-synthesized tone, shown as an Amplitude x Frequency x Time perspective plot (X = time; Y = amplitude; Z = frequency, with the fundamental harmonic plotted in the background).

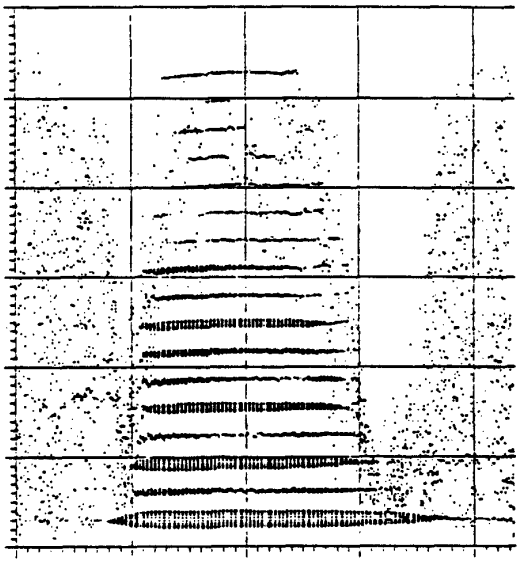


FIGURE 1b: Complex data for the re-synthesized tone, shown above in the form of a spectrographic plot (X = time, with 1/10 second lines; Y = frequency, with KHz lines; Width of bars = relative dB to -40).

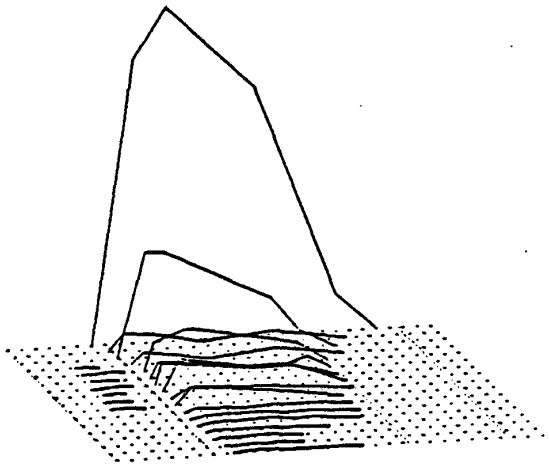


FIGURE 2a: Line segment approximations for synthesis, shown as an Amplitude x Frequency x Time perspective plot (X = time; Y = amplitude; Z = frequency, with the fundamental harmonic plotted in the background).

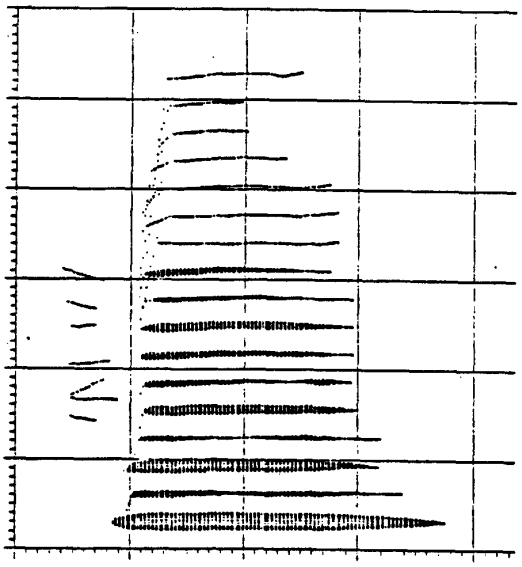


FIGURE 2b: Line segment approximations for synthesis, shown above in the form of a spectrographic plot (X = time, with 1/10 second lines; Y = frequency, with KHz lines; Width of bars = relative dB to -40).

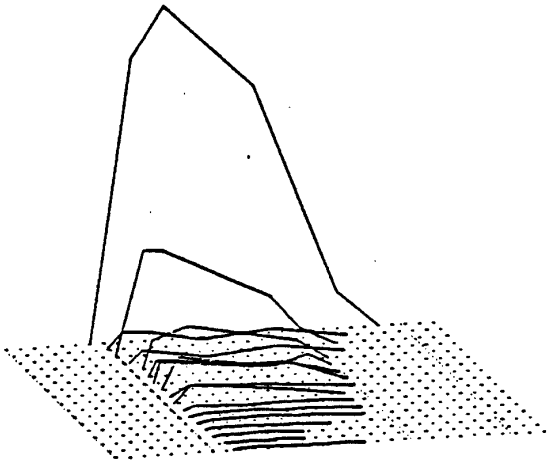


FIGURE 3a: Cut-attack approximations for synthesis, shown as an Amplitude x Frequency x Time perspective plot (X = time; Y = amplitude; Z = frequency, with the fundamental harmonic plotted in the background).

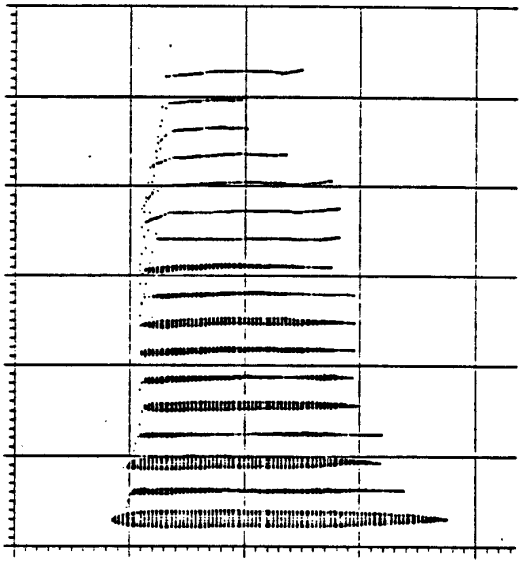


FIGURE 3b: Cut-attack approximations for synthesis, shown above in the form of a spectrographic plot (X = time, with 1/10 second lines; Y = frequency, with KHz lines; Width of bars = relative dB to -40).

5) a tone which substituted constant frequencies for the time-variant frequency functions, but used the same time-variant line-segment approximations to the amplitude functions as in case (3) above. We shall call this the *constant frequencies approximation*, and it is shown for the above tone in Figures 4a and 4b.

The stimuli were recorded on tape after being played back from the computer through a 12-bit digital-to-analogue converter at a 25 KHz sampling rate. The recording was done on a Sony 854-4 tape recorder, quarter-track monaural at 7.5 ips, on Maxell U50 1/4-inch low-noise tape. In cases (2) through (5), a small amount of noise was added to the signal, which was derived from blank tape hiss, in order to simulate the existing tape noise on the original signals. (It is important to note that this simulation was not entirely adequate, and enhanced discriminability was expected between the original condition and any other synthesized condition. Due to the lack of an isolation chamber in sufficient proximity to the computer, the original tones had to be tape recorded first; otherwise we would not have run into this methodological problem).

Playback was accomplished by means of a Dynaco PAT-4 preamplifier and Dynaco Stereo 120 amplifier connected to an Altec 804-E loudspeaker, and was done in a moderately reverberant room approximately 12 x 12 feet in dimensions. Playback level was measured at a distance of 3 feet from the speaker with a B&K sound level meter for the Eb clarinet tone and this data is given in Table 1a. This level was generally a comfortable, moderate level for all listeners.

*Listeners and Procedure*

Sixteen listeners at Stanford University were employed for this experiment. Listeners were musically sophisticated, some actively involved in advance instrumental performance and others in conducting or musical composition. Several listeners had much experience with the production of computer music, and had well-trained ears with respect to synthesized timbres.

Listeners were placed 10 feet from the speaker, directly facing it. This arrangement was made on a diagonal across the room. Listeners came for two hour-long sessions, and took a total of 573 trials. In the first session, the first 30 trials were practice runs, while in the second session, the first 15 were practice.

The trials were structured in an AAAB discrimination paradigm. On each trial four tones were played: three were *identical* and the fourth was *different*. Of course the position of the *different* tone was randomly selected. The first and last two tones were separated in times of onset by 1 second, while the second and third tones were separated by 2 seconds, giving the impression of two *pairs* of tones played in succession. Each trial was preceded by a low-frequency knock-like tone lasting for 50 milliseconds.

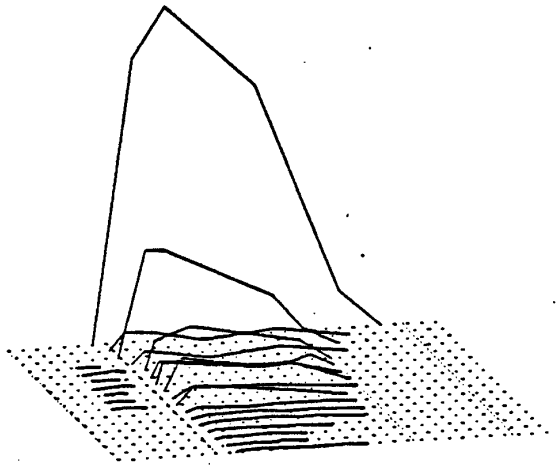


FIGURE 4a: Constant frequencies approximations for synthesis, shown as an Amplitude x Frequency x Time perspective plot (X = time; Y = amplitude; Z = frequency, with the fundamental harmonic plotted in the background).

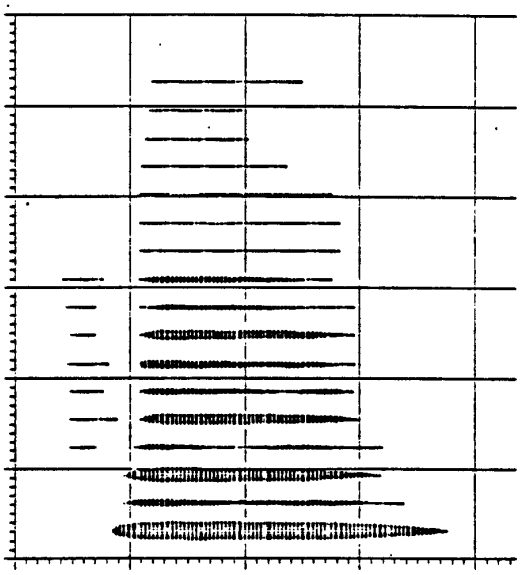


FIGURE 4b: Constant frequencies approximations for synthesis, shown above in the form of a spectrographic plot (X = time, with 1/10 second lines; Y = frequency, with KHz lines; Width of bars = relative dB to -40).

| | List A | List B |
|---|---|---|
| *Filter (Hz)* | S+N / N (dB) | S+N / N (dB) |
| *125* | 67 / 50 | 59 / 42 |
| *250* | 98 / 48 | 91 / 37 |
| *500* | 85 / 41 | 76 / 31 |
| *1K* | 84 / 37 | 78 / 27 |
| *2K* | 88 / 29 | 79 / 26 |
| *4K* | 78 / 26 | 66 / 24 |
| *8K* | 55 / 24 | 42 / 24 |
| *16K* | 34 / 31 | 28 / 28 |

TABLE I (Lists A & B): Measurements taken of Eb Clarinet (CLAR 1) by B&K sound-level meter in two experimental laboratories. Levels were measured in dB with octave-band filters (center frequencies are given here in the left-hand column) and presented for the conditions of signal+noise (S+N) and noise alone (N).. List A applies to the experiments discussed in setions A, C and D, while List B applies to the experiments discussed in section B.

There was 2 seconds of silence between the warning knock and the first tone of the trial; 4 seconds of silence followed each trial before the warning knock for the next trial, in which listeners could make their responses. Trials were presented in a random order.

All possible pairwise combinations of the applicable tonal conditions per instrument note were used to compose the trials, so there were 10 basic pairwise trial-combinations for the 9 instrument notes having 5 tonal conditions, and 6 basic pairwise trial-combinations for the other 7 instrument notes having only 4 tonal conditions. This made a total of 132 different trial-combinations. Each possible trial-combination was randomly repeated four times, making the total of 528 non-practice trials in the two sessions.

Listeners were told to rate how *equally* a sequence of 4 notes appeared to be played. *Equal* was defined in terms of the notes having *identical qualities of articulation*, or playing style. They were told to consider the 4-note sequence as *two pairs* of notes separated by a longer rest. They were further told that often there would be *one note* which was *different* from the others, and if they detected such a different note, they were to: 1) indicate in which pair the note was located, and 2) rate its degree of difference from the other notes, relative to the differences heard in all other trials, from 1 to 10. The first judgment was a measure of *discrimination* between different combinations of stimulus types, and the second judgment was a *distance estimation* of the differences heard between stimulus types.

*Results*

The collected data were averaged over the four repeated measurements per trial-combination for each listener. The averaged discrimination score equalled the number of times the different pair was correctly identified plus 1/2 times the number of times there was no answer on a trial, this sum divided by four. The averaged distance estimation equalled the sum of the estimates on the correct trials, divided by the number of correct trials in that trial-combination. The mean subject scores were then averaged to obtain mean scores for discrimination and distance estimation per trial-combination, and these scores were used in the computations which follow.

The mean discrimination scores appear in Table 2 and the mean distance estimates are given in Table 3. The correlation between the discrimination scores and distance estimations per trial-combination was .92. This is presented graphically in Figure 5, in which the ordinate represents the distance estimates and the abscissa represents the discrimination scores.

The distance estimates were then treated with a multidimensional scaling algorithm which takes individual differences in input matrices into account during the computation [Carroll and Chang, 1970; see Appendix C]. A unique rotation is derived for the final solution, determined by maximizing the variance accounted for by the axes, and a weight space is also derived, into which the individual input matrices are represented by their respective weights on the dimensions of the stimulus space.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| TRP | .727 | | | | BSN | .594 | | | |
| | .742 | .586 | | | | .617 | .656 | | |
| | .867 | .844 | .766 | | | .836 | .859 | .688 | |
| | .898 | .797 | .617 | .570 | | — | — | — | — |
| FHRN | .531 | | | | CLAR 2 | .500 | | | |
| | .648 | .563 | | | | .891 | .680 | | |
| | .703 | .688 | .602 | | | .875 | .828 | .758 | |
| | .750 | .531 | .547 | .602 | | .914 | .828 | .742 | .859 |
| TRB | .672 | | | | EHRN | .727 | | | |
| | .609 | .531 | | | | .750 | .570 | | |
| | .727 | .672 | .648 | | | .867 | .648 | .625 | |
| | — | — | — | — | | — | — | — | — |
| CEL 1 | .781 | | | | SAX 1 | .641 | | | |
| | .906 | .805 | | | | .883 | .805 | | |
| | .914 | .828 | .633 | | | .930 | .898 | .727 | |
| | — | — | — | — | | .938 | .742 | .703 | .805 |
| CEL 2 | .680 | | | | SAX 2 | .742 | | | |
| | .906 | .742 | | | | .898 | .695 | | |
| | .938 | .820 | .680 | | | .984 | .844 | .656 | |
| | — | — | — | — | | .961 | .883 | .711 | .766 |
| CEL 3 | .797 | | | | SAX 3 | .828 | | | |
| | .836 | .570 | | | | .930 | .578 | | |
| | .906 | .641 | .563 | | | .999 | .891 | .922 | |
| | — | — | — | — | | .883 | .742 | .570 | .930 |
| OBOE 1 | .617 | | | | FLUTE | .672 | | | |
| | .641 | .477 | | | | .859 | .563 | | |
| | .664 | .633 | .641 | | | .930 | .758 | .711 | |
| | .820 | .734 | .719 | .859 | | — | — | — | — |
| OBOE 2 | .648 | | | | CLAR 1 | .648 | | | |
| | .625 | .602 | | | | .617 | .602 | | |
| | .719 | .703 | .641 | | | .844 | .742 | .648 | |
| | .617 | .547 | .578 | .734 | | .820 | .789 | .664 | .711 |

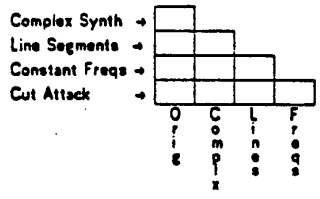



TABLE 2: Mean Discrimination Scores [section A] for tonal conditions presented with 16 tones. One matrix appears for each tone labelled above. 5 x 5 matrices are given for those 9 tones which had all 5 tonal conditions (see text), while 4 x 4 matrices appear for the 7 tones that had only 4 tonal conditions. The form of each matrix in terms of the pairs of tonal conditions judged is given in the grid immediately above.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| TRP | 2.746 | | | | BSN | 2.479 | | | |
| | 3.713 | 2.280 | | | | 2.002 | 2.368 | | |
| | 5.567 | 4.006 | 3.152 | | | 4.267 | 3.396 | 3.125 | |
| | 4.689 | 3.148 | 2.720 | 1.838 | | — | — | — | — |
| FHRN | 1.695 | | | | CLAR 2 | 1.250 | | | |
| | 2.577 | 2.047 | | | | 4.242 | 2.950 | | |
| | 3.297 | 2.831 | 1.836 | | | 4.577 | 4.479 | 3.308 | |
| | 2.740 | 1.594 | 1.440 | 2.104 | | 4.657 | 4.274 | 3.265 | 4.292 |
| TRB | 2.463 | | | | EHRN | 3.107 | | | |
| | 2.479 | 1.597 | | | | 3.780 | 2.116 | | |
| | 3.508 | 2.757 | 2.733 | | | 4.511 | 1.909 | 1.947 | |
| | — | — | — | — | | — | — | — | — |
| CEL 1 | 3.729 | | | | SAX 1 | 1.957 | | | |
| | 4.319 | 3.249 | | | | 4.965 | 3.825 | | |
| | 5.247 | 3.774 | 2.432 | | | 5.849 | 4.783 | 2.908 | |
| | — | — | — | — | | 4.931 | 3.463 | 2.894 | 3.169 |
| CEL 2 | 3.528 | | | | SAX 2 | 3.793 | | | |
| | 5.229 | 4.494 | | | | 5.626 | 3.231 | | |
| | 6.597 | 4.603 | 3.226 | | | 6.943 | 3.818 | 3.247 | |
| | — | — | — | — | | 6.309 | 4.921 | 3.121 | 3.170 |
| CEL 3 | 3.297 | | | | SAX 3 | 4.568 | | | |
| | 4.797 | 2.267 | | | | 5.516 | 1.783 | | |
| | 3.940 | 2.591 | 2.451 | | | 7.512 | 5.590 | 4.511 | |
| | — | — | — | — | | 5.128 | 3.073 | 2.849 | 4.697 |
| OBOE 1 | 2.016 | | | | FLUTE | 3.237 | | | |
| | 2.046 | 1.515 | | | | 4.720 | 2.327 | | |
| | 2.635 | 2.047 | 2.510 | | | 5.488 | 3.260 | 2.671 | |
| | 4.018 | 3.025 | 3.441 | 4.130 | | — | — | — | — |
| OBOE 2 | 2.091 | | | | CLAR 1 | 2.657 | | | |
| | 2.130 | 1.657 | | | | 2.365 | 2.813 | | |
| | 3.013 | 2.492 | 2.473 | | | 4.036 | 2.641 | 3.058 | |
| | 2.405 | 2.235 | 3.163 | 2.895 | | 4.241 | 3.946 | 2.654 | 3.581 |

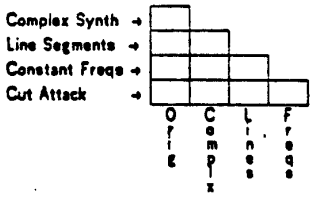



TABLE 3: Mean Distance Estimations [section A] for tonal conditions presented with 16 tones. One matrix appears for each tone labelled above. 5 x 5 matrices are given for those 9 tones which had all 5 tonal conditions (see text), while 4 x 4 matrices appear for the 7 tones that had only 4 tonal conditions. The form of each matrix in terms of the pairs of tonal conditions judged is given in the grid immediately above.

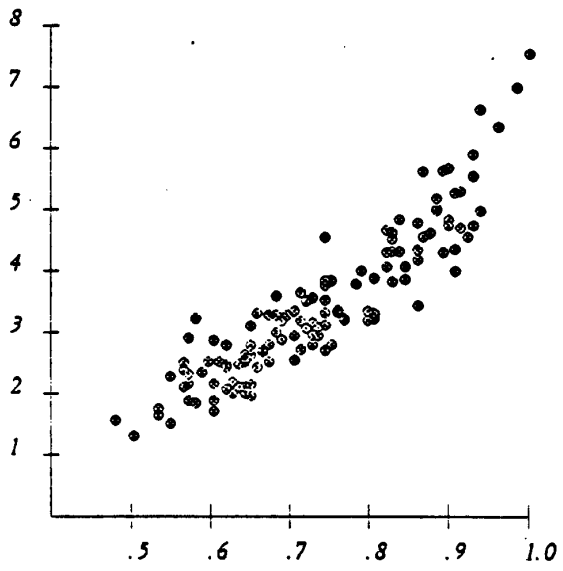



FIGURE 5: Discrimination Scores (X-axis) versus Distance Estimates (Y-axis) as obtained in section A. Mean values for each tonal condition are plotted (see text for explanation).

In the case of the data in this study, the individual input matrices represent the 4 x 4 or 5 x 5 distance estimation matrices for the 4 or 5 respective tonal conditions per instrumental note. There were 16 total matrices, one per instrumental note. A one-dimensional solution was obtained for the common 4 tonal conditions for all 16 instrumental notes. Consistent with the ordering of the degrees of modification in the 4 conditions, the ordering along the dimension was *original, re-synthesis, line segment approximation,* and *constant frequencies approximation.*

Somewhat more informative was the solution in two dimensions obtained for all 5 tonal conditions common to 9 of the instrumental notes. The stimulus solution is given in Figure 6. The vertical dimension is equivalent to the one-dimensional solution above, and is related to the degree of modification from the original tone. Note that the location of the *cut-attack approximation* with respect to the *line segment approximation* is consistent with this interpretation. The horizontal axis is interpretable with respect to the preservation of the low-amplitude inharmonicity in the initial attack segment. Apparently this segment of the attack was quite important perceptually, in that the *cut-attack approximation* tended to be more discriminable and was rated to be a greater distance from other conditions, on the whole. Table 4 gives mean discrimination scores and distance estimates derived from the 16 instrumental notes.

An examination of the weight space [see Appendix C], and a closer analysis of the input matrices, reveals that all 16 instrumental notes did not behave in exactly the same way with respect to the set of modification conditions imposed upon them. In that the matrices are small enough to allow direct interpretation, we will not refer to the weight space but rather look directly at the matrices themselves in order to obtain detail on specific interactions between tones and conditions of modification.

There is a wide range of variance between the 16 instrumental notes in those trials that paired the *original* tone with another tonal condition. In part, this must be related to the differential amounts of noise which came from the original tape and the process of digitization of the signal. For instance, the cello tones all especially suffered this noise, since they had been produced and recorded at a much lower amplitude. Attempts to simulate the tape noise by adding blank tape hiss to synthetic tones met only with partial success, and varied considerably between the 16 instrumental notes.

Comparisons of the differences in behavior for the 16 tones in cases which did not involve the *original* tone cannot be attributed to any additional differences in background noise, however. Looking at the input matrices, there are significant differences in the importance given to the *cut-attack* segment, and to the effect of the *constant frequencies approximation.* In the case of the *cut-attack approximation,* for instance, there was good discriminability and large distance estimates for the bass clarinet, while the French horn showed the opposite tendency. In the *constant frequencies approximation,* the soprano saxophone was dramatically altered, but the English horn showed little change.

.675

.772 .627

.856 .769 .682

.845 .733 .650 .760

DISCRIMINATION

2.788

3.782 2.532

4.812 3.436 2.849

4.346 3.298 2.839 3.320

DISTANCES

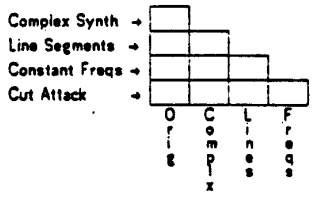



TABLE 4: Overall Discrimination Scores and Distance Estimations [section A] for tonal conditions averaged over all 16 tones. The form of the matrices in terms of the pairs of tonal conditions judged is given in the grid immediately above.

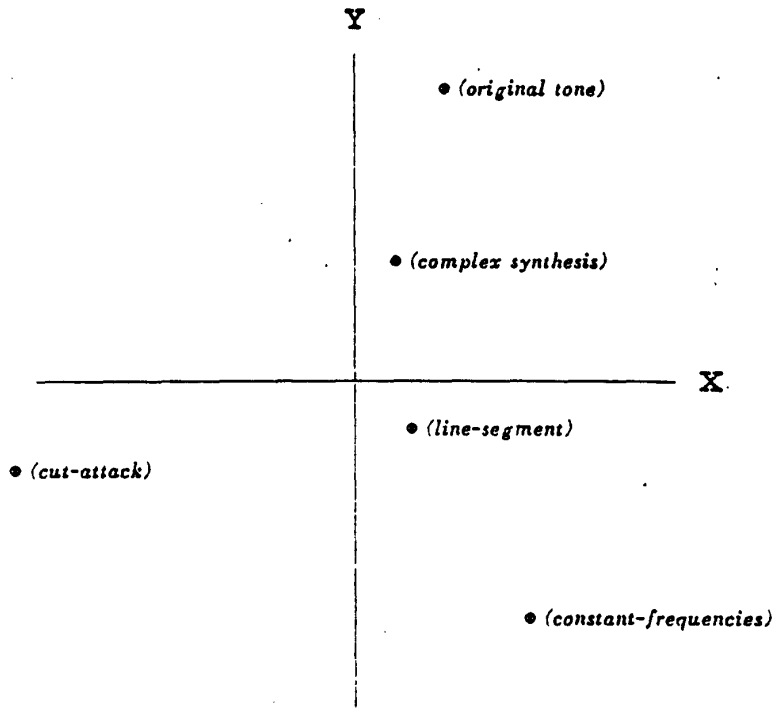



FIGURE 6: Two-dimensional scaling solution for Distance Estimates as obtained in section A. The five tonal conditions scaled are labelled above (9 tones served as the basis for this scaling). See text for an interpretation.

*Conclusions*

Of primary interest in this study is the orderly arrangement of the discrimination scores and distance estimates with respect to the modifications made on the instrumental notes. This suggests that listeners responded to the magnitudes of the physical simplifications incurred by the notes in terms of being able to discriminate between conditions. It also shows that the estimates of subjective distances between conditions was highly correlated with the discriminability of the conditions, which lends credence to a multidimensional scaling of the distance estimations.

In looking at the structure of the distance estimations, the greater magnitude of difference between the *cut-attack approximation* and all other conditions highlights the extreme importance of the onset pattern of instrumental tones often referred to in the literature [Luce, 1963; Saldanha and Corso, 1964; Berger, 1964; Wedin and Goude, 1972]. The interpretation of dimensions in the stimulus space in the appropriate multidimensional scaling solution reflects the importance of the preservation of precedent low-amplitude inharmonicity during the attack.

The differences between the tonal conditions for the instrumental notes may also be interpreted with respect to discriminability. It is important to note initially that all 16 subjects, who had well-trained musical ears, reported that the differences between tones were very hard to detect, and furthermore, that the absolute subjective distances between the tones was extremely small. Tones were almost in always heard to be life-like, but having different articulations or styles of playing. It generally was *not* the case that the original was heard to be natural and the synthetic versions were not.

The discriminability of the *original* tone from the *re-synthesis* was better than expected, averaging .68. The methodological necessity of first recording the tones and then digitizing them was unfortunately a confounding variable in this case, due to its addition of varying amounts and types of noise in the process. The high degree of variance in this pair of conditions may reflect the differential success of the analysis-synthesis technique, which also had to deal with this noise, or it may be largely a function of the inability to simulate the time-variant noise of the *original* signal in the other conditions. No rigorous conclusion can be made with these data.

The difference between the *complex synthesis* and the *line segment approximation* in terms of discrimination averages .63, and indicates that this is a hard case to distinguish. This finding suggests that the highly complex microstructure in the time-variant amplitude and frequency functions given by the analysis is not absolutely essential to the timbral quality of the tone, and such a drastic data reduction as was performed in this case will do little harm. This extends the success of Risset [1966] with approximating the trumpet tone now to all families of harmonically-structured instruments. It also gives the researcher in timbre perception a very powerful tool for the production of stimuli, in that such simplified tones have more clearly defined physical properties.

The discriminability between the *constant frequencies approximation* and both the *complex synthesis* and the *line segment approximation* suggest that this degree of simplification is too highly discriminable in many cases to be of general use. Certain instrumental notes inherently had little frequency shift, such as the English horn, while others had considerably more, as with the soprano saxophone. With the latter sort of tone, timbral differences will be heard with this degree of modification.

The discriminability of the *cut-attack* condition is essentially parallel to the distance estimations, and, as stated above, was the most discriminable case between all other conditions.

## B. Perceptual Equalization of Synthesized Music Instrument Tones for Pitch, Loudness and Duration

*Introduction*

The perception of musical timbre is a little understood topic in the field of psychoacoustics. The most common definition given is essentially one of exclusion: *Timbre is that attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar.* We might add the parameters of duration and spatial location to the set of factors which negatively define timbre. Left open to the investigator of timbre perception is a whole gamut of physical factors that may be important in distinguishing two sounds which are played at the same pitch, loudness, duration and location.

In this study we are interested in the process of *controlling* the first three of the above factors which have been excluded from timbre: *pitch, loudness and duration.* This is of ultimate importance to timbre research, so as to be able to avoid confounding dimensions in stimuli used to study the already far too multidimensional attribute of timbre. Control for these factors is sought in the form of equalization.

While finding no studies in the literature of audition which have specifically investigated the influences of pitch, loudness or duration on the perception of timbre, there have been a number of studies which indirectly show that these parameters do interact with timbre. Investigations on timbre and pitch found significant perceptual changes in timbre with good-sized steps in pitch [Slawson, 1968; Plomp and Steeneken, 1971]. Experiments on the perception of loudness for tones which vary in spectral and onset characteristics showed that loudness was indeed dependent on these timbral cues [Gjaevenes and Rimstad, 1972]. Studies of speech perception have found that duration serves as a cue for the identification of synthetic vowels [Ainsworth, 1972].

In addition, informal studies at our laboratory have revealed that slight inequalities in these three physical factors will most often be heard as dissimilarities in *timbre* when the perceptual task is oriented to ratings of timbre. This is true even for experienced listeners and trained

musicians. Very slight mistunings in pitch will be heard as timbral differences in *brightness* or *dullness*. Small loudness differences are interpreted in the same manner. Duration inequalities as small as 10 milliseconds could be heard, and could definitely present a confounding variable to judgments of timbral similarity.

Unfortunately, no *a priori* techniques now exist for the equalization of the pitch, loudness or perceived duration of complex, time-variant, natural tones. We therefore have resorted to an empirical study to perform such equalizations. Listeners in a computer-controlled experiment matched these parameters for synthesized musical tones which were derived from analyzed instrumental notes. Not only has a set of normalized stimuli been obtained for future studies in timbre perception, but valuable data has been collected for modelling the influences of combined spectral and temporal factors on the perception of pitch, loudness and duration in natural tones.

### *Experiment 1:* Equalization of Duration and Intensity

*Stimuli*

The stimuli were derived from the same 16 instrumental notes used in Section A. They were generated by the analysis-based additive synthesis earlier described [see Appendix A], after considerable simplification of the results of analysis. The analyzed time-variant amplitude and frequency functions for each harmonic of the original tone were reduced to small numbers of line segments before being used for synthesis. They were, in fact, identical to the 16 *line segment approximation* tones in Section A.

In this experiment, the Eb clarinet tone was taken to be the standard tone for the equalizations. Its measured intensity level at 3 feet from the speaker, using a B&K sound-level meter, is given in Table 1b. This was found to be a moderate and comfortable level for all listeners. Its physical duration was 330 milliseconds, and its fundamental frequency averaged 313 Hz.

All 16 instrument notes were synthesized at incremental values of intensity and physical duration. The changes in duration had no effect on the frequencies of the tones. This was possible because in the analysis-based synthesis technique: 1) the duration of a tone is determined by the rate at which amplitude and frequency functions are read by the synthesis algorithm, however, since the values in the frequency functions do not change unless intentionally altered, there is an alteration of duration with no change in frequency; 2) the frequencies of a tone may be affected by setting any ratio with respect to the originally analyzed functions, hence transposition without a necessary change in duration is accomplished; and 3) the amplitude could be scaled down arbitrarily.

16 values of physical durations were synthesized, in 10 millisecond steps from 250 to 400 milliseconds; and 16 values of intensities were generated for each duration, varying in steps

increasing from .6 to .8 dB in size respectively from 0 to -10.5 dB, relative to the maximally intense signals. Each instrument note therefore existed in 256 versions, a 16 x 16 matrix of duration versus intensity. These tones were stored digitally, in the computer.

The experiment took place in a fairly dead room approximately 16 x 22 feet in size. Playback was accomplished directly from the computer through a 12-bit digital-to-analogue converter. The analogue signal was passed through a Scully 280 preamplifier and a Dynaco Stereo 70 amplifier, and was transduced by an Altec 804 loudspeaker.

*Listeners and Procedure*

Ten listeners at Stanford University were employed for this experiment, 3 of whom took the experiment a second time, making a total of 13 data sets obtained. Listeners were musically sophisticated, some actively involved in advance instrumental performance and others in conducting or musical composition. All listeners had experience with the production of computer music, and had well-trained ears with respect to synthesized timbres.

Listeners were placed 10 feet away from the speaker, directly facing it. This arrangement was made on a diagonal across the room. Listeners participated in a variable-length session, determined by their own pace, usually lasting between 2 and 3 hours. There were a total of 64 trials, hence four replications per tone. Several practice trials were given before the experiment began, enough to familiarize each listener with the procedure.

The basic component of a trial consisted of a four-note pattern, the physical onsets of successive notes being at 1 second intervals. The notes were played in an ABAB order for the first appearance of the pattern, where A = the *standard tone* and B = the *variable tone*. The physical duration and intensity of B was randomly selected from the 16 possible values for each of the two parameters. The order of trials was randomly selected.

Listeners were instructed to indicate to the computer, by means of a small set of keys on a silent teletype keyboard, any change that they desired to make on the duration or intensity of the variable tone so as to better equalize its perceived duration and loudness with the standard tone. Upon hearing a four-note pattern, the listener could do one of the following operations: 1) simply repeat the same pattern again; 2) re-order the pattern, e.g. if it was ABAB then it would become BABA; 3) increase or decrease the intensity of the B tone, and indicate the magnitude of the stepsize; 4) increase or decrease the duration of the B tones, and indicate the magnitude of the stepsize; or 5) indicate that the desired matching had been accomplished by terminating the present trial.

Listeners could spend as much time as desired in any one trial, and could respond at their own pace. The exact step-size used in changing a parameter was selected by the computer at random to be either one or two units of difference with respect to the magnitude of stepsize indicated by

the listener. The finest stepsize equalled the 16 discrete steps synthesized. Listeners could magnify the selected stepsize by a factor of 3 or 6, and they were told of the random selection process. In the event that they hit a boundary of any parameter, e.g. desired a tone longer than the longest synthesized duration, they were suitably warned (this was not often the case).

*Results*

The resulting equalizations are presented graphically in Figure 7. There were four replications for each instrument note, each judgment having both a match for loudness and for duration. The means of the loudness and duration judgments were derived and plotted for each listener within the individual duration by intensity matrices shown in Figure 7. Note that there are a total of 16 matrices, one plotted for each variable tone which was subjected to equalization.

Listeners' average scores are labeled with respect to the key which appears in the figure, while the overall average is plotted as the center of the grid-like cross patterns in the matrices. The standard deviation of the duration and intensity dimensions is given by the extensions of the crosses along the respective axes. In a search of the literature for loudness matching experiments, we found that the variance obtained in these matches was low.

Note that the equalization of the standard tone, the Eb clarinet (CLAR 1), with itself as the variable tone was the only case in which there was an *a priori* match based on physical identity. The accuracy of this match was quite good, and the standard deviations for both dimensions was the smallest of those obtained for the 16 instrument notes.

The most informative way of viewing these results is shown in Figure 8, where the 16 tones are plotted in terms of their equalized durations and amplitudes in the three-dimensional Amplitude x Frequency x Time graphs [see Appendix B].

*Experiment 2:* Equalization of Frequency

*Stimuli*

The 16 tones equalized in Experiment 1 for loudness and perceived duration were used as the basic instrument tones for this experiment. In this experiment, the Eb clarinet tone was again taken to be the standard tone for the equalizations.

All 16 instrument notes were synthesized at incremental values of frequencies. 24 values of frequencies were synthesized for each note, in .5 Hz steps of the fundamentals from the average fundamental frequency of 306.5 Hz to 318 Hz. These tones were stored digitally, in the computer.

All other features of the experimental stimuli were as described above in Experiment 1.

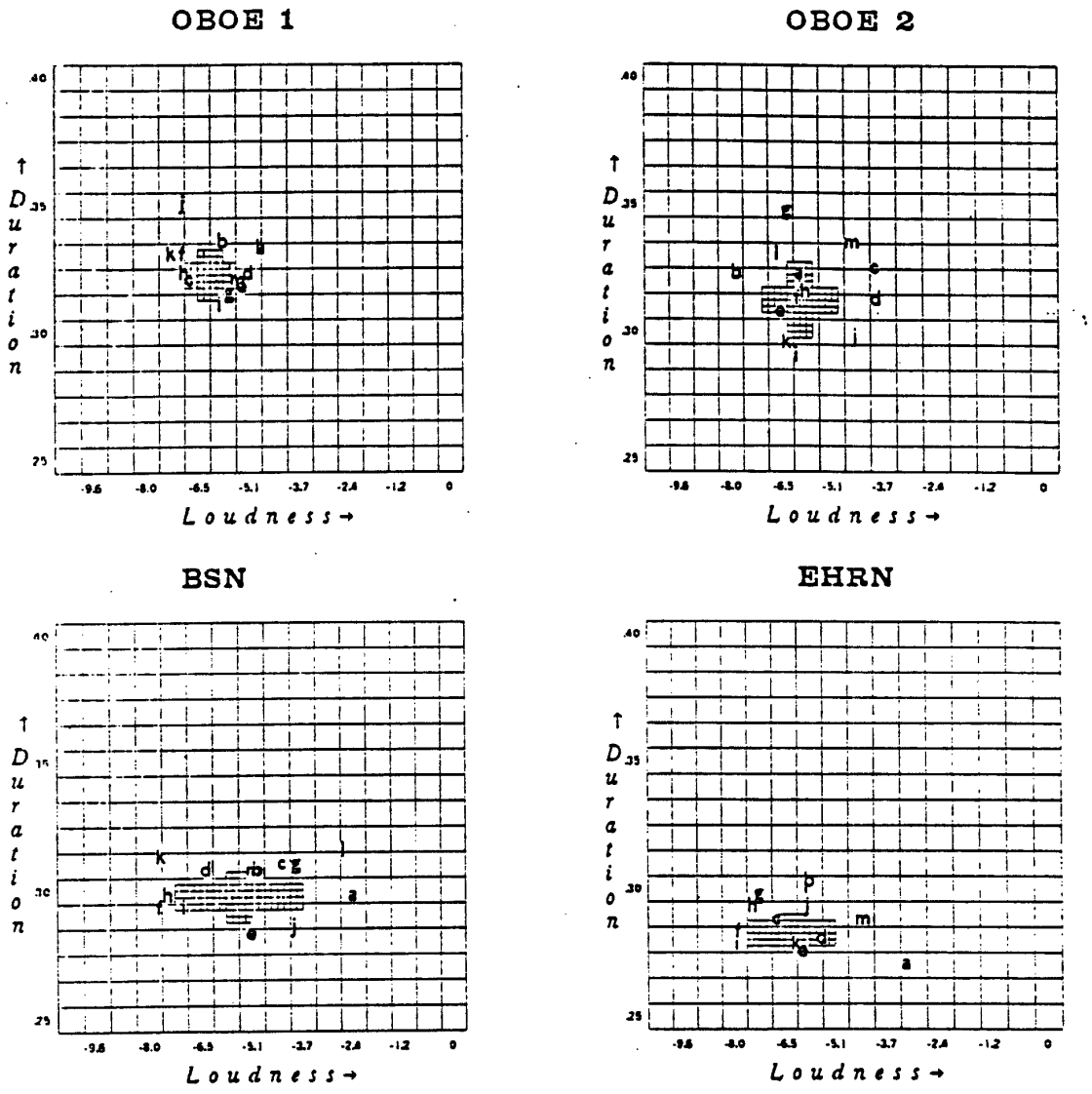



FIGURE 7 (part 1): Loudness - Duration equalizations [Exp 1, sec B]. Results for four tones plotted above. Loudness matches (X-axis) are plotted with resect to physical dB relative to the most intense version of the tone. Duration (Y-axis) is plotted in seconds. The center of the cross-grid represents the overall average and the length of extension gives the standard deviation along the two axes. Average matches for each of the listeners are plotted in lower case letters (13 data sets).

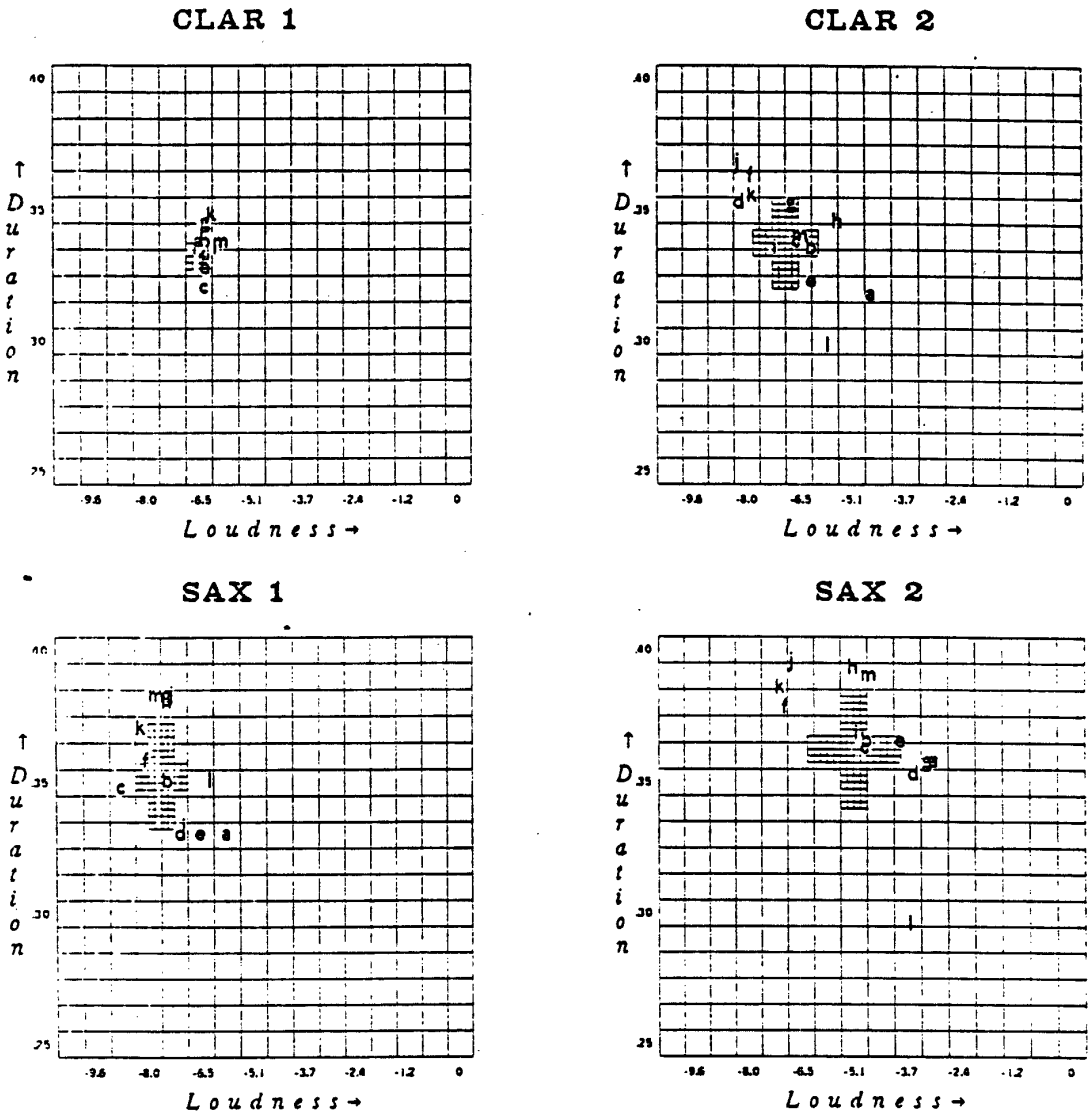



FIGURE 7 (part 2): Loudness - Duration equalizations [Exp I, sec B]. Results for four tones plotted above. Loudness matches (X-axis) are plotted with resect to physical dB relative to the most intense version of the tone. Duration (Y-axis) is plotted in seconds. The center of the cross-grid represents the overall average and the length of extension gives the standard deviation along the two axes. Average matches for each of the listeners are plotted in lower case letters (13 data sets).

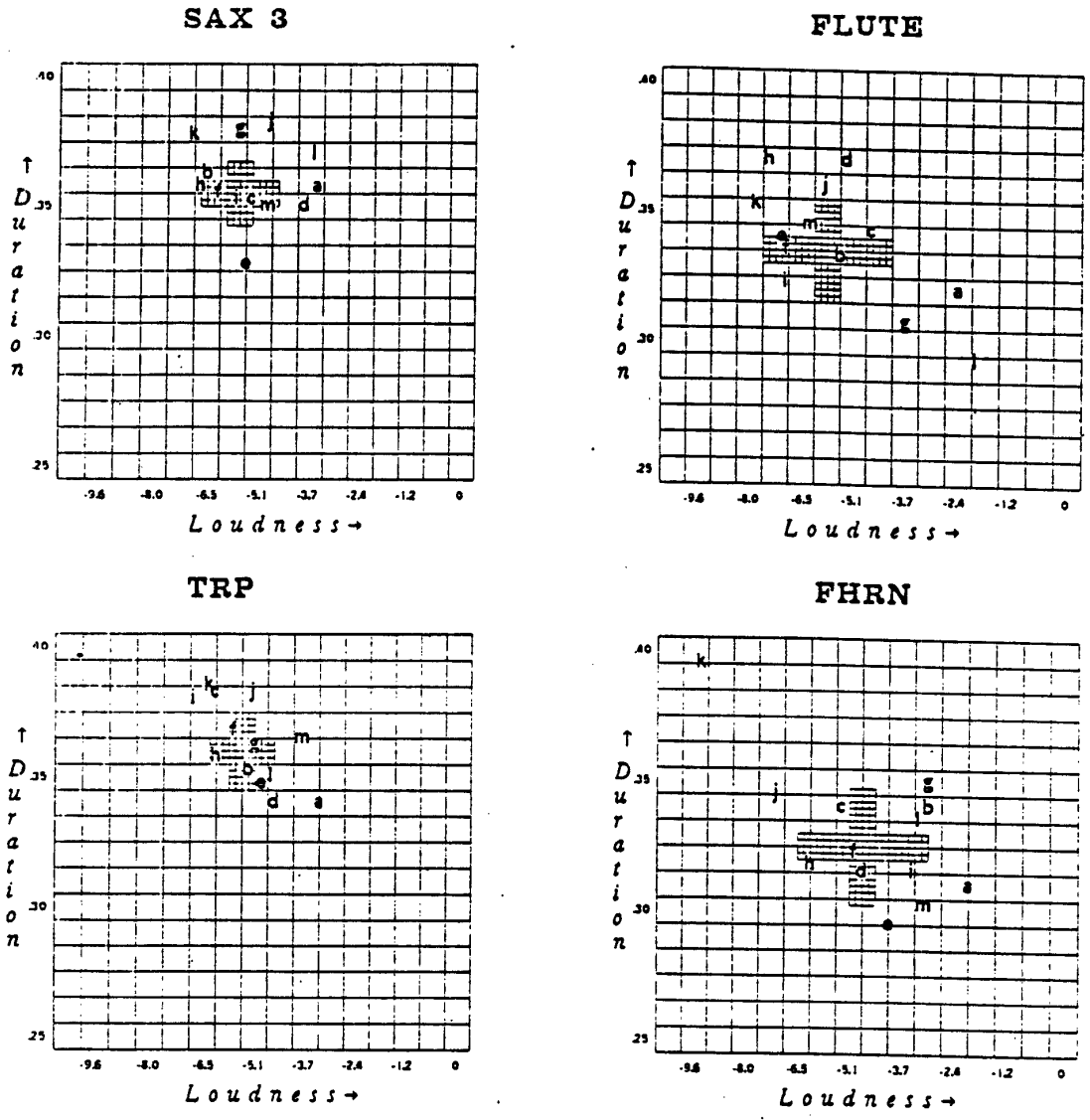



FIGURE 7 (part 3): Loudness - Duration equalizations [Exp 1, sec B]. Results for four tones plotted above. Loudness matches (X-axis) are plotted with resect to physical dB relative to the most intense version of the tone. Duration (Y-axis) is plotted in seconds. The center of the cross-grid represents the overall average and the length of extension gives the standard deviation along the two axes. Average matches for each of the listeners are plotted in lower case letters (13 data sets).

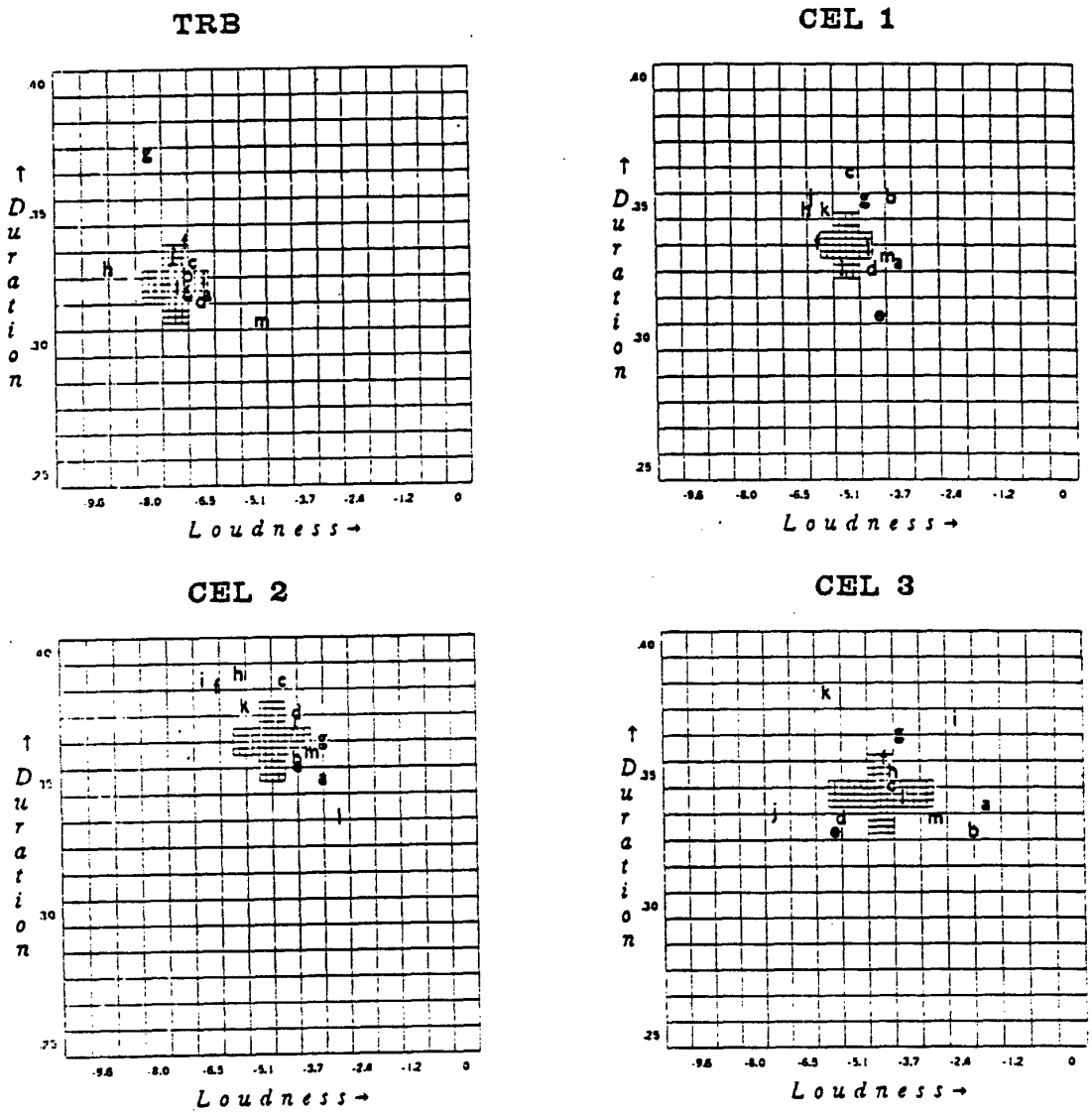



FIGURE 7 (part 4): Loudness - Duration equalizations [Exp 1, sec B]. Results for four tones plotted above. Loudness matches (X-axis) are plotted with resect to physical dB relative to the most intense version of the tone. Duration (Y-axis) is plotted in seconds. The center of the cross-grid represents the overall average and the length of extension gives the standard deviation along the two axes. Average matches for each of the listeners are plotted in lower case letters (13 data sets).

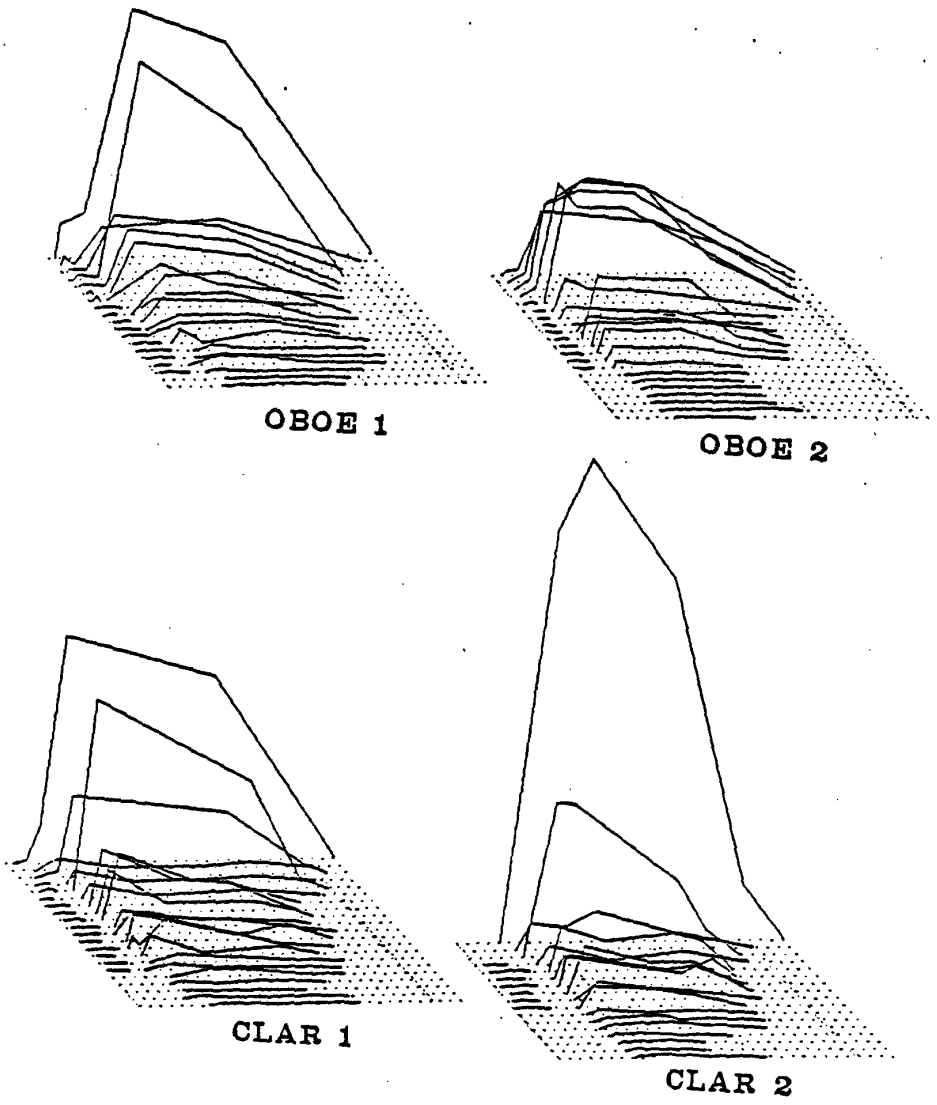



FIGURE 8 (part 1): Equalized tones for Duration and Amplitude [Exp 1, sec B]. The equalized physical properties of the tones are represented by Amplitude x Frequency x Time plots (X-axis = time with 1/10 second divisions demarked along floor grid; Y-axis = amplitude; Z-axis = frequency with the fundamental harmonic in the background).

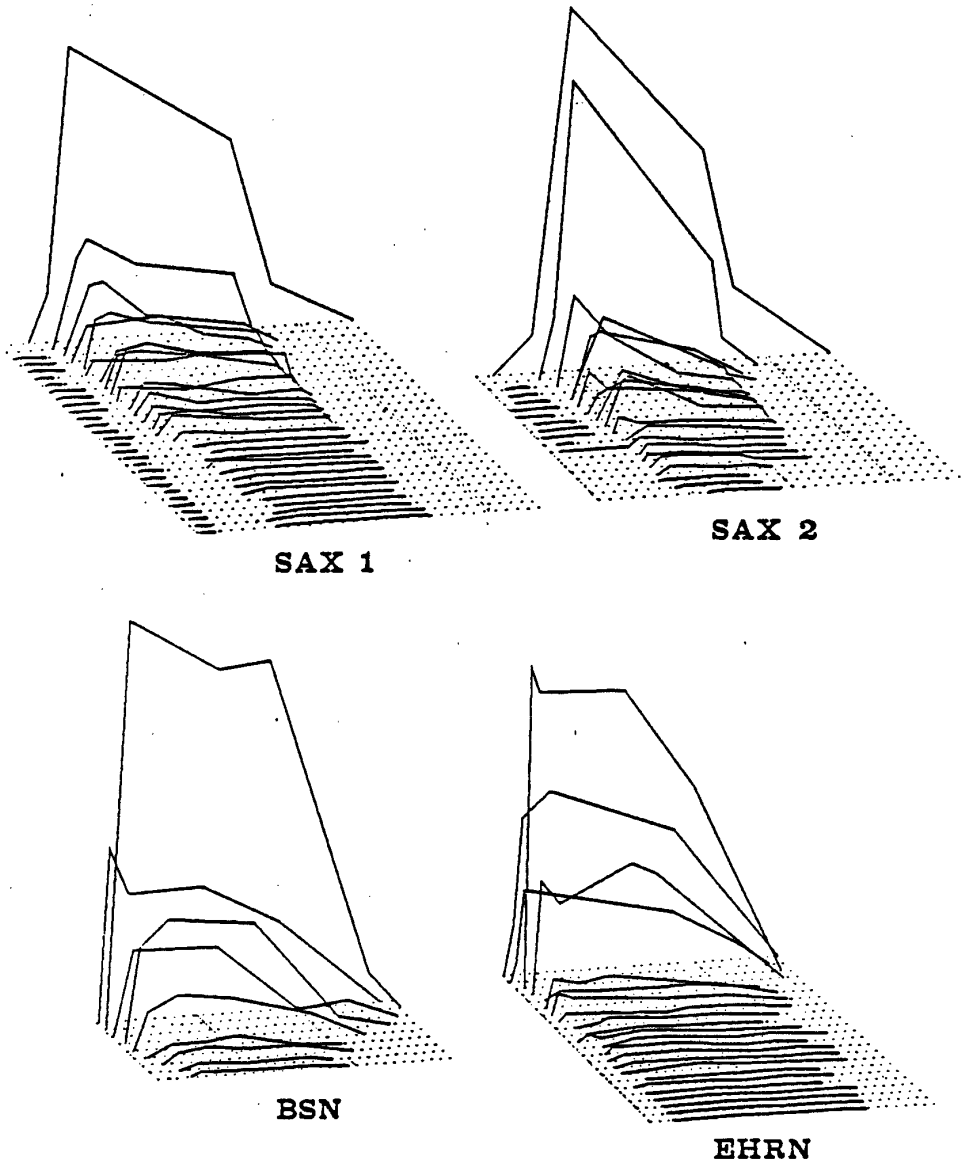



FIGURE 8 (part 2): Equalized tones for Duration and Amplitude [Exp 1, sec 8]. The equalized physical properties of the tones are represented by Amplitude x Frequency x Time plots (X-axis = time with 1/10 second divisions demarked along floor grid; Y-axis = amplitude; Z-axis = frequency with the fundamental harmonic in the background).

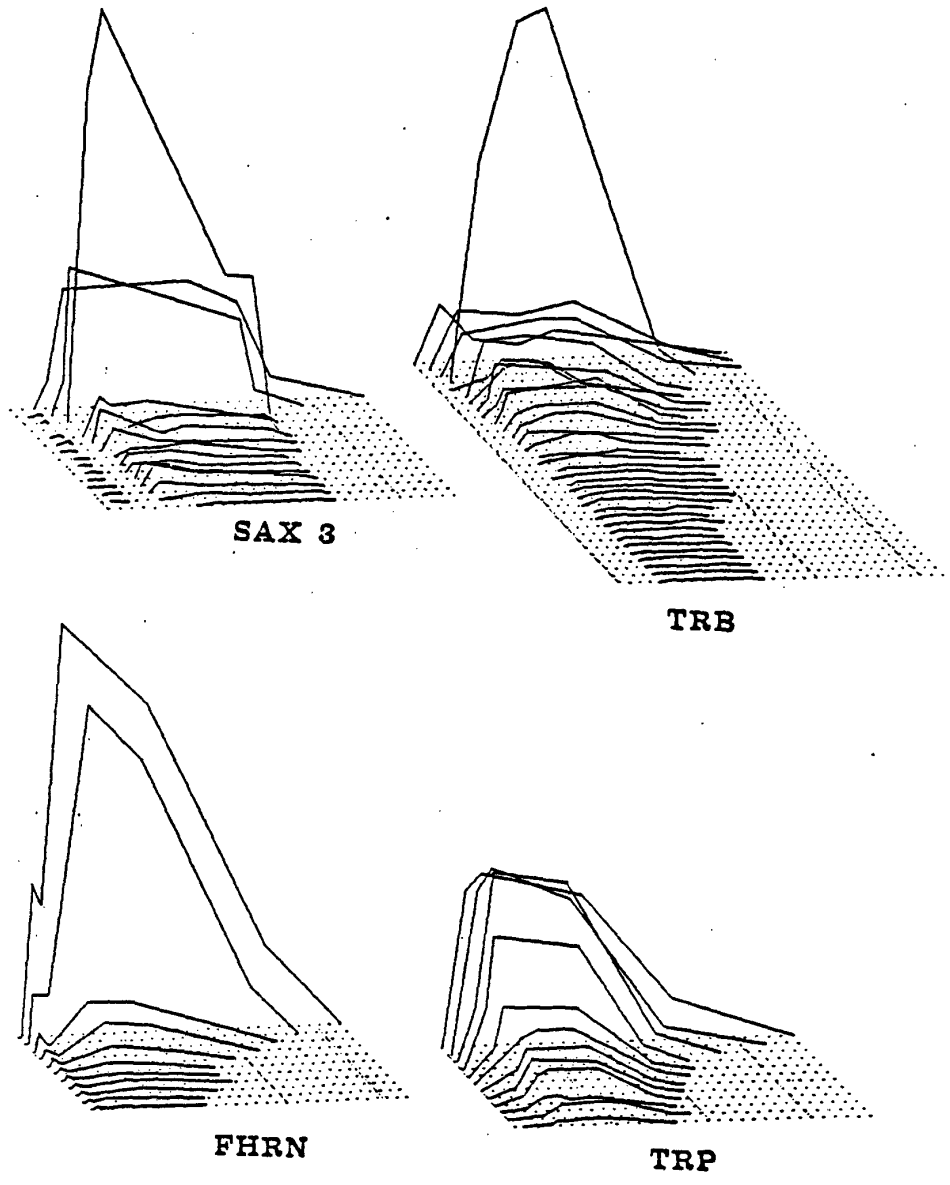



FIGURE 8 (part 3): Equalized tones for Duration and Amplitude [Exp I, sec B]. The equalized physical properties of the tones are represented by Amplitude x Frequency x Time plots (X-axis = time with 1/10 second divisions demarked along floor grid; Y-axis = amplitude; Z-axis = frequency with the fundamental harmonic in the background).

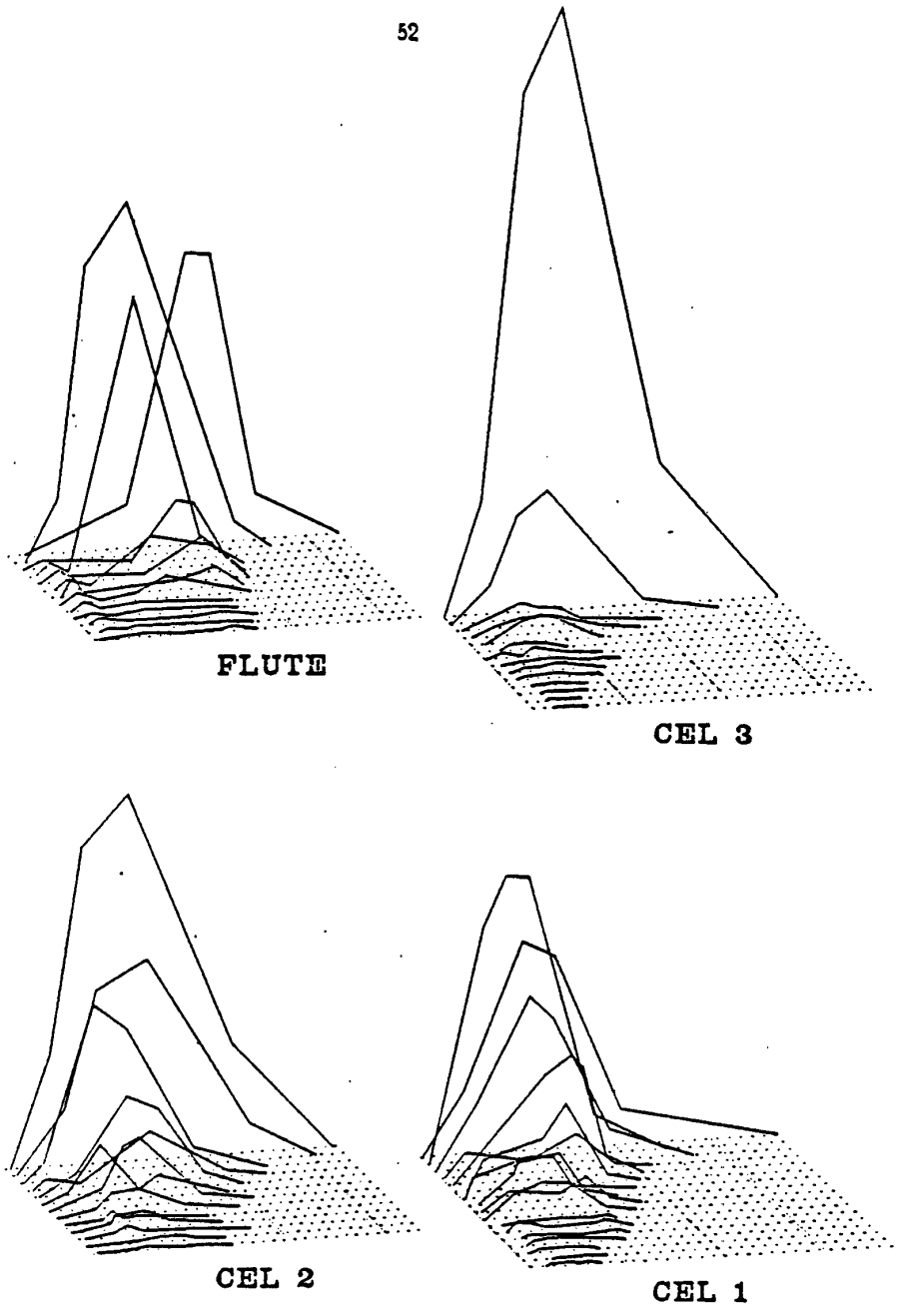



FIGURE 8 (part 4): Equalized tones for Duration and Amplitude [Exp 1, sec B]. The equalized physical properties of the tones are represented by Amplitude x Frequency x Time plots (X-axis = time with 1/10 second divisions demarked along floor grid: Y-axis = amplitude: Z-axis = frequency with the fundamental harmonic in the background).

*Listeners and Procedure*

Ten listeners at Stanford University were employed for this experiment, one of whom took the experiment a second time, making a total of eleven data sets. Listeners were musically sophisticated, some actively involved in advance instrumental performance and others in conducting or musical composition. All listeners had experience with the production of computer music, and had well-trained ears with respect to synthesized timbres. Most features of the study were identical to those described above for Experiment 1, excepting the following changes. This experiment was a one-dimensional equalization of frequencies, in terms of perceived pitch, so listeners manipulated only one parameter of sound, rather than two. The experiment therefore tended to take much less time than the first one, lasting on the average only one hour.

*Results*

The resulting equalizations are presented graphically in Figure 9. There were four replications for each instrument note, each judgment being a match in perceived pitch. The means of these judgments were derived and plotted for each listener within the individual frequency charts shown in Figure 9. Note that there are a total of 16 charts, one plotted for each variable tone which was subjected to equalization.

Listeners' are labeled with respect to the key which appears in the figure, while the overall average is plotted as the center of the grid-like rectangle in the charts. The standard deviation of the frequency judgments is given by the extension of the rectangle along the abscissa.

Note that the equalization of the standard tone, the Eb clarinet (CLAR 1), with itself as the variable tone was the only case in which there was an *a priori* match based on physical identity. The accuracy of this match was quite good, and its standard deviation was the smallest of those obtained for the 16 instrument notes.

*Conclusions*

Equalizations were performed for 16 tones with a standard tone in the perceptual dimensions of pitch, loudness and duration. The stimuli were complex, time-variant, computer-generated tones synthesized on the basis of analyzed natural music instrument notes.

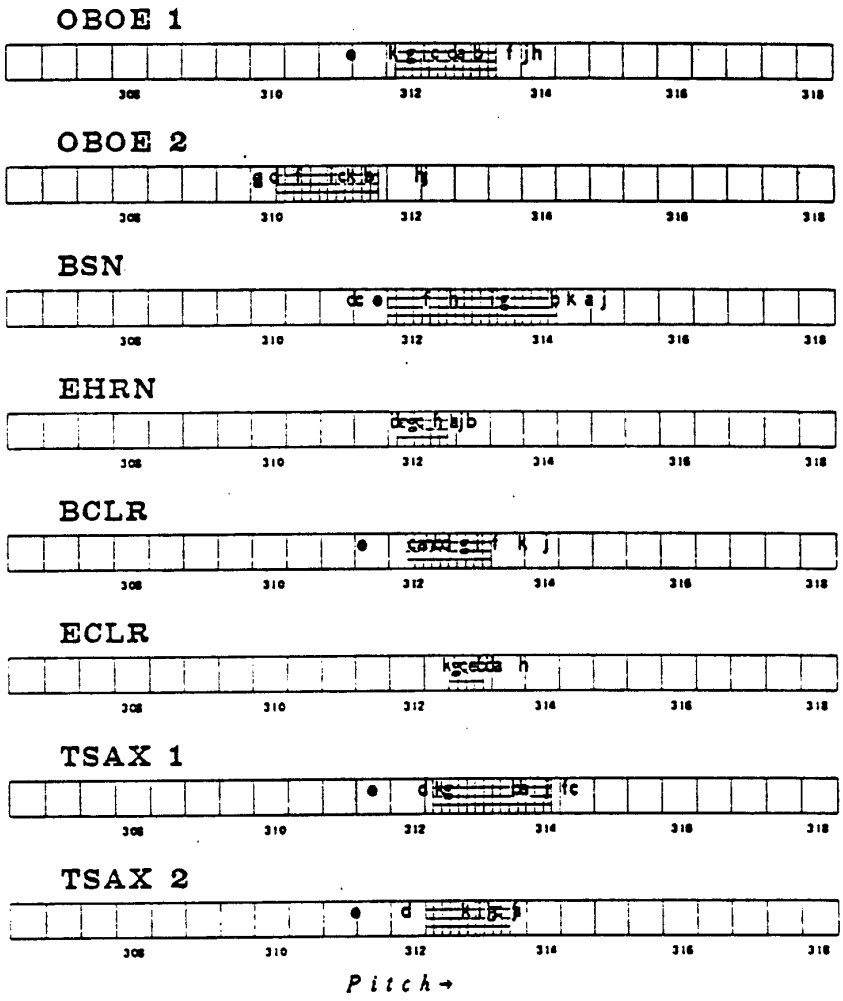



FIGURE 9 (part 1): Pitch matches for 16 tones [Exp 2, sec B]. Results for eight tones are plotted above. Pitch matches (X-axis) are given with respect to the average frequency of the fundamental in Hz. The center of the grid represents the overall average and the length of extension gives the standard deviation. Average matches for each individual are plotted in lower-case letters (11 data sets).

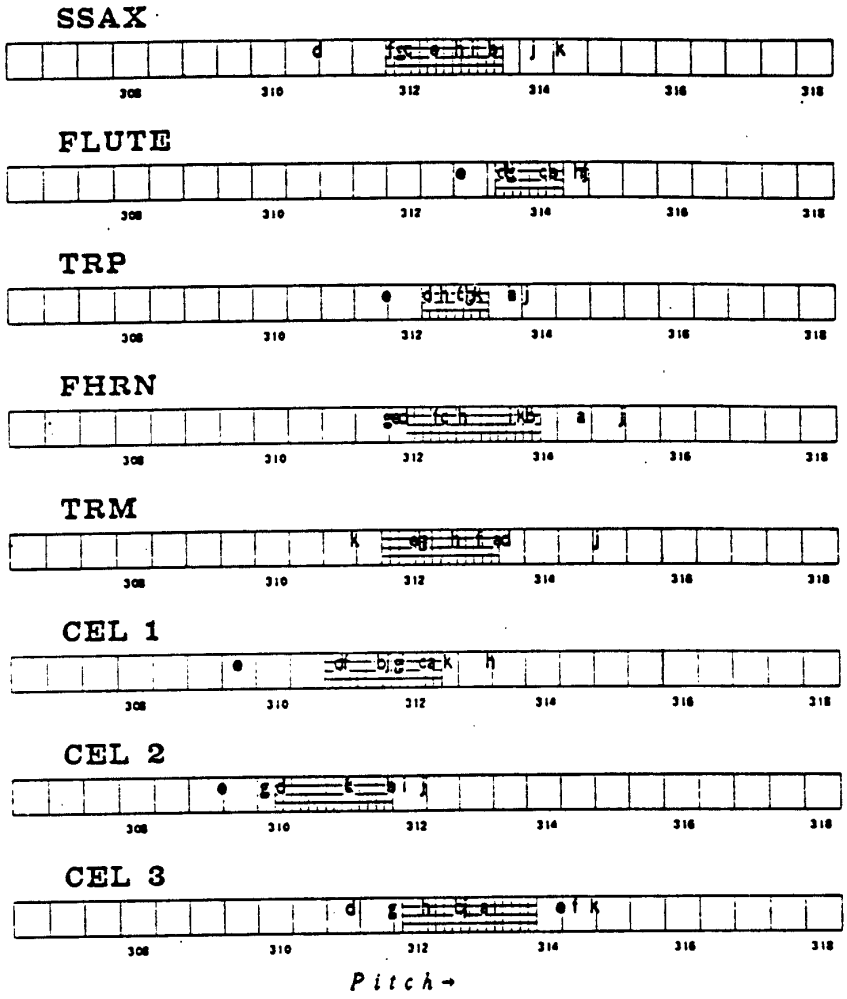



FIGURE 9 (part 2): Pitch matches for 16 tones [Exp 2, sec B]. Results for eight tones are plotted above. Pitch matches (X-axis) are given with respect to the average frequency of the fundamental in Hz. The center of the grid represents the overall average and the length of extension gives the standard deviation. Average matches for each individual are plotted in lower-case letters (11 data sets).

In that there exists no technique capable of equalizing such complex stimuli on an *a priori* basis, the present empirical approach was pursued for two reasons: 1) to obtain a set of matched stimuli for studies on the perception of musical timbre, and 2) to obtain data which could be useful in attempts to model the perception of these matched attributes of tone for such complex, naturalistic stimuli. In that we are presently pursuing a model for loudness perception in the case of time-variant tones, we will briefly examine this data in terms of known loudness theories. The most comprehensive theory of loudness has been developed by Zwicker and Scharf [1965] for an evaluation of the loudness of complex steady-state signals. This model attempts to deal with these various factors which influence loudness: 1) the sensitivity of the ear at different frequencies, 2) the locations of stimulation at different frequencies along the cochlea, 3) the effects of critical bandwidths and 4) the effects of masking patterns with respect to the above. We computed loudness scores on the basis of this model and indeed found that the spread of these scores around that of the *standard tone* used in the matches was approximately the same as the overall spread of individual listeners' matches around the average matches for each of the tones. This would indicate that the Zwicker model may be of use in automatic equalizations, and we plan to pursue it further. Of particular interest is the significance of the exact spread of matchings found in this experiment, and whether there are interactions to be found with the time-variant properties of the tones (say, speed of attack).

We also plan to look at the cues which may be responsible for the discrepancies in pitch matches. Previous investigators have found some small influences of spectral shape on the perceived pitch of steady-state tones [Fletcher, 1934; Lichte, 1955], where tones having higher-frequency energy would be heard as being higher in pitch. The data which we have collected suggests that matters are considerably more complex with natural tones. For example, the two oboes have similar distributions of energy, but one was consistently judged to be lower in pitch by almost 2 Hz by all listeners. We plan to look at the microstructure of the frequency information for the 16 tones to evaluate the bases of these matches, and to try to apply recent notions of spectral dominance in pitch perception [Ritsma, 1967] to narrow in on particular cues for the musical pitch of these examples.

Similar plans will be followed to evaluate the relationship of perceived duration to the time-variant timbre. As a result of this experiment, it seems clear that there is much work yet to be done before achieving general models for the pitch, loudness and durations of complex, time-variant tones. This paradigm for the collection of data in the case of time-variant stimuli marks a first step in that direction. It is clear that this can only be done with a computer, and that, ideally, a real-time synthesis process should be utilized in which the listener, using knobs that are continuously monitored by the computer, could change parameters within a continuum and thus avoid arbitrary boundaries and discretely pre-computed stepsizes and values.

Research should investigate not only the cues which exist for such matches, but it should also look at the relative difficulties which confront the listener in matching different pairs of tones. The non-uniformity of difficulties in equalization across different stimulus pairs can be seen in

the range of individual and group variance in this experiment. Some pairs of tones were much harder to equalize than others, and this must have a direct relationship to the nature of the cues for such matches.

## C. Multidimensional Perception of Synthesized Musical Timbres

*Introduction*

In discussing the attributes of complex tones, Licklider [1951] concludes that "until careful scientific work has been done on the subject, it can hardly be possible to say more about timbre than that it is a 'multidimensional' dimension." In the last decade, investigators of timbre have taken marked steps in the direction of finally being able to deal with the multidimensionality of their subject. These recent improvements in timbre research are largely the result of technological advances in the use of digital computers which have given the investigator powerful new means for the analysis and synthesis of complex, time-variant music instrument tones [Luce, 1963; Risset, 1966; Freedman, 1967, 1968; Beauchamp, 1969; Mathews, 1969; Chowning, 1973; Moorer, 1973], and for the analysis and presentation of complex, multidimensional data structures of the type that may be collected from studying the perception of timbre [Plomp, 1970; Wessel 1973, 1974].

A major aim of research in timbre perception is the development of a theory for the salient dimensions or features of classes of sounds. Given that timbre is clearly a *multidimensional* attribute of sound, the computer-based techniques of perceptual data analysis which fall under the category of multidimensional scaling, hereafter abbreviated MDS, seem especially well-suited for examining the complex aspects of timbre perception [Shepard, 1962a, 1962b; Kruskal, 1964a, 1964b; Carroll and Chang, 1970; see Appendix C for a brief description].

In this investigation we have collected two sorts of psychological distance measurements between 16 synthesized musical instrument notes. One measurement was a direct evaluation of the timbral similarities for all pairs of the 16 instrument notes. Here, the psychological distance of two tones is the inverse of their similarity. This data was treated with MDS algorithms to obtain an easily interpretable picture of the major dimensions in the psychological distance structure between the tones.

The other measurement taken was of the accuracy of listeners in associating specific names with the notes, in a learning task with feedback. In this case the psychological distance of two tones is inversely related to the number of confusions that occurred between them in learning their respective names. This data was compared to the MDS of the similarity structure of the tones.

The instrument notes were previously equalized for perceived pitch, loudness and duration, in order to eliminate confounding dimensions from the judgments on timbre. We feel that this step, which has generally been missing in previous studies of timbre perception, was necessary

in that timbre is classically defined to exclude the dimensions of pitch, loudness and duration, and, furthermore, that these dimensions could interact with the derived dimensions sought for timbre perception.

A second important feature of the stimuli which made them novel in timbre perception research of this kind is that they are computer-generated by an analysis-based synthesis algorithm. This allowed for their equalization as stated above, but also means that the investigator exactly determined the physical properties of the tones. This may be the most reliable manner of specifying the physical content of complex timbres, rather than a *post facto* analysis of stimuli. Of even greater significance to the cause of specifying the physical properties of the tones was the simplification done to those physical properties for the synthesis of these tones.

### *Experiment 1:* Multidimensional Scaling of Timbral Similarities

#### *Stimuli*

The stimuli in this study were derived from the equalization experiments in Section B. They were generated by the analysis-based additive synthesis detailed earlier [see Appendix A], after considerable simplification of the results of analysis. The analyzed time-variant amplitude and frequency functions for each harmonic of the original tone were reduced to small numbers of line segments before being used for synthesis. They were, in fact, identical to the 16 *line segment approximation* tones in Section A.

The stimuli were recorded and played to the listeners as previously described in Section A.

#### *Listeners and Procedure*

Twenty listeners at Stanford University were employed for this experiment, 15 of whom took the experiment a second time, making a total of 35 data sets collected. Listeners were musically sophisticated, some actively involved in advanced instrumental performance and others in conducting or musical composition.

Listeners were placed as indicated in Section A. Data sets were collected in an hour session. Each trial consisted of a warning knock, the two tones for comparison, and a decision interval. The warning knock preceded the first tone by 2.5 seconds, 1.5 seconds separated the two tones, and 6 seconds were given for the listener to make a judgment. There were a total of 270 trials, 30 of which were practice trials. The remaining 240 trials consisted of the n*(n-1) possible pairs of 16 tones, given in both directions. Trials were presented in a random order.

Listeners were told to rate the *similarity* of the two tones relative to that of all other pairs of

tones heard. They were instructed that the first 30 pairs were practice, and that they could change their rating strategies during that time. The similarity rating was made on a scale of 1 to 30, and this scale was presented to listeners as having three general ranges: 1) 1-10 = *very dissimilar*, 2) 11-20 = *average level of similarity*, and 3) 21-30 = *very similar*, relative to all pairs.

*Results*

The similarity judgments for each listener were put in a 16 by 16 matrix. In that no strong asymmetries were found in ratings with respect to the direction of presentation of pairs of tones, each set of two trials which contained the same pair of notes was averaged to form a new half matrix of similarities. The matrices were treated individually with a nonmetric MDS algorithm [Kruskal, 1964a, 1964b] and treated as a group with an individual differences MDS technique, hereafter termed INDSCAL [Carroll and Chang, 1970]. Two and three dimensional solutions were specified in each case. Individual solutions were compared with group solutions. Using nonmetric MDS techniques, between 20 and 30 runs were made per individual, in order to avoid local minima. The best solutions were tested for congruency, and accepted if they were well-matched.

In addition to subjecting the similarity matrices to group and individual MDS, they were also treated with a hierarchical clustering algorithm [based on Johnson, 1967; see also Appendix C]. The principle of maximum distances, or the so-called *diameter method* was found to give the most interpretable results. The clustering scheme produced an analysis of the similarity data which was independent of the dimensional reduction generated by the MDS algorithms. Matrices were treated individually with the hierarchic clustering scheme, and a group matrix was also analyzed. The group matrix was derived by averaging the *rank orders* of the ratings in each individual matrix. A comparison of individual solutions with the group clustering showed that this clustering was highly representative of the common trends in the individual data matrices. The two analyses were used in conjunction with one another to interpret the data [see Shepard, 1972].

The group two dimensional INDSCAL solution is shown in Figure 10. The abbreviations adopted for the 16 tones are: O1 and O2 = the two oboes; EH = the English horn; BN = the bassoon; C1 = the Eb clarinet and C2 = the bass clarinet; X1 and X2 = the two saxophone tones (*mf* and *p* respectively); X3 = the soprano sax; FL = the flute; TP = the trumpet; FH = the French horn; TM = the muted trombone; S1, S2 and S3 = the cello tones (labelled strings: *sul ponticello*, normal bowing, and muted *sul tasto*, respectively).

Embedded in this solution is the clustering analysis for the group matrix. Strengths of clustering appear within the curved forms outlining each respective level of clustering. An examination of the relationships between the clustering and the spatial solution for two dimensions reveals conflicting findings. FH appears in the wrong spatial location with respect to both clustering analysis and interpretability.

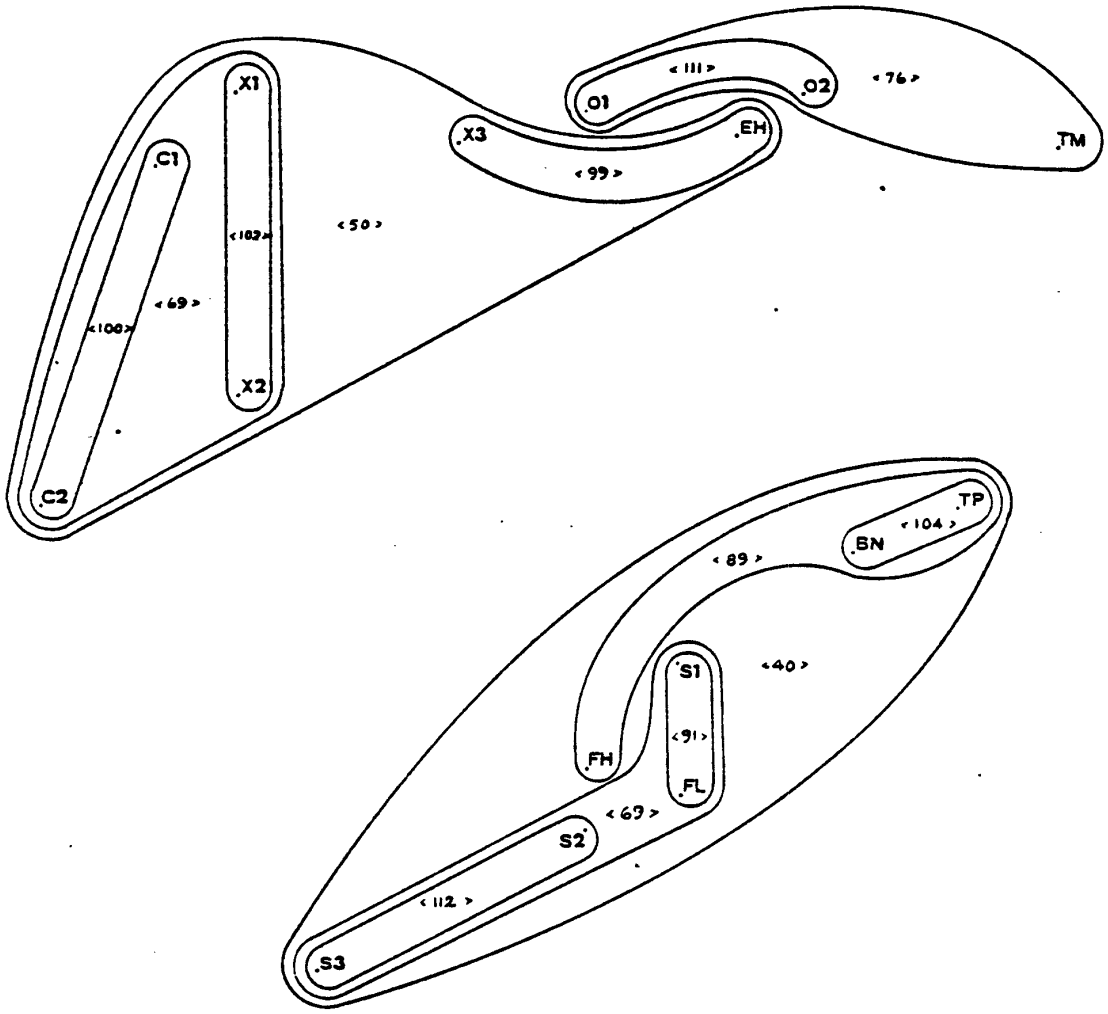



FIGURE 10: Two-dimensional group configuration for 35 similarity matrices [Exp 1, sec C] generated by individual differences multidimensional scaling program (INDSCAL [Carroll and Chang, 1970]). The results of a Hierarchical Clustering analysis [Johnson, 1967] are represented in the configuration space, as curved forms around different clusters. The levels of clustering are given inside the respective curved forms that surround each cluster. (The levels are quantitatively related to a rank-ordering of 120 items = N•(N-1)/2 pairs.) Abbreviations for stimulus points: O1,O2=oboes; C1,C2=clarinets; X1,X2,X3=saxophones; EH=English horn; FH=French horn; S1,S2,S3=strings; TP=trumpet; TM=trombone; FL=flute; BN=bassoon.

Another mismatch occurs as the justified and interpretable clusterings of O1-O2 and X3-EH are not well represented in the spatial solution, nor are the respective clusterings of C1-C2 and X1-X2.

The lack of orderly clustering in the space, the wide discrepancies between the individual two dimensional solutions, and the general lack of interpretability of the two dimensional spatial solution, suggest that the data cannot be represented in the two dimensions generated by the scaling algorithms employed. A somewhat bolder statement would be that the similarity data structure must be represented in more than two dimensions. The group three dimensional solution is displayed in the perspective plot of **Figure 11**. We will refer to the dimensions as with the following labels, whereby X = the horizontal axis spanning the width of the space, Y = the vertical axis, and Z = the horizontal axis spanning the depth of the space. The distances of the stimuli are given by their relative sizes, and the two dimensional projections of the configuration on the wall and the floor. The hierarchical clustering analysis is also represented in this spatial solution, this time by lines connecting the components of the clusters. The primary level clustering, having the greatest clustering strengths, are represented by unbroken straight lines connecting the elements. The next level, weaker clustering is represented by dashed lines, and the highest level, weakest clustering is shown by dotted lines in the space.

Note that in this spatial solution the discrepancies between the clustering analysis and the spatial locations have been resolved, and that there is a much better agreement between the two analyses. In addition, the congruency comparison of the individual non-metric three dimensional solutions with this group solution showed that it was highly representative of most individual configurations in all dimensions, and in at least two dimensions for the others.

More importantly, the three dimensional solution is interpretable in terms of the physical properties of the stimuli. To aid in making the psychophysical evaluation of the space, two dimensional projections have been prepared by locating the physical analyses of the stimulus tones at their positions in the space. The Time x Frequency x Amplitude plots of the tones are given in this manner in **Figures 12a and 12b**, and the spectrographic displays of the tones are placed in corresponding positions in **Figures 13a and 13b**.

A physical interpretation can be attempted of the three dimensional space in terms of the spectral energy distribution and the temporal features of the tones. The Y axis can be interpreted with respect to the spectral energy distribution. On the one extreme, FH and S3 have the properties of narrow spectral bandwidth and a concentration of low-frequency energy. On the other extreme, TM has a very wide spectral bandwidth and less of a concentration of energy in the lowest harmonics; O1 and O2 have significant upper formant regions about the 10th harmonic.

The X dimension would appear to relate to the form of the onset-offset patterns of tones, especially with respect to the presence of synchronicity in the collective attacks and decays of upper harmonics.

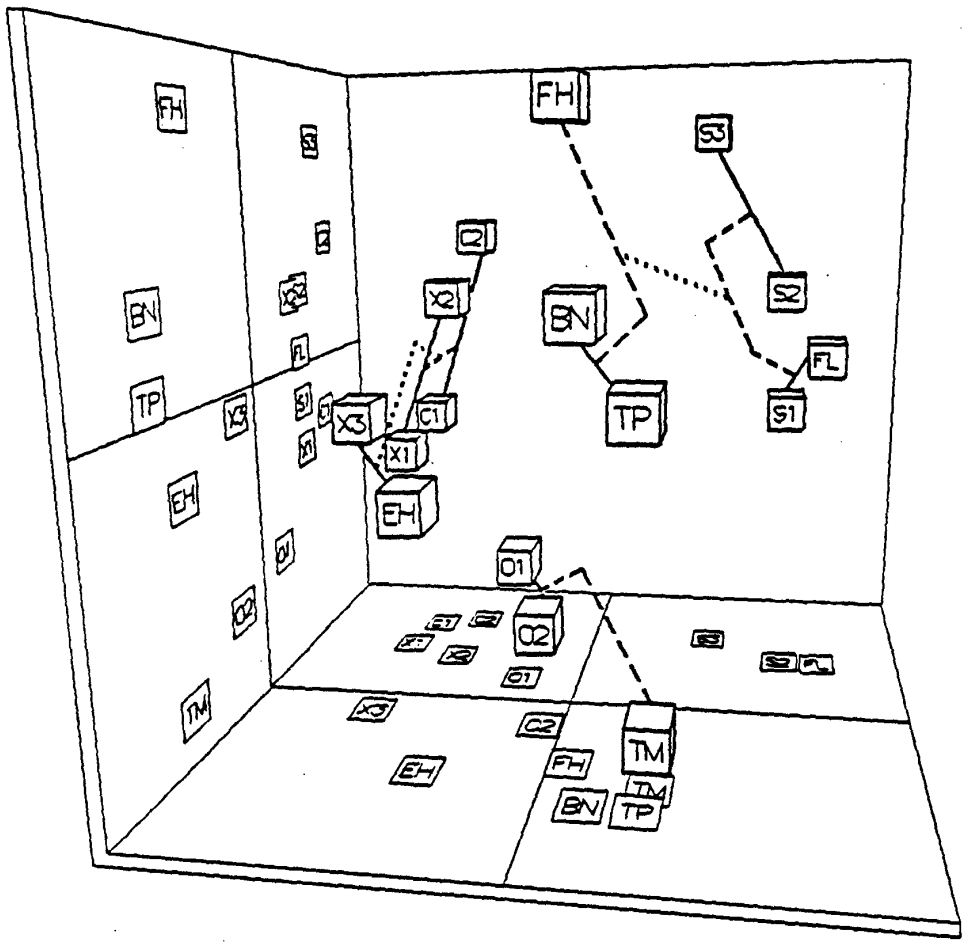



FIGURE 11: Three-dimensional group configuration for 35 similarity matrices [Exp 1, sec C] generated by individual differences multidimensional scaling program (INDSCAL [Carroll and Chang, 1970]). Hierarchical Clustering analysis [Johnson, 1967] is represented by connecting lines, in clustering order solid - dashed - dotted. Notice the two-dimensional projections of the configuration on the wall and floor. Abbreviations for stimulus points: O1,O2=oboes; C1,C2=clarinets; X1,X2,X3=saxophones; EH=English horn; FH=French horn; S1,S2,S3=strings; TP=trumpet; TM=trombone; FL=flute; BN=bassoon.

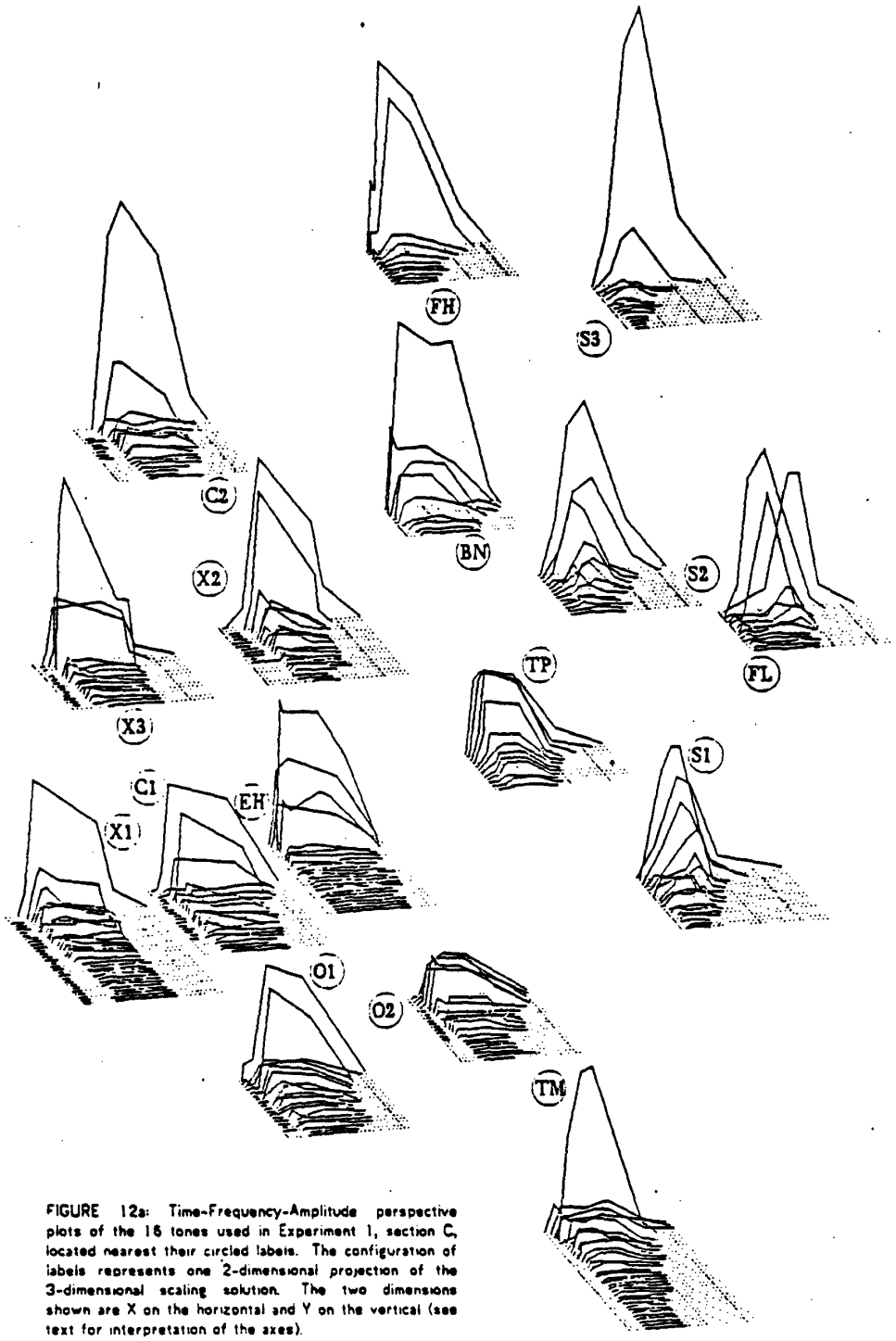



FIGURE 12a: Time-Frequency-Amplitude perspective plots of the 16 tones used in Experiment 1, section C, located nearest their circled labels. The configuration of labels represents one 2-dimensional projection of the 3-dimensional scaling solution. The two dimensions shown are X on the horizontal and Y on the vertical (see text for interpretation of the axes).

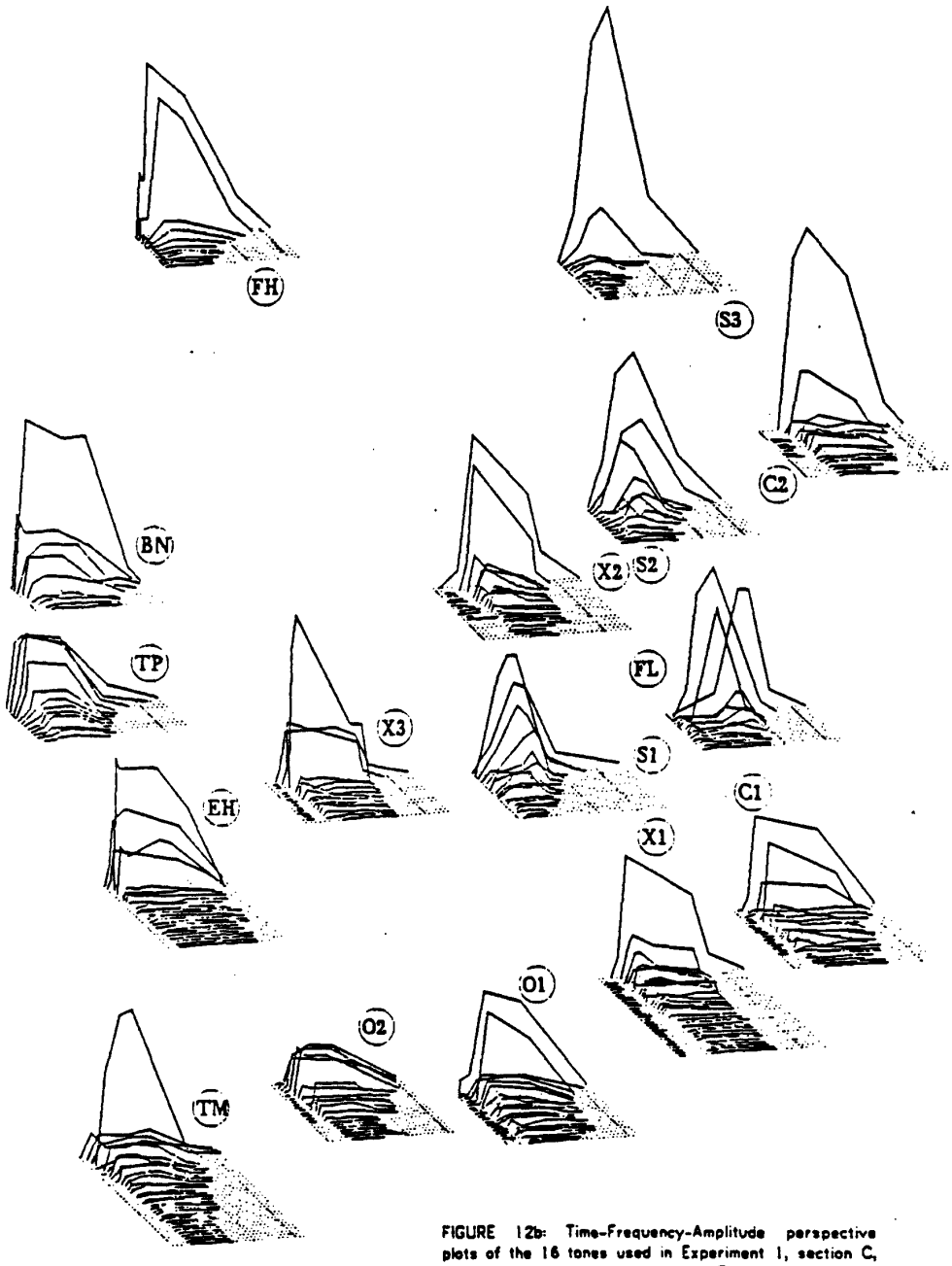



FIGURE 12b: Time-Frequency-Amplitude perspective plots of the 16 tones used in Experiment 1, section C, located nearest their circled labels. The configuration of labels represents one 2-dimensional projection of the 3-dimensional scaling solution. The two dimensions shown are Z on the horizontal and Y on the vertical (see text for interpretation of the axes).

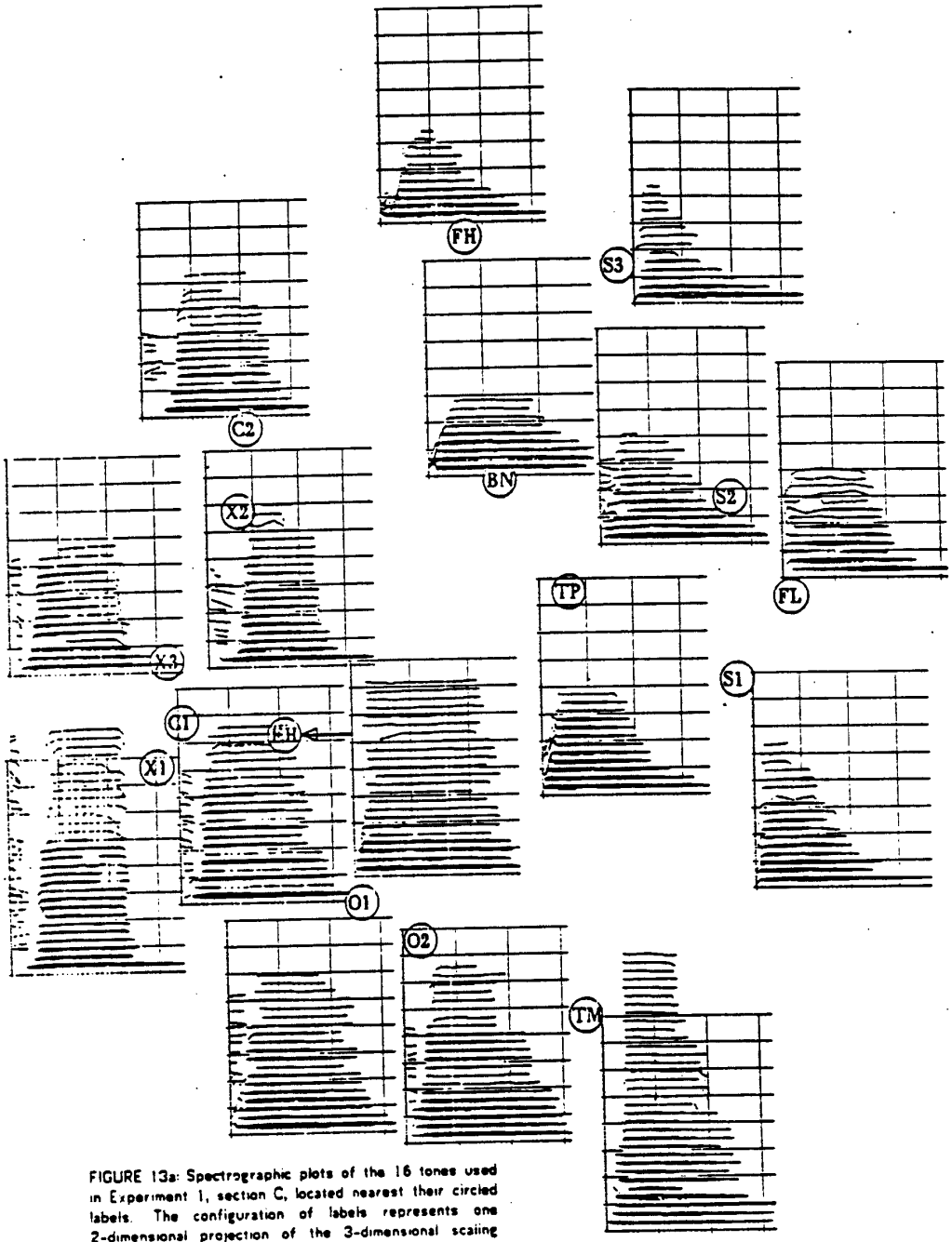



FIGURE 13a: Spectrographic plots of the 16 tones used in Experiment 1, section C, located nearest their circled labels. The configuration of labels represents one 2-dimensional projection of the 3-dimensional scaling solution. The two dimensions shown are X on the horizontal and Y on the vertical (see text for interpretation of the axes).

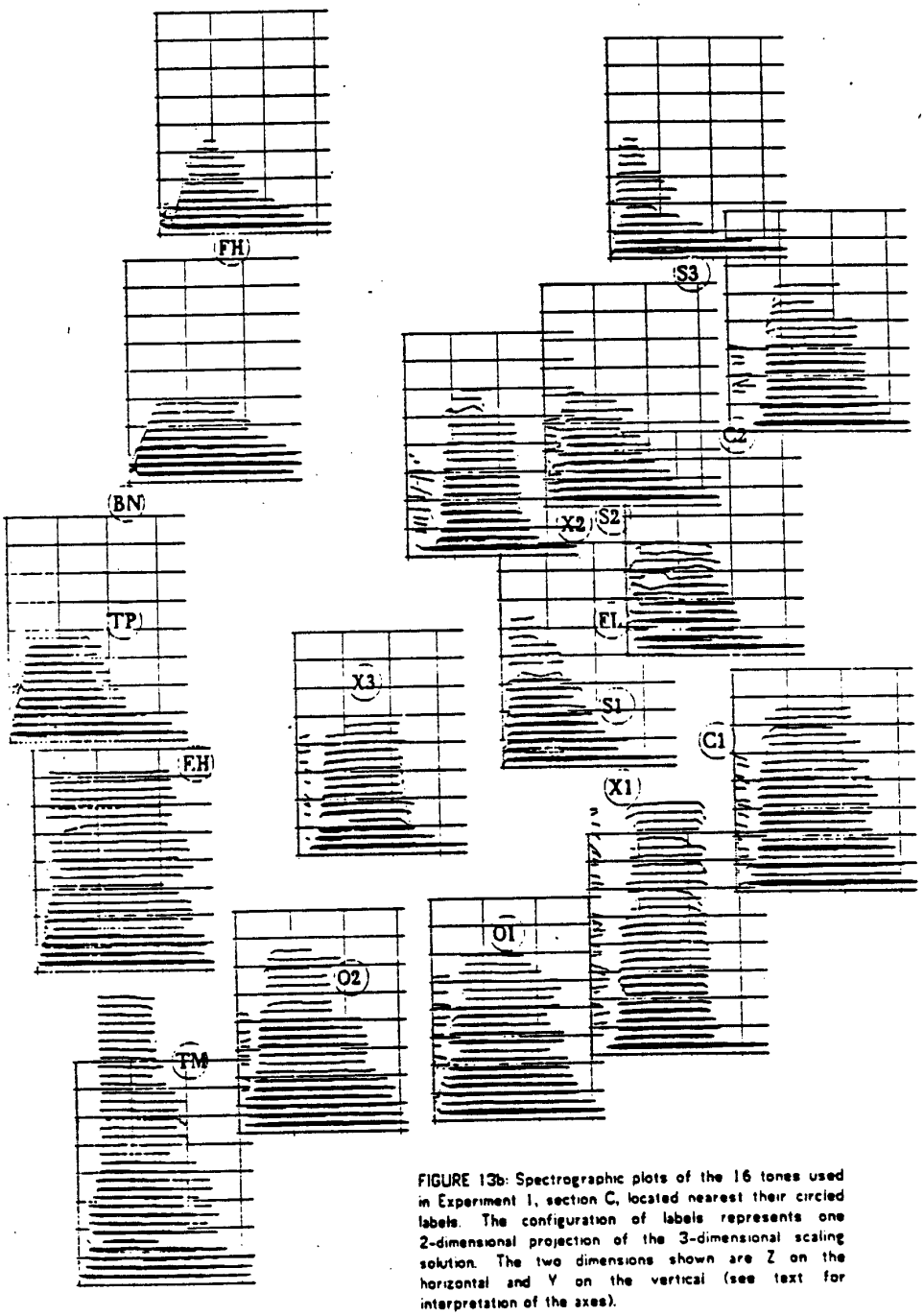



FIGURE 13b: Spectrographic plots of the 16 tones used in Experiment 1, section C, located nearest their circled labels. The configuration of labels represents one 2-dimensional projection of the 3-dimensional scaling solution. The two dimensions shown are Z on the horizontal and Y on the vertical (see text for interpretation of the axes).

The woodwinds, which are on one extreme, tend to have upper harmonics enter, reach their maxima, and often exit in close alinement. On the other extreme, the strings, brass, FL and BN, do not have harmonics which display such synchronicity, but rather tend to have tapering patterns of entrances and/or exits. It may be interesting to note that the X axis might also be partially considered to express a musical instrument family partitioning. With two exceptions, the *woodwinds* appear on the far left, the *brass* in the middle and the *strings* on the far right. The exceptions to this pattern are the clustering of BN with the brass and FL with the strings.

The Z dimension can also be interpreted in terms of temporal patterns. On this axis, the tones on one extreme, the *strings*, FL, the *clarinets* and X1, X2 and O1 display precedent high-frequency, low-amplitude energy, most often *inharmonic energy*, during the attack segment. The tones at the other extreme, which include the *brass*, BN and EH, either have low-frequency inharmonicity or, at least, no high-frequency precedent energy in the attack. The two tones on the latter side of the scale, X3 and O2, while quite near the center of the axis, show in their analyses the existence of precedent high-frequency energy. However, in an evaluation of the importance of that initial segment of the attack (the experiment in Section A), it was found that the discriminability between the existence or non-existence of that precedent high-frequency energy was .57 and .58 for the two tones, respectively. This is lower than the average discriminability of the presence of that segment for the tones on the other extreme, which was .71 accuracy.

*Conclusions*

In the present multidimensional scaling of timbral similarities between 16 time-variant musical instrument tones, *three* dimensions were required to interpret the data. One dimension related to the *spectral energy distribution*, while the other two related to the temporal pattern of the attack and decay of the tones, namely, the presence of *low-amplitude, high-frequency energy* in the initial attack segment and the presence of *synchronicity* in the attacks and decays of the higher harmonics. This interpretation is qualified with the following remarks: 1) it may be important that the energy referred to in the former dimension was *inharmonic*; 2) it is also possible that musical instrument *family* relationships were the basis of the latter dimension, as was noted above.

This study marks the first such scaling of naturalistic, time-variant tones which necessitated a three dimensional solution for interpretability. Previous studies by other investigators [Wedin and Goude, 1972 (as analyzed by Wessel, 1974); Wessel, 1974] and a pilot study performed at our laboratory used stimuli which differed from the tones employed for this study in three ways: 1) they were *longer* in duration, averaging 1 second or more rather than 350 milliseconds; 2) they were *non-processed* tones, that is, simply recordings of original tones rather than, as in the present study, computer-synthesized based on an analysis of the real tones; and 3) they were *not equalized* in the dimensions of pitch, loudness or duration.

In previous studies, the spectral distribution axis was always found. However, the MDS solutions were only in two dimensions, and the other dimension was difficult to pin down with respect to one single dimension of tone. Instead, this second dimension was interpreted as having to do with musical instrument *families*, as it showed clustering of brass, woodwind and string instruments. A physical correlation could only be made to a composite of features of attack and decay, rather than to just one acoustical dimension. It may be that in this study, these same two dimensions were also found, adding the third dimension of *high-frequency energy* in the attack. It remains an important aspect of research to determine whether the difference in dimensionality between earlier solutions and that of the present study is due to the different durations used, or the different preparations of the stimuli.

An interesting contrast exists between this scaling and the previous ones. As just mentioned, a dimension was found in three earlier solutions which consisted of instrumental family clusters. We above noted two instrumental notes in the present solution which do not conform to such a clustering, namely BN, which clusters with the brass, TP and FH, and FL, which clusters with the strings. Also note that, in the hierarchical clustering, TM grouped with the oboes, although the three dimensional solution could have it cluster with the other brass instruments as easily. It seems, then, that certain physical factors may override the tendency for instruments to cluster by family. The factors suggested here are the articulatory components of tone which occur in the attack. FL is overblown, so it has two similarities to the strings: 1) low-amplitude, high-frequency precedent (inharmonic) energy and 2) non-synchronicity of onsets. BN, likewise shares the physical properties of the brass tones with which it clusters of: 1) low-amplitude, low-frequency precedent inharmonicity and 2) tapering of onsets of the higher harmonics.

It would appear then that, just as performing instrumentalists can imitate different instruments, of course with varying degrees of success, so do instruments actually cluster in a perceptual similarity mapping. Articulatory features seem to play a central role in this non-familial clustering, but it is also important to keep in mind that the selected instrumental ranges and durations have an effect as well. Quite clearly, the use of isolated tonal stimuli is important to make note of, and it will be very important to make similar scalings for timbres that are in musical phrases, where a more complete picture of the instruments involved will emerge.

Finally, it seems necessary to reiterate the experimenter's concern for the importance of using computer-synthesized tonal stimuli in experiments on timbre perception. They have the multiple advantages of being capable of equalization in the confounding dimensions of pitch, loudness and duration, and of being exactly specified with respect to their physical features. Clearly, the potentiality of data reductions of these physical factors is of further importance in obtaining easily and perhaps even uniquely interpretable results.

It is hoped that a number of similar studies can be carried out which may test specific dimensional interpretations of this solution. With other stimuli prepared for the work outlined in Section A, for example, tones which lack the high-frequency precedent (inharmonic) energy

could be scaled instead of their counterparts in this study. Other scalings of modified stimuli, as well as stimuli from different ranges of instruments, forming different timbral subsets of instruments, and possibly occuring in musical contexts, should help to evaluate the dimensional interpretations of timbre made in this study, and ultimately, to systematically compile a psychophysical model for the distinctive features of timbre perception.

## *Experiment 2:* Confusions in the Learning of Instrument Labels

### *Stimuli*

The stimuli used in this study were identical to those used in Experiment 1, described above, consisting of 16 computer-generated timbres which were based on the analysis of instrumental notes.

### *Listeners and Procedure*

The same 20 listeners who participated in Experiment 1 also were employed for this study. All had musical training, and were familiar with the instruments of the orchestra. Many were professional orchestral players, and some had considerable experience in conducting.

This study was spread over a number of short sessions lasting about 10 minutes apiece. They were usually a beginning part of longer sessions which involved other experiments. Most listeners were able to participate in at least 4 of these sessions, many in up to 6, and a few could only be scheduled for 3 times or less.

The experiment consisted of 80 trials. On each trial listeners would hear a single note and be required to identify it in terms of the labels provided, and then get feedback on the correct answer. All listeners served in an initial session where they read the names of the notes as they heard them. In the subsequent sessions, the first 16 trials were considered practice trials, leaving a total of 64 experimental trials. Four identifications were attempted for each of the 16 tones, which were presented in a random order.

The labels were the same as the abbreviations used in referring to the multidimensional scaling study: O1 and O2 = the two oboes; EH = the English horn; BN = the bassoon; C1 = the Eb clarinet and C2 = the bass clarinet; X1 and X2 = the two saxophone tones (*mf* and *p* respectively); X3 = the soprano sax; FL = the flute; TP = the trumpet; FH = the French horn; TM = the muted trombone; S1, S2 and S3 = the cello tones (labelled strings: *sul ponticello*, normal bowing, and muted *sul tasto*, respectively).

Each trial would be preceded by a knock, which was followed 1.5 seconds later by the note for identification. There was a response interval of 4.5 seconds before the next warning knock. Listeners were required to identify the note with one of the labels above, and then to reveal the correct identity by uncovering the next entry on an answer sheet.

### *Results*

The results are presented in two simple forms. First, the overall averages in accuracy with

respect to session number is given in Table 5. There is an expected improvement in accuracy with practice.

Secondly, the overall confusion matrix is printed in Table 6. It consists of the total number of identifications which were made in all trials for all listeners. About 3 percent of the responses could not be scored, because of insufficient identification, for example, not specifying by number an instrument which formed part of a set, like the strings. Such responses were simply discarded.

*Conclusions*

A large source of the confusions found in the data was in identifying the numeric index of a tone in a subset of similarly labeled notes: S1-S2-S3, O1-O2, C1-C2, and X1-X2-X3. This was not unexpected, and the difficulties caused by the similarities of labels cannot be distinguished from the perceptual difficulties which may have been responsible for confusing the actual tones with one another.

If studies like this are done in the future, it would seem preferable to give the tones mutually exclusive labels to help eliminate this problem. It may be ideal in some situations to label them with totally neutral names, rather than particular names that may be suggestive of the tones and which would serve differentially as cues to identification.

The most interesting confusions occurred between instruments of different label types. Two of the most highly confused of such items show striking asymmetries in the matrix with respect to the tones that they were identified as being; that is, while one such tone was identified as another tone, the latter tone was not identified as being the former with near the same frequency of confusion. The two cases mentioned also relate directly to the perceptual similarities collected in Experiment 1.

The first item is X3 which was identified as EH with about the same frequency of confusion that it had in being identified as either of the other two saxophone tones, that is, a little over eight percent. The group scaling solution shown in Figure 11 places X3 in the same neighborhood as the three confused tones, about intermediate between them. The second item is BN, which was identified as FH about seven percent of the time, and as TP in about four percent of the time. Notice that in the scaling solution, the close relationship of these three tones is also manifest. The only other similarly strong case of confusion was between the oboes and EH, a confusion which occurred in all directions of stimulus-identification, and is also somewhat reflected in the similarity solution.

| Session | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Correct | 60% | 74% | 78% | 83% | 84% | 81% |
| Listeners | 22 | 20 | 18 | 15 | 12 | 6 |

TABLE 5: Improvement with training in the identification of 16 tones [Exp 2, sec C]. The Percent Correct is listed with respect to the Session Number; the total Number of Listeners that participated in each session is also listed.

*RESPONSE*

| *STIMULUS* | O1 | O2 | EH | BN | C1 | C2 | X1 | X2 | X3 | FL | TP | FH | TB | S2 | S1 | S3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| O1 | 173 | 82 | 35 | 4 | 8 | 5 | 10 | 6 | 3 | - | 8 | - | 6 | 6 | 5 | 2 |
| O2 | 115 | 218 | 24 | 3 | 1 | - | 2 | - | - | 1 | - | - | - | - | - | - |
| EH | 40 | 38 | 248 | 12 | - | - | 5 | 3 | 3 | - | 1 | - | 8 | 2 | 4 | 1 |
| BN | 1 | 4 | 8 | 305 | - | - | - | - | - | - | 14 | 26 | 9 | - | - | - |
| C1 | 1 | - | - | - | 294 | 60 | 8 | 6 | - | - | - | - | - | - | - | 1 |
| C2 | - | - | 2 | - | 77 | 258 | 10 | 12 | 6 | - | 1 | - | - | - | - | 2 |
| X1 | 1 | - | 2 | 3 | 1 | 2 | 229 | 86 | 39 | - | 1 | - | 3 | - | - | - |
| X2 | 1 | - | 2 | 3 | 6 | 8 | 67 | 231 | 39 | 1 | - | 1 | - | - | - | - |
| X3 | 6 | 9 | 29 | 4 | 3 | 2 | 30 | 42 | 236 | - | 1 | - | 3 | - | - | - |
| FL | - | - | - | - | - | - | - | - | - | 358 | - | - | - | 5 | 8 | 1 |
| TP | 1 | - | 5 | 5 | - | - | - | - | - | - | 342 | 4 | 7 | 1 | - | - |
| FH | - | - | 2 | 1 | - | - | - | - | - | 5 | 7 | 356 | - | - | - | - |
| TB | 3 | 4 | 1 | - | - | - | 1 | - | - | 1 | 9 | - | 346 | - | 1 | - |
| S2 | - | - | - | - | 1 | - | - | - | - | 3 | - | - | - | 267 | 74 | 24 |
| S1 | 6 | 2 | 3 | 6 | 1 | - | - | - | - | 7 | 2 | - | 1 | 57 | 263 | 9 |
| S3 | - | - | - | 1 | 2 | - | - | - | - | 1 | - | 2 | - | 26 | 15 | 320 |

TABLE 6: Confusion matrix from identification/learning study [Exp 2, sec C]. The 16 tones were presented and identifications attempted, with feedback. This matrix represents the total number of responses for all non-practice trials (there were 22 listeners that participated in up to 6 sessions apiece).

The potentially informative aspect of the first two asymmetrical confusions is the way in which they may relate to timbral dominance or neutrality for identification. Consider, for example, BN, which was confused with the two brass tones mentioned above, and also rated as being most similar to them in Experiment I, thus taking it out of its musical family environment of the woodwinds. This would seem to indicate that the features which existed in this particular note made it hard to identify the source as being a bassoon. It was taken from the high extreme of the instrumental range, and this certainly must have contributed to its identification with brass tones. It is often noted in musical performances that the high-range of the bassoon seems quite similar to a trumpet line in the same range, hence it is probably not the case that the findings of this study is strictly a result of the particular tones utilized. Clearly, *in-context* studies will greatly expand the data base relating to the phenomena of timbral dominance in identification.

## D. An Exploration of the Perceptual Continuity of Timbral Transitions between Familiar Music Instrument Tones

*Introduction*

The existence of *categorical perception* with familiar musical sounds, while being an interesting matter in its own right, is an important question for perceptual psychology in general. In the literature, categorical perception has been the issue of many controversies. Of particular interest is whether the existence of a categorical mode of perception for certain speech sounds may be taken as evidence for the manner in which speech is decoded [see Studdert-Kennedy, 1970].

Investigators claiming to dispute the matter have sought out other cases of categorical perception to show that it may be a feature of perceptual processing under many sets of general circumstances [Lane, 1965]. Looking at musical perception, some findings have suggested that categorical perception exists for musical intervals [Burns and Ward, 1974]. Cutting and Rosner [1974] found categorical perception for adult listeners with saw-tooth stimuli that differed in rise time, heard as plucked or bowed notes. Interestingly enough, they later found categorical perception in two-month-old infants as well, and they interpreted the results as undermining the basic notion of categorical perception [Jusczyk, et. al, 1974]. (Bever [informal report, 1974] found categorical perception in the responses of infants to musical intervals.) The possible existence of categorical perception for timbre is of particular interest in our current research because it has a bearing on some of the properties of the spatial analogy utilized for timbre.

If perception were *continuous* in the timbre space, then a set of tones generated by a regular acoustical interpolation between two familiar instrumental notes in the space would be heard as a gradually changing transition in timbre between the two extremes. If timbre perception were *categorical*, however, then the acoustically even interpolation would be perceived discontinuously, suddenly jumping from the first tone to the second tone, the tones in between all taking the identities of one of the two extreme tones, with little or no ambiguity even in the center of the transition.

A related effect of categorical perception of a transition would occur with a sequential presentation of the interpolation. In this case, categorical perception would suggest that it does not matter in which direction the interpolation is to be presented, there would be no change in the location of the sudden shift in identity of the tones from that of the beginning tone to that of the ending tone. If, however, the transition were continuous, then there may be an expected directional effect, whereby the initial identification holds on longer than it would if it were in the terminal position. There would in effect be a *hysteresis* in the identification function, the location of cross-over in identification shifting with direction of presentation.

Another measurable correlate to categorical perception would be the discriminability of adjacent tones in the interpolation. Tones which were separated by *equal acoustical intervals*, in the interpolation through multidimensional physical property space, would show some non-

monotonic discrimination function if perceived categorically. If both members of a pair of tones fell within the same perceptual category, they would be far less discriminable than if they came from two distinct categories, that is, fell across the categorical boundary in the transition.

The development of algorithms for the interpolation between two given music instrument tones is necessary in order to examine the existence of categorical perception with timbre. For this study, an algorithm based upon the additive synthesis model was formulated, in which tones are physically represented as sums of harmonically related sinusoids whose amplitude functions are time-variant and frequencies are fixed constants. For continuity with previous research, presented above, the amplitude functions were simplified line segment approximations to the originally analyzed functions. The interpolation procedure operated upon the frequencies of each harmonic and shapes of their respective amplitude functions, as will be described below.

In that an interpolation is generated through multidimensional physical space in the case of timbre, there exist many paths between two given points, and the selection of some particular algorithm is only the selection of one particular path. In the following study, it should be noted that no *ultimate* proof or disproof of categorical perception is forthcoming when interpolating in multidimensional physical domains, because of the lack of a unique solution. There are no naturally occuring interpolations between discrete musical instruments, and we have little experience with the concept of transitions in this domain. It may be important to note that this case is equally true in speech research on consonants where the interpolations are in multiple domains. The naturally occuring interpolations in speech seem to be with vowel sounds, and the uniqueness of a path is determined by the physical properties of speech production. Bearing this in mind, the following study is seen as a research project on a specific interpolation algorithm as it relates to categorical perception for timbre.

Four aspects of research have been pursued in this study. Three of these are referred to above, with the three measurable effects of categorical perception: 1) *perceptual cross-over points* in identification for sequential interpolations and the existence of directional effects; 2) *identification* functions for isolated tones and the sharpness of the boundary conditions that are measured; and 3) *discrimination* functions for pairs of equally spaced tones and the effect of location with respect to identification boundaries and categories. The fourth approach to the matter has been the measurement of *perceptual similarities* for a set of interpolation midpoints and a subset of original tones, with respect to the set used in Section C. This data was analyzed with multidimensional scaling algorithms and compared to the structures discussed in Section C.

## *Experiment 1:* Hysteresis as a Measure of Continuity in Timbral Interpolations

*Stimuli*

The stimuli consisted of 6 sets of interpolated tones, generated by a general interpolation algorithm between 4 familiar music instrument tones. These 4 familiar tones were selected from the notes described and utilized in previous research, using the analysis-based additive synthesis and data reduction techniques. The data representation of these 4 tones was an extremely simplified version, using small numbers of line segments to approximate the originally analyzed amplitude functions for the harmonics of a tone, and constant values to represent their frequencies.

The actual 4 tones selected were equivalent in representation to the *constant frequencies approximations* used in the work in Section A. They were, however, also subject to the findings of the *equalization* experiments in Section B when frequencies, amplitudes and durations were specified for synthesis. The 4 tones utilized for interpolation were C1, O1, FH and S3 of earlier studies. These were selected because of their relatively low discriminability from their *line segment approximation* counterparts used in the work of Section C, as determined by the measurements of Section A. Their respective accuracies of discrimination were .65, .64, .60 and .56, among the lowest scores obtained for this comparison.

The interpolation algorithm was called with the respective physical representations of the two end-point tones and a number between 0 and 1, giving the fraction of interpolation desired between the given extremes. This algorithm attempts to: 1) interpolate the point in time at which the maximum is reached in an amplitude function *temporally* as well as in amplitude between the two end-points; 2) perform a weighted average on the activity before the maxima of the two end-point functions to generate the attack segment of the interpolated tone, with respect to the fraction specified. If a detectable segment of activity occurs in the initial part of the attack, it will be handled as precedent activity to the constructed function, and will be scaled in magnitude and duration with respect to the weighting - this provides for the tones which have segmented precedent activity in many harmonics; and 3) perform a similar weighted average on the activity following the maxima to generate the decay segment. This process is graphically displayed in Figure 14. The constant frequency determined is the weighted average of the two end-point values for that harmonic.

Each interpolation was composed of 10 percent steps. There were 11 tones, in each transition set. Stimuli were recorded from the computer and played back to listeners with the same procedure specified in Section A.

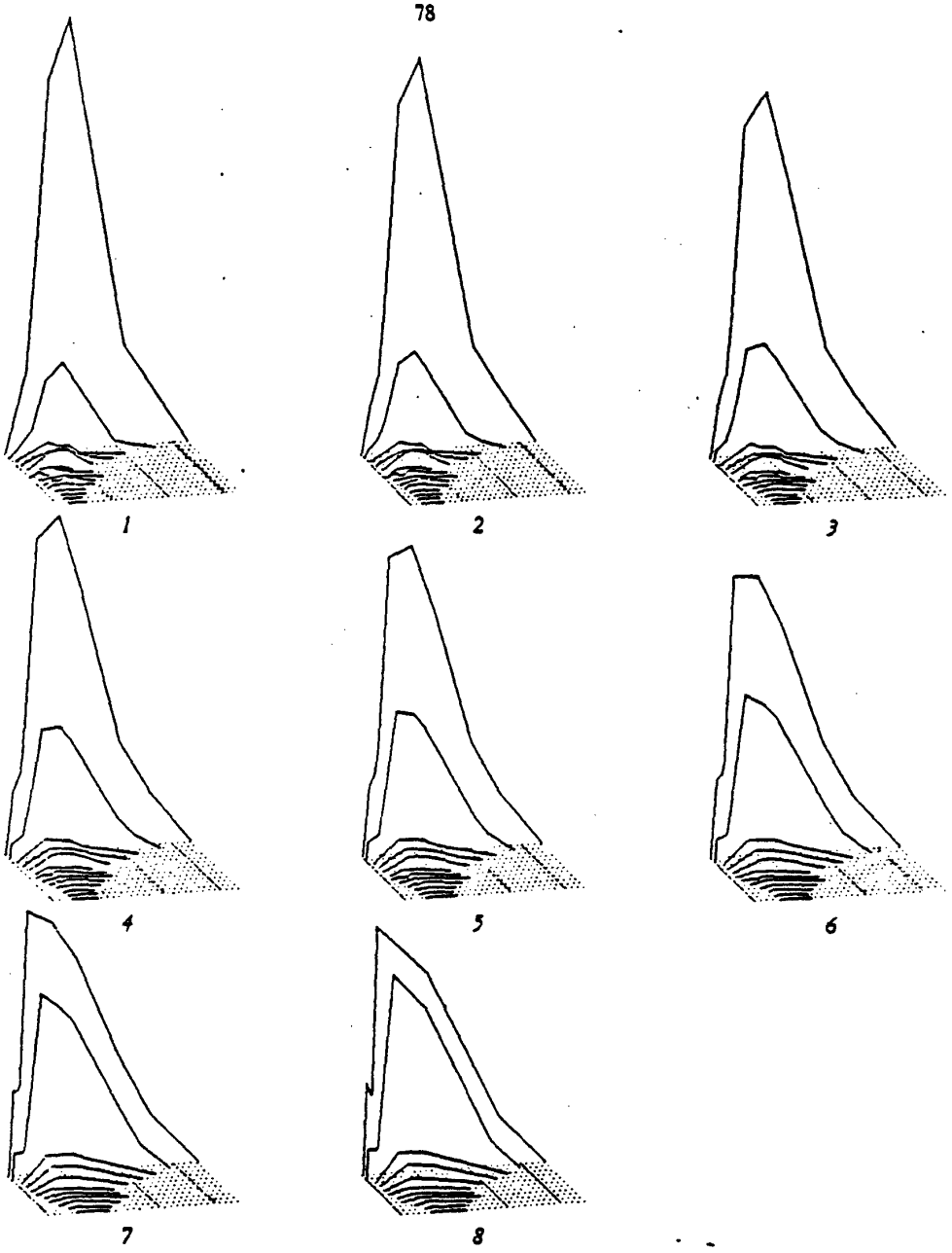



FIGURE 14: Example of timbral transition in 14% steps [basis of the stimuli used in section D]. Shown above are the end-point tones S3 (=1) and FH (=8), along with the six interpolated tones between. Amplitude x Frequency x Time plots display the physical changes in the interpolation (see text for explanation of how the algorithm works).

*Listeners and Procedure*

Eighteen listeners at Stanford University were employed for this experiment. Listeners were musically sophisticated, some actively involved in advanced instrumental performance and others in conducting or musical composition. Several listeners had well-trained ears with respect to computer-generated music, and all listeners had participated in other phases of this research on timbre perception. The physical arrangements of the laboratory were as described in Section A.

There were 60 trials in this experiment, lasting 14 minutes in all. Each trial consisted of a sequential presentation of 11 tones, an interpolation in timbre between two of the 4 familiar music instrument notes used. There were a total of 6 interpolations between the 4 tones, and taking direction as a factor of presentation, this made an overall total of 12 basic sequences. Five repeated measurements were obtained, making the total of 60 trails.

An individual trial lasted for 14 seconds. A warning knock preceded the trial by 1 second, and the 11 tones were heard at 1 second intervals. 3 seconds were given for the listener to respond to the sequence. Sequences were chosen randomly in blocks of 12.

Listeners were instructed that a sequence of 11 notes would be played, starting with one timbre and ending on another. They were told the names of the two end-point timbres, and were asked to indicate the number of the tone in the sequence which marked the *first appearance* of the second timbre by circling one of a set of numbers from 1 to 11 on the score sheet. They were also asked to indicate on a scale of 4 how *sudden* the change seemed, where 1 was extremely sharp and 4 was extremely gradual.

*Results*

Mean scores were derived for each of the 12 basic types of interpolated sequences, where direction was taken as a factor. The average cross-over points and suddenness of transition ratings are given for the 12 sequences in Table 7. The cross-over points are presented graphically in the set of diagrams in Figure 15.

The overall averages shown in Figure 15 are represented by the two marks placed above each rectangular grid appearing down the left column of the figure. Each grid represents a single physical interpolation, for both directions. The 11 vertical partitions in each grid represent the successive 11 tones in that interpolation. Various other data from the next two experiments is also plotted within these grids, some of which includes labels of the tones which were selected as end-points in the transitions. For example, the first plotted grid includes the labels c and o; this was the interpolation between the clarinet, Cl, and the oboe, Ol. The other two labels are s for the string, S3, and h for the French horn, FH.

| END POINTS | BACKWARDS | | FORWARDS | |
|---|---|---|---|---|
| | cross-over | (rate) | cross-over | (rate) |
| C1 - O1 | 2.2 | (3.1) | 6.3 | (3.0) |
| C1 - FH | 3.1 | (2.6) | 5.9 | (2.6) |
| C1 - S3 | 3.8 | (2.8) | 7.1 | (3.0) |
| O1 - FH | 2.4 | (3.0) | 6.4 | (2.8) |
| O1 - S3 | 2.8 | (3.1) | 6.7 | (3.1) |
| S3 - FH | 2.3 | (3.2) | 4.9 | (2.4) |

TABLE 7: Cross-over points in identification for the intermediate tones of interpolations [Exp 1, sec D]. Listeners indicated which tone in the 11-note sequence marked the initial appearance of the second timbre. The average perceived Cross-over points are listed above as a score from 0 to 10 (where 0 = the first tone of a sequence and 10 = the last tone of a sequence). There were 12 interpolations derived from 6 pairs of End-points (listed in the left column) which could be played either Forwards or Backwards (listed above as two sets of columns). The Rates at which transitions were perceived to change was also judged, and the averages are listed above in parentheses (where 1 = very sudden and 4 = very gradual).

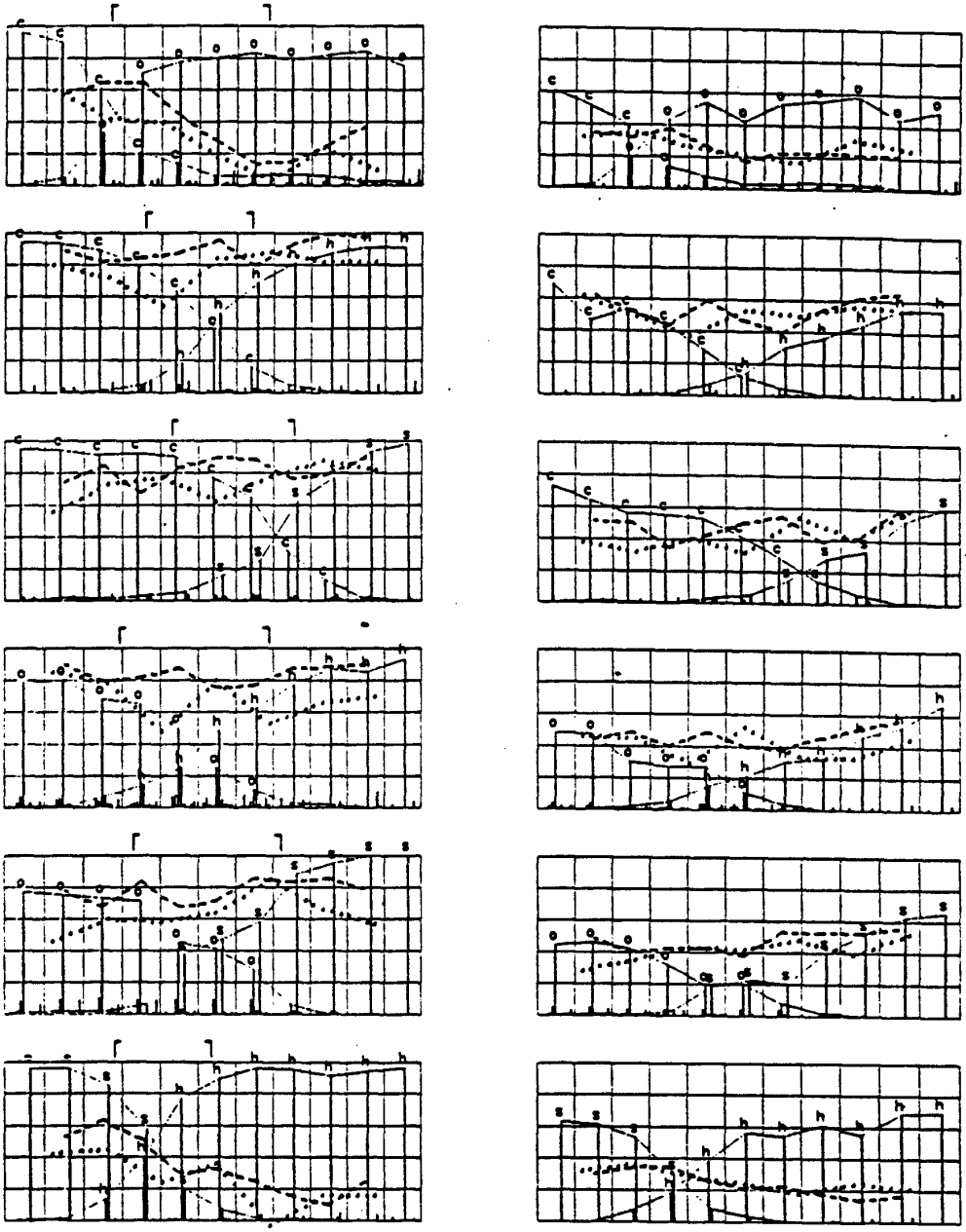


FIGURE 15. LEFT: Percent identification (Y-axis: 0-100%) of interpolated tones (X-axis partitions) between labelled endpoints in 107 steps [Exp 2, sec D]. Inner-partition histograms show responses to each tone using the 10 available identification categories (from left: flt, trp, trm, clr, obo, hrn, str, sax, bsn, eng). Discrimination functions (Y-axis: 50-100%): dashed=1/5 step, dotted=1/7 step [Exp 3, sec D]. Markers for cross-over estimates in sequential interpolations [Exp 1, sec D] appear on TOP of frames (the hat points in the direction of the starting tone). RIGHT: Goodness ratings (Y-axis: 0-100%) of interpolated tones [Exp 2, sec D]. Distance functions (Y-axis: 0-100%): dashed=1/5 step, dotted=1/7 step [Exp 3, sec D].

The two marks placed above this grid reveal the mean positions of perceptual cross-over going in the two possible directions in the transition. The hat on top of the vertical marker points toward the *first* tone presented in the interpolation and hence indicates the direction of the transition.

The main finding of this experiment is the definite presence of hysteresis in all transitions. The range of directional effects was from 25 to 45 percent of the extent of the interpolations. Another finding is that not all of the areas of cross-over are centered at the 50 percent mark of the interpolations. Finally, an examination of the suddenness ratings shows that the transitions tended to be perceived as more gradual than sharp, in that the average responses do not hover about 1.0, the *sudden* end of the rating scale.

*Conclusions*

The finding of a strong hysteresis effect with direction of the transition between two familiar timbres indicates that there is not a sharp perceptual boundary for identification of the interpolated tones. Suddenness ratings agree with this finding, generally reflecting that listeners heard the transitions as more gradual than sudden.

The lack of sharp categorical boundaries and the sensitivity to context of the identification functions would suggest that, strictly defined, these transitions are not perceived categorically. If they were, then we would expect to find sharp boundaries in the identification functions, with little discrimination between tones which fall within the same category. In this study, however, the rather gradual nature of the perceived transition suggests that there was ongoing discrimination between adjacent tones near the area of cross-over. The distance that separates the cross-over points for the two directions serves to spread this discriminability and also can be taken as a crude indication of the steepness of the identification boundaries. In these transitions, that boundary is generally very wide, hence one might conclude from this study that the interpolation is heard more continuously than categorically.

## *Experiment 2:* Identification of Isolated Tones selected from Timbral Interpolations

*Stimuli*

The stimuli used in this experiment were identical to those described above in Experiment 1. They consisted of 6 sets of interpolated tones, 11 tones per interpolation. All other details were identical to those described for Experiment 1.

*Listeners and Procedure*

Fifteen listeners at Stanford University were employed for this experiment. Listeners were musically sophisticated, some actively involved in advance instrumental performance and others in conducting or musical composition. Several listeners had well-trained ears with respect to computer-generated music, and all listeners had participated in other phases of this research on timbre perception. The physical arrangements of the laboratory were as described in Experiment 1.

There were 436 trials in this experiment, lasting 44 minutes in all. In each trial a single tone was presented for identification. The tone was preceded by 1.5 seconds with a warning knock, and listeners were given 4.5 seconds to respond following the tone. There were 40 practice trials and 396 experimental trials, making a total of 6 repeated measurements on each of the 11 tones in the 6 interpolations. The stimuli were presented in a random order with the restriction that no two adjacent trials would have the same interpolation as the source of the stimuli.

Listeners were told to select the name of the tone from a list of ten possibilities: clarinet, oboe, English horn, bassoon, flute, saxophone, French horn, trumpet, trombone or string. They were told that the tones were similar to the 16 instrument notes used in earlier experiments. They were also instructed that there would be a smaller set of tones played than the number of alternatives given for identification.

Listeners were additionally instructed to indicate on a 7-point scale the degree to which the tone corresponded to the category of identification. This latter measure was considered to be a *goodness* rating of the tone.

*Results*

Identifications and goodness ratings were averaged over the 6 replications for each of the 11 tones in the 6 interpolations. These results are presented graphically in Figure 15. The two columns containing 6 grids each represent the identification scores and the goodness ratings, from left to right. The 6 rows of identification and goodness grids represent the 6 different interpolations. As explained above, each grid has 11 partitions, each partition representing a successive tone in the transition.

Within each partition there is a histogram seated on top a set of 10 hairs, which protrude from the bottom of the grid. These ten hairs correspond to the 10 categories given for identification: flute, trumpet, trombone, clarinet, oboe, French horn, string, saxophone, bassoon and English horn, from left to right. The height of a bar above one such hair corresponds to the percentage of times that the category was used to identify the tone of that partition. The ordinate should be read from 0 to 100 percent, horizontal grid marks indicating jumps of 20 percent. For the grids in the right column, the ordinate represents the respective goodness ratings, and should be read as a corresponding percentage of the 7-point scale.

The bars which correspond to either of the two end-point tones for the immediate interpolation are labelled by the appropriate name when the percentage of identification was greater than 10 percent. These labels are: c for clarinet, o for oboe, h for French horn and s for string. To help observe the relative identification and goodness scores for the tones of a particular interpolation with respect to the two end-points, these labelled bars are connected by lighter lines. The two lines that cross within each grid in the left column display the identification function for the appropriate interpolation, the cross-over point is the categorical boundary, and the steepness of the curve corresponds to the sharpness of the categorical boundary.

*Conclusions*

The most striking result of this study is that, although listeners could choose freely from 10 alternatives to identify the inner tones of the interpolations, a clear majority of responses used only the two end-point names. It is important to recall that the listener had no way of knowing what the two end-points were that may have generated a particular isolated tone, and that this was not a two-choice task. We may conclude that the tones which were generated by the interpolation algorithm indeed were identified as one of the two tones used in their construction, and furthermore, that there is an orderly rate of change in identification from one end to the other. This could be considered categorical behavior in a loose form, where these tones were most readily perceived in terms of the features that were preserved through the interpolation, rather than taking on totally distinct and unrelated identities.

In the more narrow sense of categorical perception, however, the present study tends to indicate that these interpolations were more continuously perceived. The lack of sharp identification boundaries, and the corresponding slow roll-off in the goodness functions indicates a gradual transition rather than a sudden, categorical shift in identification. The goodness functions should be considered a percentage modifier on the identification functions, which themselves tended to be gradual in form.

It is particularly interesting to compare the two cross-over points found in Experiment 1 for each of the sequential transitions with the cross-over points found in this study which were based on identifications of the isolated tones. In all 6 interpolations, the isolated cross-over point indeed falls between the two sequentially-derived cross-over points. This would suggest that all three points measure a categorical boundary held in common to the different tasks, the outer two, sequentially-derived points show the extremes of the boundaries and the cross-over point obtained with the isolated tones presumably gives the true midpoint of the boundary. Asymmetries are found in both functions.

The identification responses which were not one of the two end-point names had some interesting properties. They originated mostly with three or four of the listeners, and had strong correlations with the similarities earlier measured for 16 instrument notes in Section C. For example, saxophone was most often given near the clarinet side of one of the three transitions

which included the clarinet as an end-point. Other common confusions include English horn for oboe, trombone for oboe and bassoon for French horn, all of which correspond nicely to the similarity judgments between those instrumental tones in Section C.

### *Experiment 3:* Discrimination of Pairs of Tones selected from Timbral Interpolations

*Stimuli*

Two sets of stimuli were used in this study. The first consisted of the same tones generated for Experiments 1 and 2, and will be referred to as the 10 percent steps in the interpolations, in that the transitions were divided into ten parts by these tones. The actual discrimination task compared alternate pairs of tones, so 20 percent steps were utilized in the test.

A second set of tones was generated at a little more than 14 percent steps, the transitions being divided into seven equal parts. These were used in a second experiment in this discrimination study to measure different sized steps. All other features of the stimuli were identical to those in the previous experiments.

*Listeners and Procedure*

The same set of listeners as participated in Experiment 2 was employed for this study. The details of the laboratory situation have been specified above.

There were 285 trials for the 20 percent steps and 225 trials for the 14 percent steps. Two sessions were run; the first session lasted for 48 minutes and tested alternate 10 percent steps, the experiment therefore measuring 20 percent steps, and the second session lasted for 38 minutes and tested adjacent 14 percent steps. The first 15 trials in each session were practice material. Five repeated measures were taken for each pair of tones in each of the respective interpolations.

The discrimination task employed was the AAAB procedure outlined in Section A. This was a multiple task consisting of a *discrimination* response and a *subjective distance estimation* corresponding to the relative degree of difference heard. All other procedural details were identical to those described in that section.

*Results*

Discrimination scores and distance estimations were averaged as described in Section A. The results are presented graphically in Figure 15, plotted as dotted and dashed lines inside the grids. Discrimination functions are plotted in the left column and distance estimations in the right column. The grids are divided into 11 partitions which represent the 11 tones that

comprise the 10 percent steps in the interpolation. The 6 rows represent the 6 different transition types used, and include the identification labels of their end-points.

The dashed lines represent the 20 percent discrimination and distance functions, while the dotted lines represent the 14 percent functions. The ordinate is to be read from 50 to 100 percent accuracy for the discrimination scores, and as a percentage of a 7-point distance estimation score. Each score is positioned on the abscissa midway between positions of the two tones employed in the task. Hence, the 20 percent score is positioned immediately between the two alternate partitions that represent that pair of tones, falling on the intermediate partition. The abscissa was divided into 14 percent divisions to plot the 14 percent scores.

*Conclusions*

Little evidence was found in this study for a corresponding peak in discrimination at the categorical boundaries of an interpolation. These results may be taken with caution, however, in that four of the transitions appear to have resulted in ceiling effects in discrimination, that is, there was high discrimination for all pairs. The two functions which did not display ceiling effects rather showed tendencies toward the floor of 50 percent, chance-level discrimination. It may be significant that this floor-level performance was in both cases firmly embedded within a categorical boundary. Only one of the two transitions shows a peak near the boundary, while there is a slight mismatch with the other interpolation. In that both discrimination functions are heavily skewed, corresponding to pronounced skewing in the identification functions, we did not uncover any classical discrimination curves as have been found in the speech literature [see Liberman, 1967].

It may be significant to note that the distance estimations revealed no tendency to peak at boundaries, but rather look relatively flat. They do not show the same ceiling or floor effects as found with discrimination scores. These results suggest a more continuous transition, rather than the classical categorical effect of little or no discrimination within boundaries but pronounced discrimination between boundaries. To make a thorough test, both larger and smaller sized steps must be used.

Aside from the above considerations, the discrimination functions do reveal some interesting properties. As was expected, the smaller sized steps were more poorly discriminated. Also, there is a striking correspondence between the average level of discrimination for an interpolation and the corresponding distance between its end-points in the similarity space derived in Section C. The two closest pairs of end-points in the space, C1-O1 and FH-S3, gave the worst discrimination scores which approached floor-level performance, while the more widely separated pairs of end-points had ceiling effects. This again seems to point out a type of continuity which exists in the *timbre space.*

## *Experiment 4:* Similarity Measurements for a Set of Naturalistic and Interpolated Tones

### *Stimuli*

The stimuli were selected to form a subset of the tones based upon natural music instrument notes used in Section C, added to a subset of the interpolated tones employed in Experiments 1 through 3. The midpoints of the 6 interpolations were selected, and the 4 points on which those interpolations were based, C1, O1, FH and S3 were selected from the stimuli used in Section C. We will refer to these interpolated midpoints with the labels of CO, CH, CS, OH, OS and HS, for C1-O1, C1-FH, C1-S3, O1-FH, O1-S3 and FH-S3, respectively.

Among the 8 other naturalistic tones used in this study, also taken from the stimuli of Section C, were C2, S2, TP, TM, EH, X1, X2 and BN. This made a total set of 18 stimuli for comparison. All other features of the recording and playback of stimuli were as described in Section A.

### *Listeners and Procedure*

Seventeen listeners at Stanford University were employed for this experiment. Listeners were musically sophisticated, some actively involved in advance instrumental performance and others in conducting or musical composition. Several listeners had well-trained ears with respect to computer-generated music, and all listeners had participated in other phases of this research on timbre perception.

The physical arrangements of the laboratory were as described in Section A. The exact procedure employed in this study was described in Section C, Experiment 1, for the collection of similarity judgments. There were a total of 336 trials, lasting for a session length of 54 minutes. Of these, 30 trials were practice, leaving a total number of experimental trials equaling 306, which is $n*(n-1)$, using both directions.

### *Results*

The similarity data collected was treated as described in Section C, Experiment 1, excepting the differing total number of stimuli. Two and three dimensional group and individual solutions were obtained from multidimensional scaling algorithms, by the individual differences scaling (referred to below as INDSCAL) [Carroll and Chang, 1970] and nonmetric algorithms [Shepard, 1962a, 1962b; Kruskal, 1964a, 1964b], respectively. The individual solutions were tested against the group solutions in those respective dimensionalities and congruency scores observed. In addition, group and individual similarity matrices were treated with hierarchic clustering analyses [Johnson, 1967; see also Shepard, 1972].

The group two-dimensional solution with embedded clustering is shown in Figure 16. An examination of the spatial solution with respect to the clustering reveals discrepancies. Notable among these conflicts between spacial location and clustering is the close spatial proximity of the three non-clustering items OS, X2 and CH, which we would not expect to be among the very closest of points in the similarity space. Likewise, the spatial proximity of X1-CO and C2-CS are unexpected and in conflict with clusterings.

As was found in Section C, the group three-dimensional solution fares much better in terms of agreement with the clustering analysis, as well as with respect to its interpretability. The perspective plot of this three-dimensional configuration is shown in Figure 17, which also includes the clustering analysis, plotted as described for Section C. Plots of the physical characteristics of the tones are put into two-dimensional projections of the three-dimensional space in Figures 18a and 18b for the Time x Frequency x Amplitude representations and in Figures 19a and 19b for the spectrographic displays.

An interpretation similar to that of the solution for the whole set of naturalistic tones in Section C was sought. A good agreement was found in explaining the psychophysical bases of two of the dimensions. As labelled in Section C, the Y axis again displays an ordering with respect to the *spectral energy distribution* of the tones, and the Z axis can be interpreted on the basis of the presence of absence of precedent *low-amplitude, high-frequency (inharmonic) energy* in the attack segment. Subjecting the configuration to a slight rotation about Z brings the Y dimension more closely in line with the ordering of tones obtained with the full set of naturalistic instruments (in spite of the fact that rotation is difficult to justify with INDSCAL, we found that just such slight degrees of rotation might occur if small subsets of subjects were excluded from the input data).

While no clear physical interpretation has been made for the X axis, when the configuration is subjected to this slight rotation, it would appear that it is almost segmented in half on the basis of whether or not tones were involved in interpolations, being either *midpoints* or *end-points*. On the right we find the 4 naturalistic end-points and the 6 midpoints, while on the left we find the remaining tones. S2 is the only exception, clustering close to S3, but nevertheless, near the dividing line.

*Conclusions*

In the present study a set of 18 musical tones were rated for pairwise similarity, and the resulting data structure treated with multidimensional scaling. Particularly important to note was the composition of the set of stimuli, 12 of which were naturalistic musical instrument notes, and the remaining 6 were derived by generating the midpoints between a particular 4 of these 12 naturalistic notes, using a timbral interpolation algorithm. The 12 naturalistic tones comprised a subset of 16 naturalistic tones whose similarities were scaled in a previous study discussed in Section C.

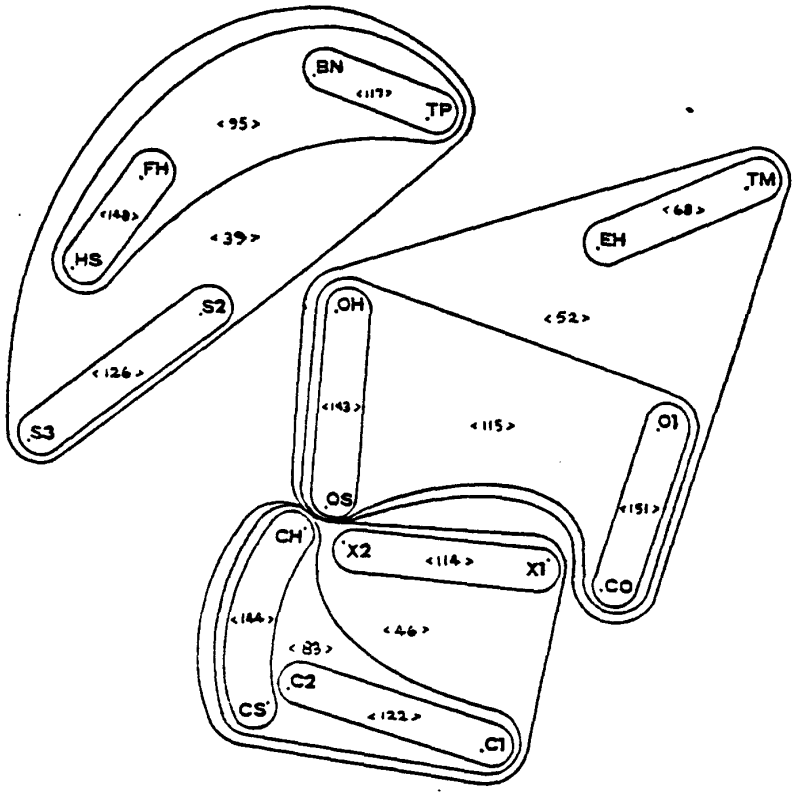



FIGURE 16: Two-dimensional group configuration for 17 similarity matrices [Exp 4, sec D] generated by individual differences multidimensional scaling program (INDSCAL [Carroll and Chang, 1970]). The results of a Hierarchical Clustering analysis [Johnson, 1967] are represented in the configuration space, as curved forms around different clusters. The levels of clustering are given inside the respective curved forms that surround each cluster. (The levels are quantitatively related to a rank-ordering of 120 items = N-(N-1)/2 pairs.) Abbreviations for stimulus points: O1=oboe; C1,C2=clarinets; X1,X2=saxophones; EH=English horn; FH=French horn; S2,S3=strings; TP=trumpet; TM=trombone; BN=bassoon; and the interpolation-derived points: CO=c1-o1; CS=c1,s3; CH=c1-fh; OH=o1-fh; OS=o1-s3; HS=fh-s3.

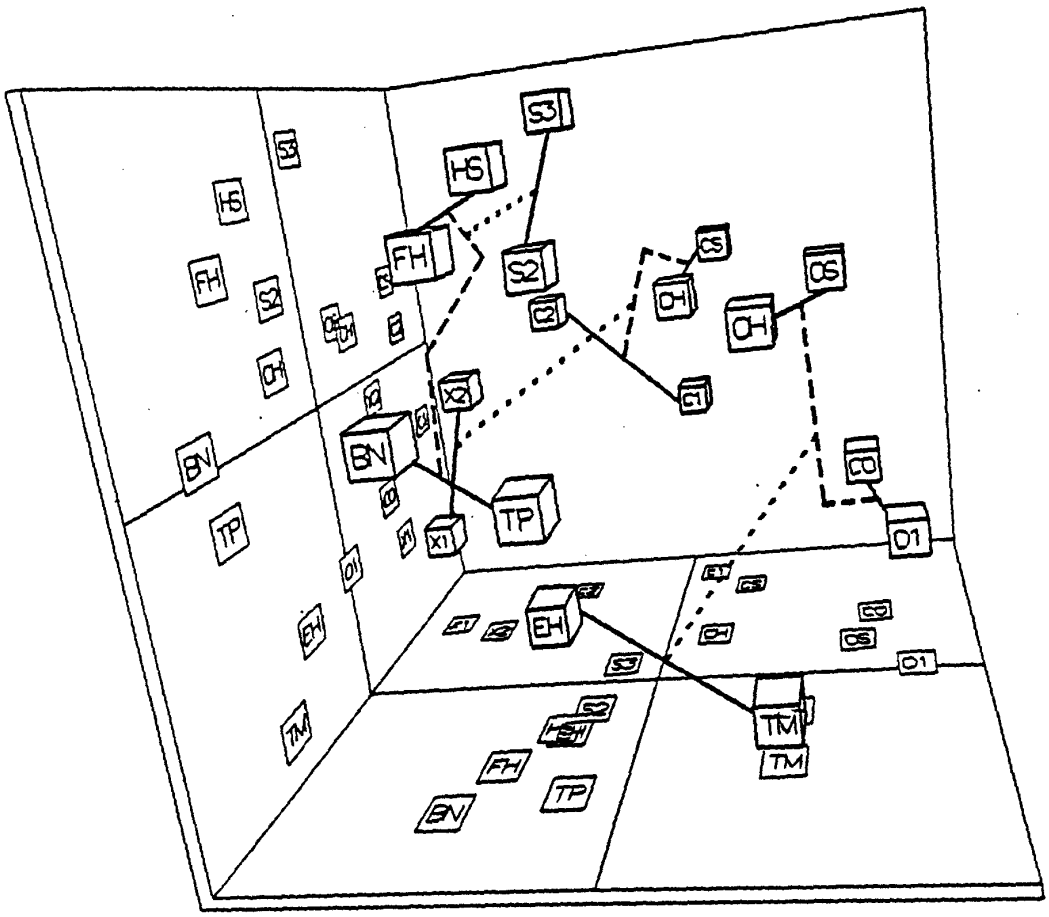



FIGURE 17: Three-dimensional group configuration for 17 similarity matrices [Exp 4, sec D] generated by individual differences multidimensional scaling program (INDSCAL [Carroll and Chang, 1970]). Hierarchical Clustering analysis [Johnson, 1967] is represented by connecting lines, in clustering order solid - dashed - dotted. Notice the two-dimensional projections of the configuration on the wall and floor. The 18 stimuli included 6 interpolated midpoints between the four end-points: O1, C1, S3 and FH. Abbreviations: O1=oboe; C1,C2=clarinets; X1,X2=saxophones; EH=English horn; FH=French horn; S2,S3=strings; TP=trumpet; TM=trombone; BN=bassoon; C0=c1-o1; CS=c1-s3; CH=c1-fh; OH=o1-fh; OS=o1-s3; HS=fh-s3.

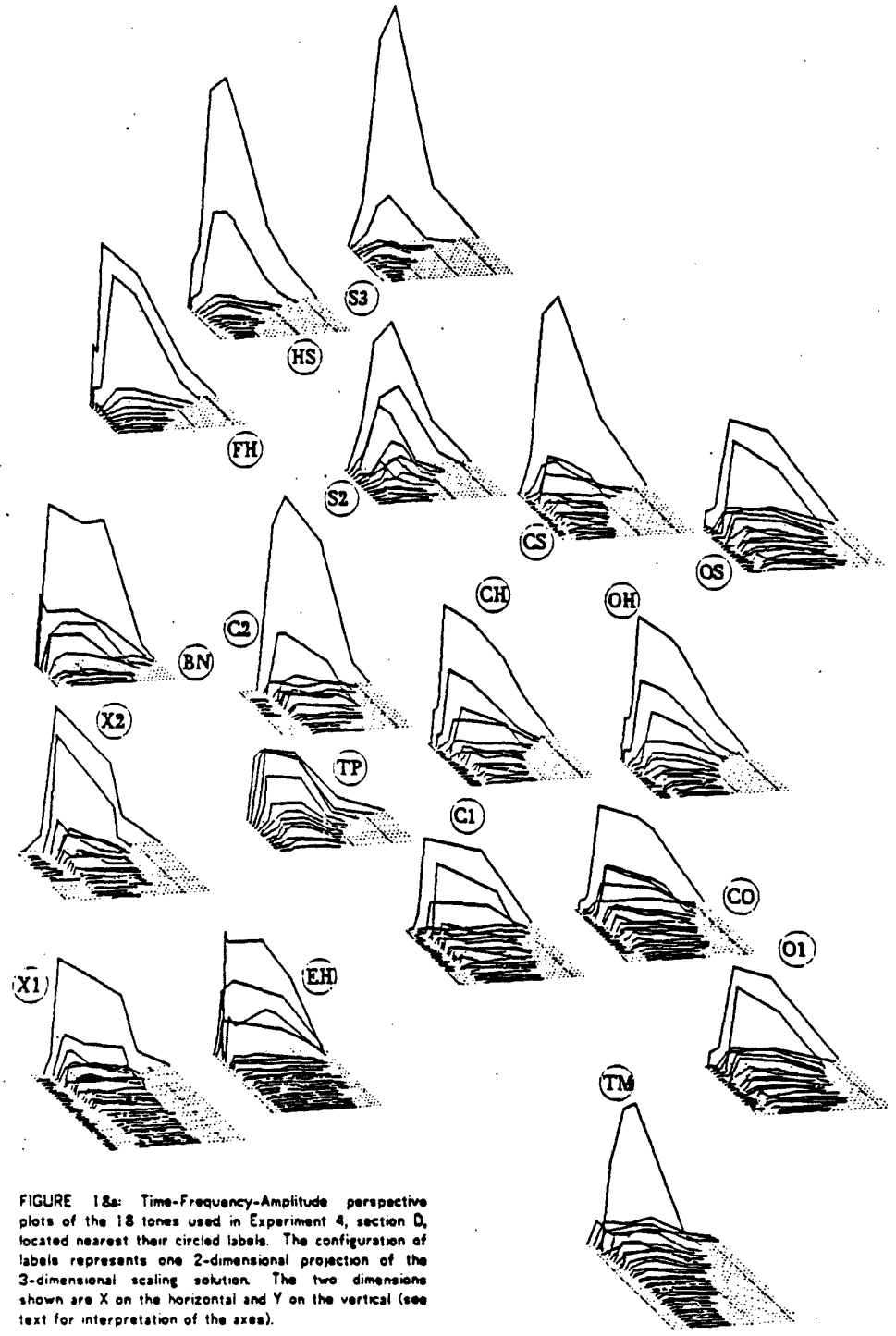



FIGURE 18a: Time-Frequency-Amplitude perspective plots of the 18 tones used in Experiment 4, section D, located nearest their circled labels. The configuration of labels represents one 2-dimensional projection of the 3-dimensional scaling solution. The two dimensions shown are X on the horizontal and Y on the vertical (see text for interpretation of the axes).

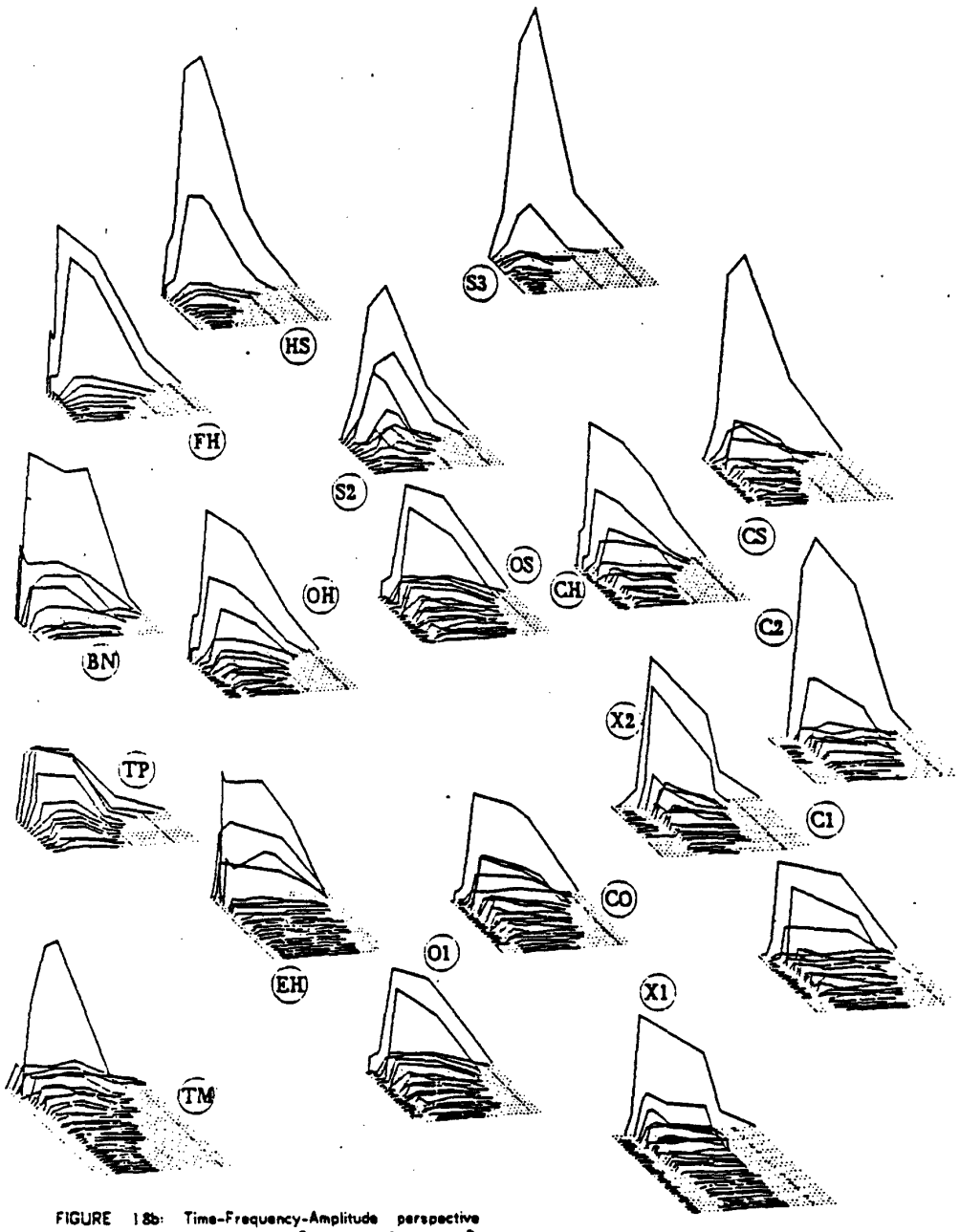



FIGURE 18b: Time-Frequency-Amplitude perspective plots of the 18 tones used in Experiment 4, section D, located nearest their circled labels. The configuration of labels represents one 2-dimensional projection of the 3-dimensional scaling solution. The two dimensions shown are Z on the horizontal and Y on the vertical (see text for interpretation of the axes).

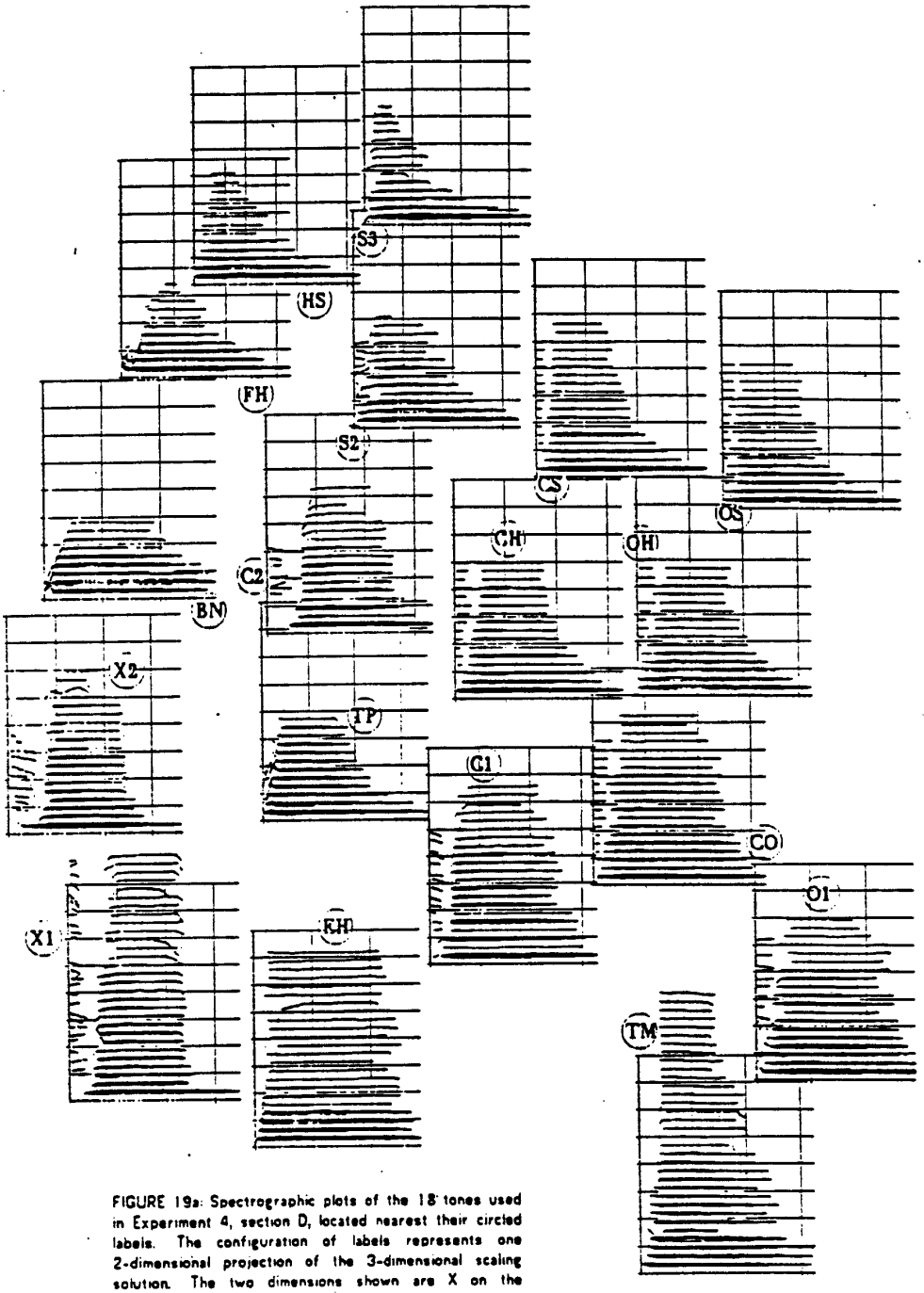



FIGURE 19a: Spectrographic plots of the 18 tones used in Experiment 4, section D, located nearest their circled labels. The configuration of labels represents one 2-dimensional projection of the 3-dimensional scaling solution. The two dimensions shown are X on the horizontal and Y on the vertical (see text for interpretation of the axes).

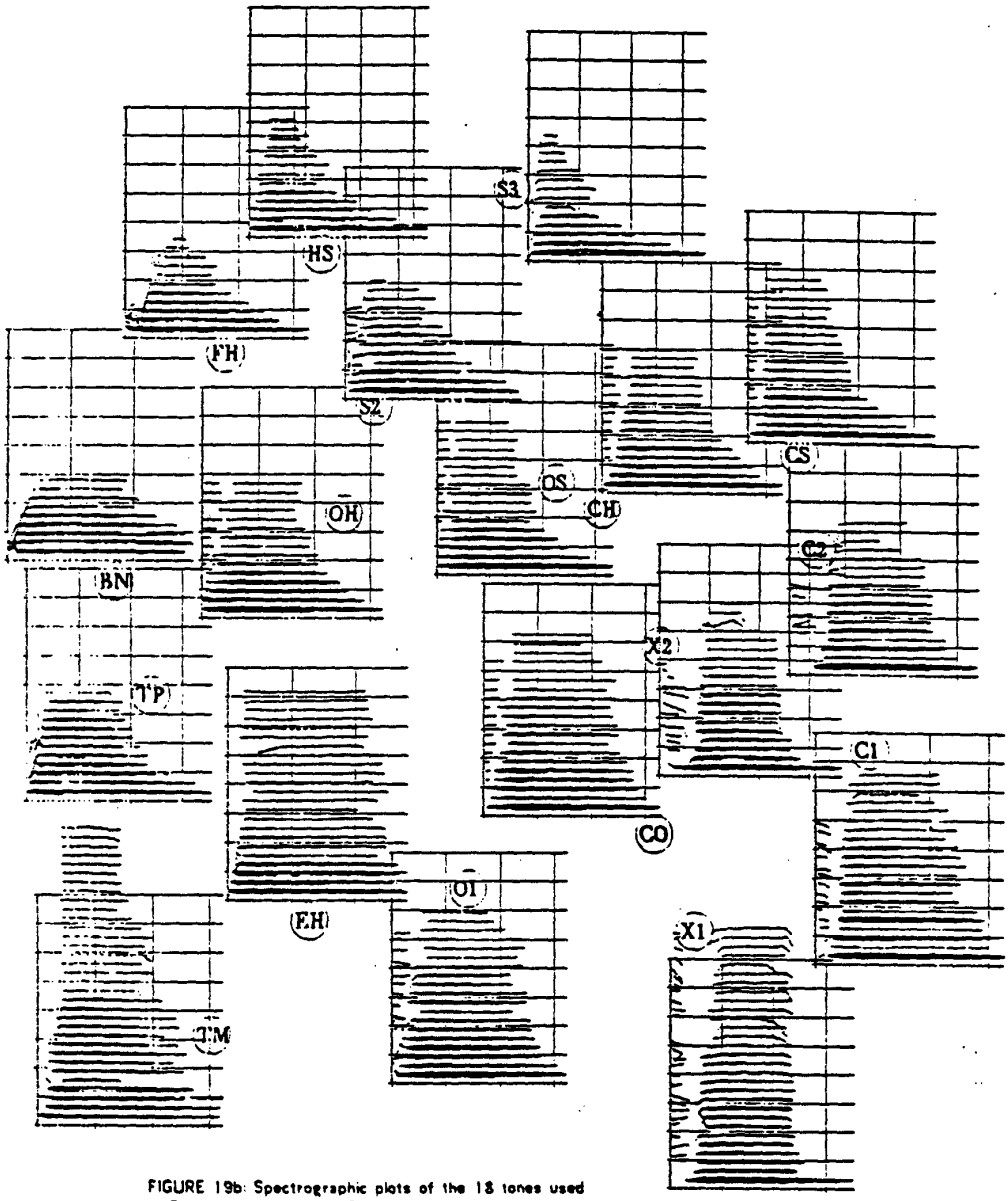



FIGURE 19b: Spectrographic plots of the 18 tones used in Experiment 4, section D, located nearest their circled labels. The configuration of labels represents one 2-dimensional projection of the 3-dimensional scaling solution. The two dimensions shown are Z on the horizontal and Y on the vertical (see text for interpretation of the axes).

A minimum of *three* dimensions was required to represent the data structure obtained in this study. Two of these dimensions were interpreted in terms of the physical properties of the tones, namely the *spectral energy distribution* and the presence of *low amplitude, high-frequency (inharmonic) energy* in the initial attack segment. These same two dimensions were uncovered in the previous work described in Section C, where a total set of 16 naturalistic tones were used.

The spatial locations of the 6 *midpoint* tones is approximately midway between their respective *end-point* tones in the three dimensional solution. This would suggest that the interpolation algorithm did indeed generate a perceptual continuum between two given notes, in that the midpoints fall midway in similarity between the two end-points. If the algorithm had produced *categorically perceived* transitions, then midpoints might have the tendency to appear extremely close to one of the two extremes, or even map onto it if categorical perception were very strong. This would especially be true if the mathematical midpoints were not identical to the categorical boundary cross-over points, as was the case in 3 of these 6 interpolations. It is interesting to note that there was a small shift in spatial locations of these 3 mathematical midpoints toward the end-point in whose category they were placed in the previous experiments in this section. However, small shifts also occur in the other cases.

The exact locations of these midpoints seem to suffer a distortion from a Euclidean interpolation, wherein they are all pushed out from the center of the space. This perhaps indicates a perceptual distinction made between these midpoints and the other naturalistic tones. This is in agreement with the interpretation suggested for third dimension, that it may represent a division of the tones between the *midpoints* on one hand, and the remaining naturalistic tones on the other (the *end-point* perhaps serving as a link between their midpoints and the other naturalistic tones). Perturbations in this timbre space with respect to the configuration derived for the 16 naturalistic tones in Section C might be explained on this basis.

The latter dimension, where tones involved in interpolations were distinguished from those tones that were not, could not be interpreted in terms of obvious physical features of the tones. It may have been the result of a higher-level distinction made between of the *interpolation-derived* midpoint tones and the other *naturalistic* tones. In that the end-points had strong similarity relationships with their respective midpoints, as well as with the other naturalistic tones, the end-points may have been the common link in the similarity space. A simpler hypothesis might consider the experiences that the listeners had in previous experiments, some of which utilized only naturalistic tones and others of which had the set of midpoints and end-points for stimuli. If the third dimension derived in this study was indeed a product of a higher-level distinction made between the tones, then it may be that the space represents a combined mapping of low-level perceptual distinctions and a higher-level, more *cognitively based* differentiation.

# IV. CONCLUSIONS AND IMPLICATIONS FOR FUTURE RESEARCH

## A. Review of Current Research

*Perceptual Measurement of Analysis-Synthesis Technique*

One phase of the research described in Section A had implications for the perceptual success of the analysis-based additive synthesis strategy. This success was defined to be the virtual indiscriminability of the synthesized tones from those original tones on which they were based. The *ultimate* test of this technique was not accomplished by the particular experiment discussed, unfortunately, in that methodological problems resulted in additional cues for discrimination between the original and re-synthesized tones.

However, in spite of these additional cues, consisting of noise added to the signal in the processes of recording and digitization, the resulting comparison between a synthetic and original tone was *exceedingly* close, according to listeners' verbal reports. There was no indication that the extremely small differences perceived in this comparison resulted from essential changes to the *timbre*, and, in fact, most differences were readily heard as minute changes in the articulation of an (*assumed-to-be*) actual instrument which was being played by the same player in the same basic style. Most of the professional performing musicians who were listeners in the experiment (who were not familiar with the computer techniques involved) commented that the very *fine* degree of differentiations made by the hypothesized performer playing the different notes were far in excess of the types of distinctions they thought possible in instrumental performance techniques.

This is clearly the best measurement of the perceptual adequacy of the analysis-synthesis strategy. It has been verified by countless public demonstrations comparing the original and re-synthesized tones. It is important to note that the isolated tonal condition could likely be the most *stringent* test of the technique. Musical listening situations, where there are notes in melodic contexts, possibly even with simultaneously sounding lines, will certainly reduce the total attention given to timbral details, and could thereby reduce the discriminability of various models for synthesis.

The most rigorous discrimination tests will have to await the right configuration of equipment, however. This is clearly of *methodological* importance. The proper environment for testing would have the actual notes performed in an acoustical module and recorded directly into a computer via a high-quality analogue-to-digital converter. The intermediate tape recording should be avoided at all costs. Then, playback would be into the same sort of module, directly from the computer.

*Simplifications of the Physical Properties of Analyzed Tones*

In this phase of the research discussed in Section A, discrimination measurements were taken with respect to different modifications imposed on the analyzed data structure of a tone before re-synthesis. These modifications were uniformly in the direction of *simplifying* the very complex time-variant amplitude and frequency functions obtained from the analysis of a tone. The approach taken was to fit the complex functions with a small number of *line segments*, thus eliminating the mini-perturbations and fine detail of the analyzed curves.

Of particular concern was the *extent* to which the data reduction process could be carried and still remain successful in terms of *indiscriminability*. Line segment fitting could ignore the low-amplitude precedent activity in the attacks of several harmonics. A more drastic modification of the frequency functions was to assume that they remained *constant*, rather than time-variant (which, by the way, remains an assumption built into most work in the areas of psychoacoustical complex-stimulus production and musical timbre synthesis).

The results of the perceptual study, in which all combinations of tonal conditions (*original*, *complex re-synthesis* and *simplified re-syntheses*) were paired for discrimination and subjective distance estimation, revealed an orderly structure in these judgments. This structure reflected both the *extent* of simplification from the original tone, and the *nature* of the modification. The data reduction which ignored the total *attack* configuration of the instrument note by not including the very low-amplitude, precedent activity in the upper harmonics, was more discriminable and perceived to be more distant uniformly across all other conditions. The simplification of the frequency functions, whereby they were made to be constant rather than time-variant, had a very discriminable effect in several cases.

The most successful modification of the complex time-variant functions measured was the *line-segment approximation* in both amplitude and frequency domains. This provides a very powerful tool for perceptual research on timbre, in that the *drastic* data reduction has eliminated non-essential physical characteristics of the tone synthesized from this data base. The data reduction for the short duration tones used in these experiments was on the order of 100 to 1; where it took about 1000 numbers to represent the complex time-variant behavior of a harmonic, it took only about 10 numbers after data reduction.

The pronounced success of this modification strongly suggests that the micro-fluctuations found in the analyzed functions have little perceptual significance with most tonal cases. It became apparent, however, that in some instances these fluctuations had to be treated carefully. A strong blip at the beginning of a note usually had to be included in the simplified model. Narrowing down the *factors* corresponding to successful versus unsuccessful modification may be very important future research directed towards uncovering distinctive features of tone.

*Equalization of Stimuli for Pitch, Loudness and Duration*

From informal observations on the effects of the various dimensions of sound upon one another, it was expected that *pitch, loudness* and *duration* might influence similarity judgments for timbre. Although psychoacousticians have not directly investigated these influences *upon timbre*, there has been much research which shows the effects of timbral properties of tone upon the dimensions in question. We therefore decided to perform an *equalization* of stimuli before proceeding with perceptual scalings of timbre, and this constituted the experiments discussed in Section B.

The results of these equalization experiments showed that, given our present state of knowledge regarding the perception of complex, time-variant tone, it would be impossible to equalize stimuli for pitch, loudness and duration on an *a priori* basis. The best model for loudness was applied with tentative success in explaining our resulting perceptual matches. However, more testing must be done to determine whether indeed the model as is (it was built for *steady-state* tones) actually would have assisted in generating a set of matched stimuli. Studies in the psychoacoustic literature would suggest that a complete model for loudness must necessarily consider time-variant phenomena [Gjaevenes and Rimstad, 1972].

In considering the other two dimensions that were equalized, *pitch* and *perceived duration*, there is no suggestion in the literature that time-variant, complex tones have been studied. At present, there is no model for pitch perception which would advise the investigator how to equalize time-variant stimuli. Furthermore, hardly any research has been done on perceptual duration, yet alone for naturalistic tones.

The lack of this knowledge comes as little surprise, however. The multidimensional interactions which must occur in such *complex* perceptual situations as the case in question, would seem to prohibit a detailed study until adequate *tools* are developed for research. The present set of studies may serve as a pointer to the opening potentials for this type of research in the future. The tools are being developed, largely because of the interest of researchers in the field of computer music and timbre perception. The multi-parameter equalization of complex naturalistic tones performed in this experiment would not have been possible without the digital computer and its related synthesis techniques.

The set of studies discussed above strongly suggests the value of an equalized set of stimuli. Experiments certainly should be done to verify or disprove this sentiment. However, until successful models have been developed for the automatic equalization of timbral stimuli, it would seem that the researcher must face the problem of an empirical equalization before employing a set of stimuli in experiments on timbre.

*Scaling the Multidimensional Attributes of Timbre*

Among the established techniques of perceptual scaling, those methods which fall in the category of *multidimensional scaling techniques* present the most powerful, general approach for an exploration of the complex perceptual attributes of timbre. The scaling accomplished with these techniques is based solely upon perceptual measurements, rather than upon known or *assumed* properties of the stimuli. In that the perceptual judgment is usually a generalized rating, like *similarity*, there is also an avoidance of specific assumptions on the nature of the verbal scales which should be employed for a particular type of stimulus.

Two studies discussed above utilized this type of perceptual scaling and analysis. One, described in Section C, had a set of 16 naturalistic instrumental notes, while the other, described as Experiment 4, Section D, mixed 12 naturalistic tones (a subset of the 16) with 6 tones derived artificially from certain of the naturalistic tones. The method of derivation, an interpolation algorithm developed to create *transitions* in timbre between two naturalistic end-points, will be reviewed in the next part of this discussion. Generalized similarity ratings were used to obtain the subjective distances between all pairs of tones in each study. These psychological distances were then treated with multidimensional scaling techniques and presented in an optimized low-dimensional Euclidean representation of their data structure.

The results of these two studies shared several common properties. Of specific interest was a comparison of these results to three previous studies which have been done in this field [a pilot study in our laboratory, 1973; Wedin and Goude, 1972 (scaled by Wessel, 1974); Wessel, 1974]. One striking difference was the inability to obtain a reasonable *two-dimensional* solution with the present two studies, since the previous three scalings were interpreted in two dimensions. This may be due to several differences in method, but three strong candidates are: 1) the equalization performed on the present stimuli; 2) the differences in duration of the present stimuli and those previously studied, the present tones being much shorter; and 3) the increased number of stimulus points in the space, from just 9 tones in most of the earlier work to 16 or more in this research.

A positive correspondence exists between all scalings of timbre, in the common interpretation of one of the obtained spatial dimensions in terms of the physical property of *spectral energy distribution* (see Section C). With the present set of two studies, another dimensional interpretation in common concerned the physical existence of *low-amplitude, high-frequency precedent energy* (possibly its *inharmonicity* being important) in the initial onset segment of the tones.

The third dimension obtained in the first study, discussed in Section C, was interpretable either in terms of a simple physical property (that of *synchronicity* in the attack-decay behavior of upper harmonics) or with respect to stimulus groupings according to instrumental *families* (the strings, brass and woodwinds). Neither interpretation was unequivocal. This may be of particular importance because the three previous scalings of timbre referred to above, while

interpreted as having one dimension that was based upon an obvious physical feature of tone (e.g. the spectral energy distribution), had another dimension which was *uninterpretable* in terms of an obvious *physical* dimension, but was considered to have to do with those properties of tone involved in the higher-order distinctions made between musical instrument *families*.

(It is interesting to note that in two additional non-auditory experiments where similarities of instruments were judged on the sole basis of the instrument *names* rather than their sounds [Wedin and Goude, 1972 (as scaled by Wessel, 1974); Wessel, 1974] the resulting similarity data showed a much greater weight given to the *family* axis in the internal representations of instruments; hence it may be considered to be a more *cognitive* dimension of tone.)

In the second study, described in Section D, no obvious physical interpretation of the third dimension was readily suggested. Rather, a higher-level distinction seemed to be made between the tones on the basis of their involvement with interpolations. The interpretation of the third dimension in this study as having to do with a higher-level distinction made between interpolation-derived tones and naturalistic tones seems to parallel the existence of a family distinction interpreted for the studies above which utilized only naturalistic tones.

If the third dimension derived in this latter study was indeed a product of a higher-level distinction made between the tones, then it may be that the (multidimensionally-scaled) timbral similarity space represents a combined mapping of low-level perceptual distinctions and a higher-level, more *cognitively based* differentiation between stimuli. The scalings for sets of naturalistic tones suggest a *hybrid* space, where some dimensions are based on low-level perceptual distinctions made with respect to obvious physical properties of tone, while other dimensions can be explained only on the basis of a higher-level distinction, like musical families. In our second study, where interpolations between naturalistic tones were included, the family structure seemed to have been superseded by the distinction made between the interpolation-related tones and the others.

These studies taken as a whole would suggest a *timbre space* in which low-level properties are combined with higher-level family or other group distinctions in making the generalized similarity judgment. This would not be totally inexplicable, in that stimuli having components of *familiarity* might be expected to manifest such components in a generalized judgment like similarity. Perhaps a scaling of timbre which could be *totally* interpretable in terms of the physical properties of the stimuli might be a study in which any possible sort of cognitive-type perception of tone has been effectively eliminated. One such study may have been a scaling [Plomp, 1970] of single repeated periods taken from steady-state portions of tones (the resulting stimuli being virtually unrecognizable as musical instruments). Here, the solution derived was totally interpretable with respect to three physical properties, equivalent to the features in vowel perception.

Most probably, the best way to proceed with timbral multidimensional scaling will be to restrict

the stimuli to smaller subsets of instruments. Here, the possible cognitive-level attributes of tone encountered will be features of articulation, playing style, and so forth. It is possible that these contexts may serve to uncover still other psychophysical correspondences in timbre perception, perhaps even to the higher-level attributes of articulation and style.

The *anomalies* found with respect to family grouping in the first study of this series are worth noting. Many listeners reported strong initial misidentifications of the anomalous cases, an overblown flute indeed sounding like a *string* instrument, a high-register bassoon really mistaken for a *brass* instrument. Carefully worked out parallel studies in the confusions in identification of instrument tones should be of great use in interpreting scaling solutions. Unlike the identification/learning study used in this research, which was designed for multiple purposes, we suggest initial experiments without feedback.

*Interpolation between Naturalistic Timbres*

A set of studies assessed the properties of a general interpolation algorithm that was designed to produce a set of tones which formed a transition between two familiar extremes; the studies are detailed in Section D. The interest in this algorithm was the light which could be shed on the *continuous* versus *categorical* nature of the perceived transition between known end-point tones. It is important to realize initially that the algorithm used was but one of several *possible* ways to generate tones on the basis of two given notes, since interpolation was through a multidimensional physical space. Hence the studies assessed a *particular* algorithm that was employed.

Experiment 1 examined the identification of interpolated notes in sequential presentations. There were strong *hysteresis* effects. The point of cross-over in identification, defined by the number of the tone which marked the initial appearance of the second familiar timbre, was heavily dependent upon the *direction* in which the sequence was presented. The identification of the first-played tone tended to hang on beyond the point at which it was heard to first enter when the sequence was played in the *opposite* direction. This classical hysteresis effect suggested the lack of *sharply* defined categorical boundaries. Independent judgments indicated that these timbral transitions were more *gradual* than *sudden*, and this supported the notion that the interpolation algorithm generated perceptual continua.

Experiment 2, in which isolated interpolated tones were identified and rated in a multiple choice task, also gave no indication that the categorical transition was sharp. It rather revealed gradually changing categorical boundaries, which correlated with gradually decreasing ratings of the goodness of the tones in fitting those categories used for identification. (Detailed in many listeners' introspective reports was the perception of a *combination* of the two tones at the center of the interpolation, either both sounding simultaneously or one tone having partial properties of both. A pursuit of this fascinating perceptual phenomenon, whereby a single set of harmonics has the multiple cues for two tones, may uncover important facets of timbre

perception regarding the nature of cues for identification. Note that implicit in this ambiguous tone is a more fundamental but less rigorously defined sort of categorical perception, which must *resolve* an artificial tone in terms of *known* sources.)

Experiment 3 tested for the existence of peaks in discrimination at categorical boundaries, a key factor in the *classical* definition of categorical perception [Liberman, 1967]. There was no clear trend of peaking in the discrimination and distance estimations of listeners. The studies suffered ceiling effects, where discrimination was very high for all pairs tested in four of the six interpolations, and the opposite floor effects existed in the other two. In the latter two, there was some indication that discrimination was lowest within an expansive categorical boundary, which does relate to the classical effect. No clear results were obtained, however, due to the great skewing of the boundaries in the latter two cases. The distance functions were uniformly flat, indicating a continuously perceived transition. There was no direct indication in the three studies that the interpolation was perceived in a categorical manner, but it was rather suggested that the interpolation was perceived as a continuum.

The multidimensional scaling study discussed above, Experiment 4 from Section D, also tended to indicate a continuous-type perception of this interpolation algorithm. The placements of the tones within the spatial solution, especially of the interpolation-derived *midpoints* with respect to their *end-points*, was in agreement with the notion that the interpolation generated a continuum. The relative distances of the 4 end-points and respective midpoints was completely in accordance with the relative discrimination functions for the different interpolations. End-points which were *close* in the space generated *poorly-discriminated* interpolations, while end-points that were more *separated* in the space generated *well-discriminated* interpolations. The relative distances of the naturalistic tones in this scaling solution were similar to those obtained in the first scaling of the set of 16 naturalistic tones, although perturbations of the overall structure were found.

The interpretation of the space generated in the second scaling study included a dimension which was thought to represent the distinctions made by listeners between the interpolation-derived tones and the naturalistic tones, with the end-points of the interpolations linking the similarity space together. Informal verbal reports by some listeners indicated that the tones which occurred in the center of several of the interpolations were different from the naturalistic end-points (perhaps sounding like a pair of notes or a hybrid tone). This supports the above interpretation of the space. Careful studies should determine the nature of those interpolations which generate hybrid transitions and those which do not, using this algorithm. This would be useful in locating timbral cues, designing new algorithms for interpolation and possibly further testing categorical perception for timbre.

## B. Summary of Contributions and Speculations

*Methodological Contributions*

A primary methodological contribution to timbre research made in the course of this work has been the extension and elaboration of the analysis-based additive synthesis technique. A wide range of timbres were encompassed by this technique, and the potential power, or even necessity, of the strategy of synthesis for experiments in timbre perception was made known. Heightening the utility of this approach was the equalization and manipulation of stimuli that were made possible.

Adding still further along these lines to the tools of the investigator of timbre perception was the exploration of data reduction techniques. The complex results of analysis, certainly gives the researcher a full knowledge of the physical properties of stimuli synthesized on the basis of the analysis. However, it presents far too much physical information to allow for unambiguous psychophysical models that are based upon physical correlations to perceptual structures. Any empirically based data reduction of the complex analyzed functions used for synthesis will simultaneously simplify the job of the researcher trying to determine meaningful physical parameters and directly address the nature of those perceptually salient features in tones. There will be no change in perceptual structures based on simplified tones if the data reduction has not eliminated salient acoustical information. In our data reductions, using small numbers of line segments to approximate functions two orders of magnitude more complex, resulting simplified tones were barely discriminable from tones synthesized on the basis of the more complex functions.

Graphical tools were developed to facilitate an exploration of the physical properties of complex, time-variant tones. The perspective plot of Amplitude x Frequency x Time for a single tone reveals temporal relationships between the harmonics. A computer spectrograph was also developed to assist in an examination of events in the frequency domain, such as formants. Various on-line manipulations also facilitated the display of such information, such as being able to visually move about the three-dimensional perspective plot or being able to interactively specify the floor level intensity in the spectrograph.

Likewise, tools for the interactive exploration of three-dimensional data structures generated by multidimensional scaling techniques proved invaluable in this work. Interactive displays which allowed for a rotation of configurations, in combination with a stereoptic apparatus, revealed all three dimensional relationships within a configuration. Congruency-matching as an added feature to this display, whereby two different configurations were rotated to best fit and simultaneously displayed, added further information about the relationships between different scaling solutions. Finally, the ability to generate three-dimensional representations as two-dimensional hardcopy, as well as to place physical analyses of stimuli into two-dimensional projections of the space, greatly facilitated the data interpretation process.

Another contribution to specific future research on timbre, as well as to model building with respect to the pitch, loudness and duration of time-variant tones, was the equalization of a set of 16 musical instrument tones. Several timbre studies in the future may be based upon these tones. In addition, research which will attempt to extend models for the perception of the three other attributes of tone listed above will initially be based on these equalized tones. More fundamentally, tests may be performed exploring the applicability of the notion that equalization is necessary to examine timbre perception, and these tones may serve an integral part in that research. Various facets of the equalization procedure also may be examined in detail, including the success of equalizing the pairs of tones which were not directly compared in the present experiment. Other procedures may grow out of this research.

A final methodological contribution to timbre research was a specific algorithm for the interpolation between two given naturalistic tones which was not simply based upon a linear mixing of waveforms, but attempted to take time-variant processes into account. Refinements of this approach will serve to further investigate the nature of timbre spaces, the existence of categorical perception in timbre and the types of cues which lead to timbral identifications.

*Theoretical Contributions and Speculations*

Several facets of a psychophysical model for timbre were explored in this research. Successful data reductions narrowed the range of analyzable physical properties having perceptual significance. Particularly important were the patterns of temporal activity residing in the attack of a note, such as low-amplitude energy preceding the main build-up of harmonics, or blips in the initial onsets of certain brass and woodwind tones. Deviations from these patterns, which nonetheless could successfully be represented in grossly simplified forms, were very detectable. The importance of temporal activity in the frequency domain was found in certain cases, especially with notes that have inharmonicity in their onsets, such as strings, or notes that have pronounced internal pitch shifts in onset and offset, as is the case with many woodwind instruments.

Variance with respect to the salience of these features was found amongst the 16 instrument notes measured. The lack of uniformity in the more drastic simplifications of tones suggests a binary-type distinctive feature model of timbre, especially as it relates to cues for identification. Methodological techniques outlined above, especially the data reduction and discrimination measurements coupled with multidimensional scaling of measured distances, should serve as one approach to enumerating these features. Outright identification measurements would be an important parallel approach.

The multiply-defined meaning of timbre, on the one hand speaking of the cues for the *identification* of a familiar instrumental source, i.e. that which makes a trumpet a trumpet, and on the other hand talking about the *qualities* of tone independent of its identification or lack of identification with respect to source, has found its way into this research effort. The particular

discrimination studies which examined data reduction techniques were exploring both aspects, and the multidimensional scaling of timbral similarities also seemed to uncover the two essences of timbre perception. The former studies revealed some reductions which would perturb only the type of articulation of an instrumental tone, while other reductions would seem to move a tone into a different timbral category. For instance, removing the precedent low-amplitude, high-frequency energy on some woodwinds would resemble the difference in articulation between a tongued and non-tongued attack of the same instrument. Removing certain blips in a brass tone had the effect of making it a much more ambiguous note. Relating to the constant frequencies approximation, in many cases tones were simply disturbed slightly in tonal quality, but not shifted in their apparent source, e.g. simply a brighter sound. In other cases, as with a particular saxophone note that had noticeable pitch shift, the constant frequencies version tended to sound much more like an English horn.

These examples share the common aspect of relating known features of tones to their sources. Certain sources may naturally display particular modifications, such as brightening of tone or tonguing of attack, which will be taken into account when particular acoustical modifications are applied to their output. The listener will still identify the tonal source. Other sources may have such characteristic qualities, such as the attack blips of brass, inharmonicity of strings or pitch shifts in certain other tones, that modifications which change these tones may incidentally move them out of particular categories either into a more ambiguous realm or actually into another category.

The multidimensional scalings of timbral similarities also revealed the multiple nature of timbre. Two dimensions were uncovered in this work which related to obvious, common properties which could be compared between a set of heterogenous tones, coming from different instrumental families. They may be considered as having to do with general tonal qualities which supersede particular identification functions of perception. Yet another dimension has been uncovered in the five similarity studies done with timbre. This dimension relates to instrumental family in four cases, and in the fifth, to a distinction between tones having to do with an interpolation algorithm and the remaining naturalistic tones. The commonality in all five cases is a higher-level perceptual process being mapped onto the similarity judgment, a process which could relate to identification-level behavior.

Finally, in considering the theoretical implications of the notion here developed of timbral interpolations, yet another way of looking at timbre in terms of a spatial model is suggested. A given scaling of timbre, in generating a representational space, produces a visible structure relating the similarities of one particular set of stimuli. If the space, that is, the structure inherent in the similarity data, can be linked *totally* to the physical qualities of the stimuli, and indeed exhausts the set of salient physical attributes of the stimuli (as has occasionally been possible outside of timbre research) then a psychophysical map of timbre has been obtained. If the determined physical qualities were actual acoustical *continua*, then this map would have the *ideal* characteristics that a point could be constructed within the context of the physical

dimensions of tone, and its location would be *predictable*. Likewise, one should be able to *generate* a known tone on the basis of its coordinates in the space, since the dimensions of the space would be specifiable totally in terms of acoustical continua. An interpolation in this sort of idealized psychophysical timbre space would then be a matter of moving along axes at perceptually equalized increments, and there would be an *a priori* interpolation technique.

We have ample intuition that the sort of idealized timbre space described above may not in fact be obtainable. The research which has sought out this psychophysical mapping of timbre has found no such scheme. Rather, composite dimensions are uncovered, like the host of cues which related to spectral energy distribution, or even more combinatorial, the musical instrument family axis. If we assume for the moment that the problem does not lie so much in the lack of suitable analytic techniques as it does in the inherently complex and multi-purpose mechanisms involved in timbre perception (or facial recognition-perception, for that matter), then we are forced to redefine the nature of a timbre space. The present state of research indicates a timbre space that indeed reflects such *cognitive* facets as familiarity and recognition. There does seem to be more than one line through the timbre space, and it seems likely that no interpolation based upon the physical properties of two familiar points in such a space will be capable of generating a third actual point that happened to map midway between them in similarity (with possible trivial exceptions that do not involve multidimensional interpolations). Given the non-absolute stature of the prospect of timbral transitions, it may not be appropriate to come to any final conclusion regarding categorical perception and timbre. There may be continuous interpolations through a categorical space, depending upon the nature of the algorithm employed and the context of the judgment. The converse may also be true.

The model of interpolations derived from these studies was one in which certain types of physical attributes of tone seemed to present continuous perceptual transitions, while others seemed to be of a more discrete nature. The model of timbre perception suggested is a composite model where salient physical properties of tone may correspond to continuous perceptual dimensions or may relate to more discrete, binary-type distinctive features. Global spectral effects, for example, would seem to be more continuous, as has also been found with vowels in speech research. Details of the attacks of notes, which are important for the identification of source, seem to be of a more discrete nature. Certain of the interpolations seemed to be dominated by a strong discrete feature of one or both end-points. For instance, the French horn had a salient blip in its attack function. An incremental transition seemed to result in a discrete function, whereby either a blip was heard or not. If heard, it could dominate an otherwise neutral timbre. In the presence of a second end-point having a strong feature also detected, the listener might hear two tones simultaneously sounding, or of an odd sort of hybrid tone. Such impressions were frequently reported.

Observations of numerous interpolations using the present algorithm revealed that the overlap of distinct features of the more binary type could be controlled. Hence one might be able to predict the quality and identification of interpolated tones from the functions used to control

the contribution of various features. For instance, with a pair of naturalistic end-points having quite different physical features in their attacks, the strengths of these features may be weighted in the process of interpolation. With perceptually preserved, overlapping sets of features, central tones may appear to be multiple, sounding either like a hybrid tone or a pair of the end-point tones playing together. This fascinating effect is created, remember, on the basis of exactly *one* set of harmonics that preserves multiple features. Musical listening, unlike speech perception, allows for simultaneity of voices, so the lack of categorical-type perception as found in speech (where a tone is resolved to a unique and singular source) may be a product of what is possible in the *musical mode of perception.*

Integrating the above, it would appear that timbre perception involves several different levels of processing, activated in different contexts for different tasks. A primary adaptive reason for such perceptual processing relates to identification of source characteristics. Distinctive features, possibly of a binary-type, may be the goal of processing on this level. This task may call forth various other higher-level, stored information concerning relationships between perceived entities. The sort of perceptual behavior suggested may be of the categorical type. More low-level comparisons of immediate stimulation on the basis of shared acoustical dimensions is possible, and constitutes another major domain of timbre perception. Gross features of tone shared by all relevant stimuli would be the basis of qualitative comparisons that lie outside of the domain of specific identification. This sort of comparison might also be the ultimate basis for building timbral identities. Continuous or discrete features may be involved, such as spectral energy distribution, more often a continuum than not, or the presence of low-amplitude precedent energy in the attack, which may be more binary for perception (further research is required). It seems appropriate to model dynamic timbre space, subject to the transformations of context, as have been seen even in the two experiments included in this research. Different subspaces may bring out the influences of different features, as perceptual processes focus in on more detailed comparisons of some attributes than others when appropriate to the stimulus set and the perceptual task.

## C. Suggestions for Future Research

### *Data Reduction for the Location of Distinctive Features*

A very important direction to continue in future research involves a systematic approach to the simplification of physical parameters of tones for re-synthesis. This provides a powerful means for uncovering distinctive features for timbral identification and salient physical properties for tonal quality.

The perceptual measurements employed to date in conjunction with the data reduction scheme have included discrimination and subjective distance estimations. Future research may benefit by looking at the effects of tonal modifications on identification. This would point more directly at the types of cues that are active in determining the source of a tone, and could systematically

lead to a distinctive features model for timbre. Note that the goal of this type of extension of the data reduction scheme would be to specify such cues for recognition and/or tonal quality rather than the more limited goal pursued in this research, to simplify the set of physical properties in order to assist in psychophysical modeling.

New approaches to data simplification should be initiated in future research. One sort of reduction which seems very promising is a simplified modeling in the spectral domain. The effects of critical bandwidths in hearing would strongly suggest that higher harmonics may not have to be treated with the sort of individual accuracy given to date. Significant data reductions may be possible, basing syntheses upon simplified models of spectral evolution which approximates the actual physics of tones, and this would be subject to perceptual experimentation.

It is clear that any type of successful data reduction will serve not only to simplify the physical properties of tone but will simultaneously reduce the complexities involved in timbre research on all fronts. The significance of this line of research cannot be overstated.

*The Use of Different Analysis-Synthesis Schemes and Forms of Data Display*

Closely related to the search for new forms of data reduction should be a parallel effort to explore the usefulness of new types of analysis-synthesis techniques and related graphical displays of physical information. One scheme which may be of future use is that of subtractive synthesis based upon an analysis for the tonal properties of spectral evolution. This approach has been used with great success in the field of speech analysis and synthesis, and could lead to new insights if successful with musical timbre. Displays of information would be oriented towards the time-variant behavior of spectral energy distribution, and a perspective plot of time-slices through formant activities may be most informative. In synthesis, data reduction would occur in that small numbers of time-variant filters may be adequate to model the perceptually salient aspects of tone.

This approach may broaden the domain of timbres which may be studied. Recall that the analysis-based additive synthesis technique has only been applicable to tones with *harmonic* overtone structures, that is, to tones whose partials are at approximately integer multiples in the frequency domain. An important step for the additive synthesis approach will be to seek out new forms of analysis that are capable of dealing with strongly inharmonic spectra. Current plans are being formulated to extend the use of digital filters for the analysis of inharmonic spectra. This will increase the domain of timbres capable of exploration to include many other important subsets of musical sounds.

*Scaling Different Sets and Subsets of Timbres*

It will be necessary that perceptual research on timbre look at other timbral domains, in addition to the harmonic-based instrument types presently being investigated by most researchers. Such sounds as percussive tones, which are composed either of strongly inharmonic partials or even of bandwidths of noise, should be investigated to increase the generality of models for timbre perception. Similar broadening of research suggests the scaling of stimuli having different durations and pitches than those which have been examined in this research.

In addition, research could also benefit from studying different subsets of tones within a particular timbral realm, such as harmonic overtone structures. Different features are expected to emerge by studying different contexts for perceptual judgments on timbre. For example, new dimensions for differentiation would be expected to come into play if a subset of orchestral instruments coming from a single family was scaled. Features relating to articulation and style of playing may come to light using various sets of stimuli.

Furthermore, the effects of timbral context on the structure of perceptual judgments may be examined by the use of overlapping sets of stimuli. The two studies performed in this research revealed an interpretable structural transformation given different contexts for the overlapping subset of stimuli. In this regard, it may be informative to look at subsets built on the stimuli already scaled in this work, for example, a set of all brass and/or woodwind instruments, or a set of clarinets and/or saxophones. Telescoping through timbral domains may simultaneously reveal telescopic context effects.

Finally, more along the lines of hypothesis testing, sets of stimuli should be scaled where there is some intent to predict the resulting structure. An example which is suggested by the first scaling study in this research would be to scale the same set of 16 tones, using the versions in which the precedent low-amplitude energy in the attack segments has been deleted. This would directly relate to one of the dimensions interpreted in the spaces thus far obtained. Another scaling, suggested by the second study, would include interpolation-based midpoints but not their end-points. This may give insight as to the hypothesized role played by the end-points as serving to link differentiated elements.

*Analysis and Synthesis of Melodic Contexts*

An important step which must be taken in methodological development is the analysis and synthesis of connected notes. This necessarily will precede carefully controlled work on the perception of timbre in musical contexts. Recall that the analysis techniques employed to date for timbre have been inherently designed for and applied to isolated notes. There exists no analytic scheme which is directly applicable to musical phrases. The following approach has been suggested for dealing with this problem [Moorer, 1974]:

Since there are no analytic techniques for directly analysing continuous waveforms over articulation boundaries as exist for separated, stationary notes, we must approach the problem in a somewhat circuitous manner.

The heterodyne filter is capable of tracking strong transients if the frequency is held nearly constant, so that many different instruments and articulations may be analysed just by having the player hold the pitch constant while articulating. To extend this to articulation at different frequencies, we must know the following things about the behavior at the transition:

1) Is there a discontinuity at the articulation point? If there is no discontinuity, then the frequency must change smoothly between notes. We may use the linear predictor to determine the spectral envelope during the transition period and a pitch detector to determine the period at each point in time. This does not give us the inharmonicity of the signal, if there is any. We will have to generalize on the basis of what we know about the tones when the pitch is held constant. If the instrument is slightly inharmonic, we must preserve that inharmonicity while the frequency is changing.

If there is a discontinuity at the articulation point, we may consider the complex as two separate tones and perform the analysis on each tone individually. This requires that we determine the position of the discontinuity and separate the notes by hand. We may discover the discontinuity and locate its position by examining the derivative of the sound waveform. We may obtain the derivative of the sound waveform by passing it through a wide-band digital differentiator.

2) Does the amplitude go to zero, or nearly zero at the articulation point? If it does, then we may separate the notes into isolated tones and perform the analysis separately. If it does not go to zero, and does not have a discontinuity, then we must analyse for spectral envelope and fundamental frequency, as described above.

Although it may turn out that there are some kinds of articulation that we cannot synthesize because of inadequate analysis techniques, there will certainly be a great variety of articulation types that are amenable to analysis. We expect that most tongued wind instruments will accept analysis easily. Strings when the direction of bowing is reversed, or when the bow is lifted from the string should be analyzable. The only case that will not be directly analyzable will be *legato* tones, where the pitch is changed without articulation.

It remains possible that many of the properties necessary for the *simulation* of connected passages will be amenable to simplification. This will greatly reduce the process of stimulus specification in the physical domain, and lend greater control to the synthesis method. Furthermore, it may turn out that empirically-based studies on the perceptual simulation of connected phrases will be the major approach to this whole issue.

*Timbre Perception in Musical Contexts*

Just as it was necessary to move from steady-state tones to time-variant musical tones in timbre research, so it seems clear that investigations must now move toward timbre perception in musical contexts. A major concern in timbre perception voiced above, yet not satisfied in this research, was that information derived from studies of the perception of isolated tones may not be representative of more typical musical perception.

The great difficulty met by many musically-trained listeners in attempting to identify the instruments which produced isolated tones testifies to the rather limited nature of current research. It is certain that other dimensions of timbre perception come into play when listening to a contextual passage of music. Not only do specific idiomatic traits of instruments occur, but more global data bases for perception are presented. One such data base which is opened to the listener is the pattern of resonances traced as the instrument is played through a register. This sort of information must be significant in normal listening situations.

Another type of research which may greatly benefit from perceptual measurements in musical contexts is the data reduction work. Here, the effects of musical context on attention to timbral details must be systematically explored. Tools should be developed in this regard for the simulation of articulations and patterns of phrasing. The ability to detect differences in musical contexts will be a more relevant sort of measure than discrimination studies on isolated tones.

*The Further Development of Interpolation Algorithms*

The area of timbral interpolation is quite open to development. The single algorithm employed in this research may suggest related approaches. The model for interpolation derived from using this method was one in which certain sets of features may be continuously varied while others seem to be of a more discrete nature.

Control of the overlappings of properties of the end-point tones appears to be possible. In effect, this makes the perceptual nature of the interpolated tones predictable and specifiable. Control of overlap was found, for example, in the use of logarithmic versus linear interpolations for the harmonic amplitude functions. Linear interpolations were used in this research, because overall loudness was best preserved for the transition. Logarithmic interpolations had the potential to control the rate at which particular features would decrement or increment. Low-level, precedent energy in one end-point only presents an good illustration of this. When interpolated linearly, this feature was detected often past the center of the transition. Logarithmic interpolations, depending upon the base used, could reduce the amplitude of this energy below threshold at points quite ahead of midway through the transition.

Central tones in the latter case tended to sound much more neutral with respect to identification. This suggests that, in the near future, interpolations may not only provide interesting means for assessing the continuity of a timbral dimension, but may also serve as a tool for uncovering distinctive features and manipulating them between extremes. More work should be done with the present type of algorithm, perhaps examining the use of logarithmic interpolations. Further explorations should be made in constructing other sorts of algorithms for timbral transitions. These may likely be based upon other models for synthesis.

# V. APPENDICES

## APPENDIX A: Analysis-Based Additive Synthesis

In additive synthesis we physically model a complex sound waveform as a sum of sinusoids with slowly time-varying amplitudes and phases. The process of synthesis involves specifying the amplitude and phase (equivalently, amplitude and frequency) for each component sinusoid as it varies with time throughout the duration of the tone. We will generally refer to this specification as being a time-varying function, amplitude or frequency, for a component sinusoid. These sinusoids are added together to produce the complex waveform. Equation (A 1) summarizes this formulation.

$$(A1) \quad F_\alpha = \sum_{n=1}^{M} A_n \sin(\omega_n \alpha h + \theta_n)$$

Notation:
- $\alpha$ is the sample number
- $h$ is the time between consecutive samples
- $F_\alpha$ is the sampled, digitized waveform at time $\alpha h$
- $A_n$ is the amplitude of the nth partial tone and is assumed to be slowly varying with time
- $\theta_n$ is the phase of the nth partial tone and is assumed to be slowly varying with time
- $\omega_n$ is the radian frequency of the nth partial tone

One can see from this model that if we can determine the functions $A_n$ and $\theta_n$ of a tone from a musical instrument, we can then synthesize an approximation to the waveform $F_\alpha$ from those functions by use of equation (A 1). To determine the functions $A_n$ and $\theta_n$ of a music instrument tone, we must assume the frequencies of the partial tones, $\omega_n$, are nearly harmonically related. By harmonically related, we mean that the tone has a fundamental frequency, $\omega$, and that the frequencies of all the partials of the tone are integer multiples of the fundamental frequency. That is, the frequency of the $n^{th}$ partial, $\omega_n$, is approximately $n\omega$.

Note that equation (A 1) could have been formulated with time-varying frequencies and constant phases. This formulation is equivalent and for all practical cases, the one can be derived from the other. We will speak interchangeably of the 'phases' of the harmonics and the 'frequencies' of the harmonics as a function of time. In the context of the analysis of tones, it is most natural to produce the phases of the harmonics as functions of time.

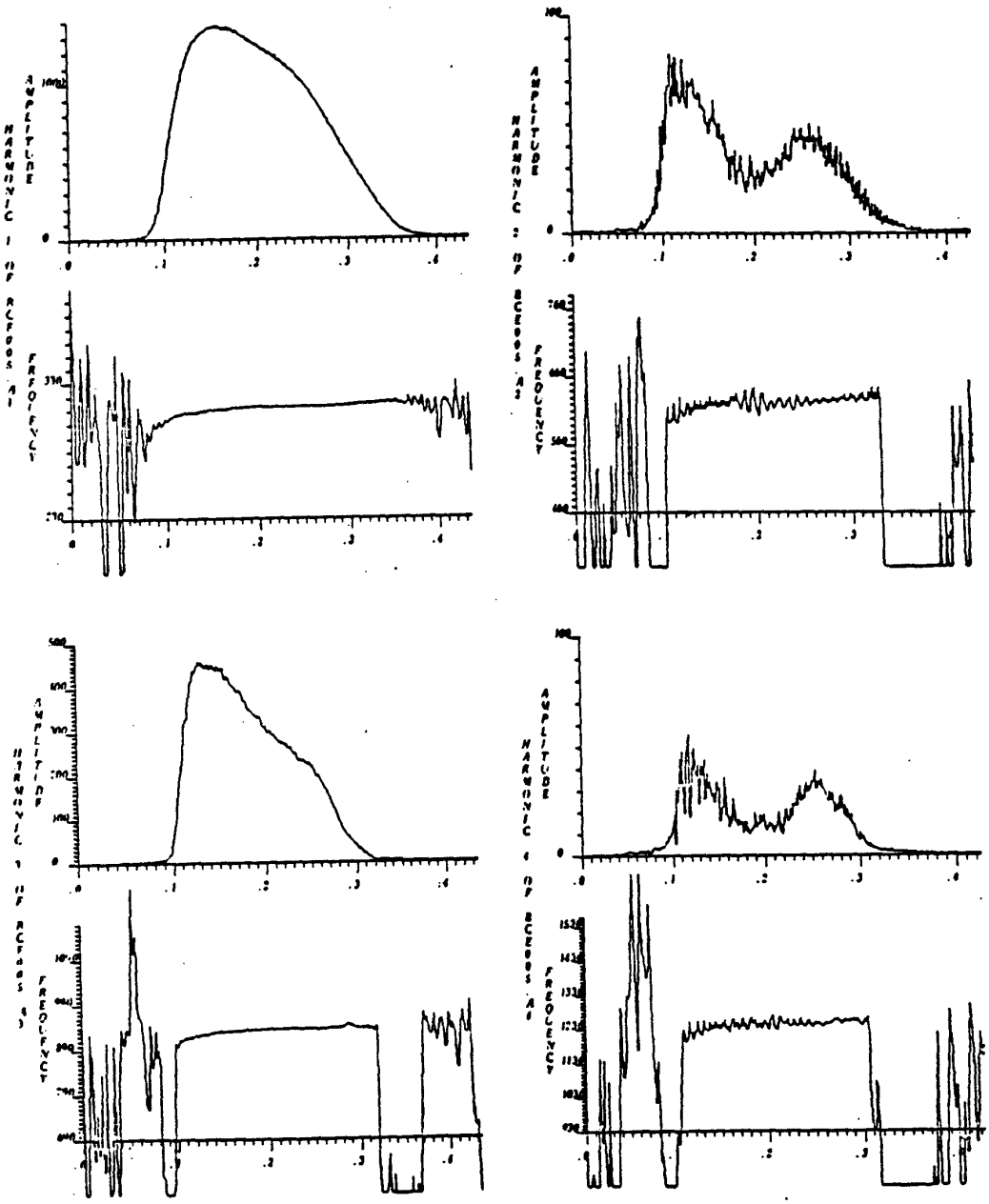


FIGURE A1(part 1): Time-variant Amplitude and Frequency functions for the individual harmonics of a tone (the X-axis = Time in seconds; the Y-axis = Amplitude or Frequency, as labelled). These functions resulted from an application of the heterodyne filter [Moorer, 1973] to the complex tone at the harmonic frequencies specified. The first four harmonics are shown above.

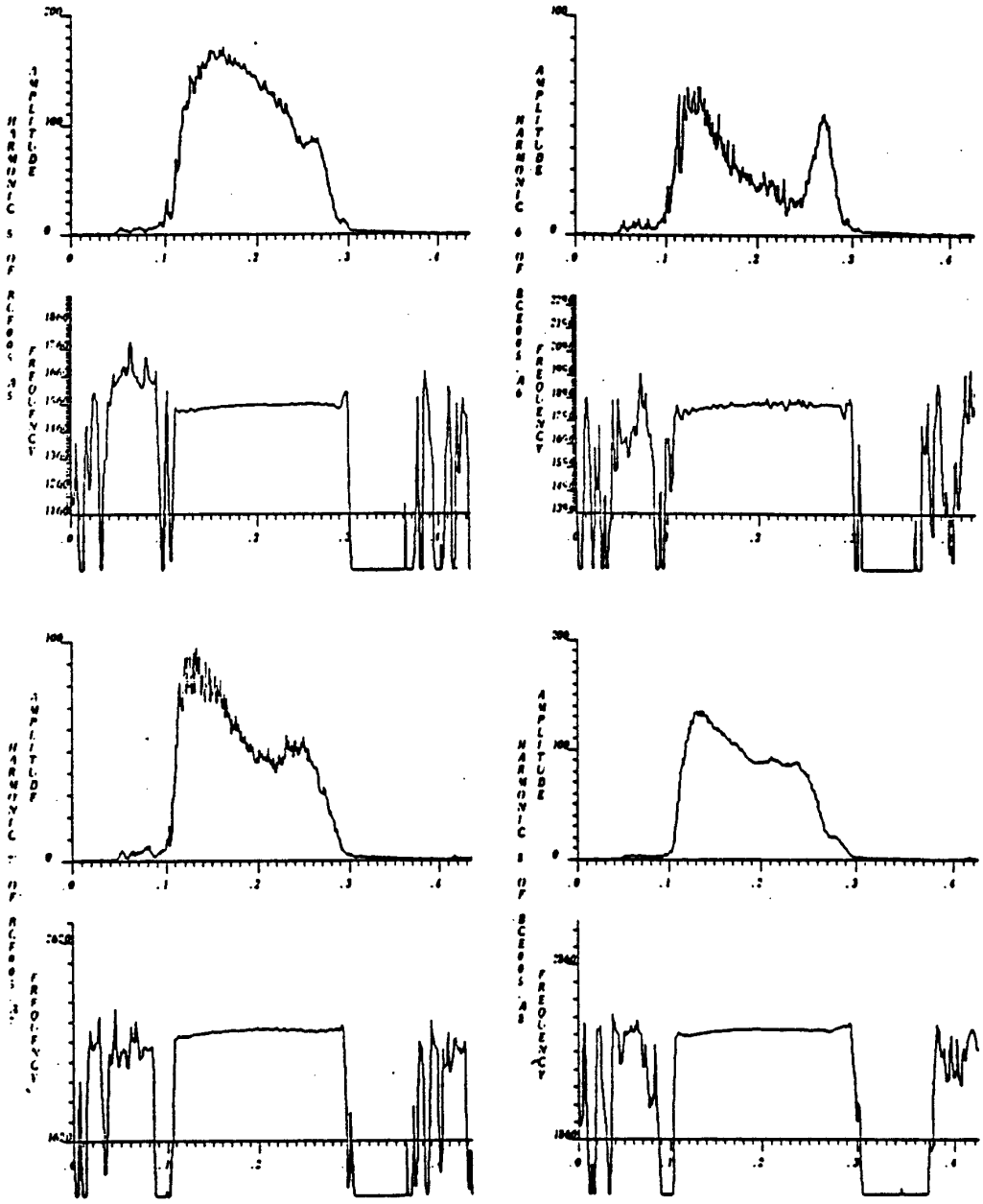


FIGURE A1(part 2). Time-variant Amplitude and Frequency functions for the individual harmonics of a tone (the X-axis = Time in seconds; the Y-axis = Amplitude or Frequency, as labelled). These functions resulted from an application of the heterodyne filter [Moorer, 1973] to the complex tone at the harmonic frequencies specified. Harmonics five through eight are shown above.

For intuitive purposes, however, it is more instructive to view displays of the frequencies of the harmonics as functions of time, and we will therefore usually refer to the frequency (rather than phase) functions of harmonics. In Figure A1 we present an example of a set of time-varying amplitude and frequency functions for eight component sinusoids of a tone. We would use these functions in additive synthesis to control the amplitudes and frequencies of eight components of a complex tone. In fact, they would constitute the first eight harmonics of the tone, their average frequencies approximately begin multiples of 310 Hz. Note that these functions were actually the result of the computer-analysis of a real tone, which was tape recorded and then digitized.

To be more explicit, we should point out again that $A_n$ and $\theta_n$ are functions of time. We will indicate this in the following text by appending the subscript $\alpha$ to each of these functions:

$$A_n = A_{n\alpha} \quad \text{and} \quad \theta_n = \theta_{n\alpha}$$

This brings out the time dependence more clearly.

It should be pointed out that this is not a Fourier series representation despite its outward similarity. The Fourier series approximates a periodic function with a sum of orthogonal sinusoids with fixed frequencies and phases and (possibly) exponential amplitudes. The model described in equation (A1) is not necessarily periodic, has time varying phase, and has time varying amplitudes that are not necessarily exponential. The individual components are certainly not orthogonal in the general case.

The following description of the analysis technique is a capsulization of Moorer [1973]. We have avoided use of continuous analysis and have defined all our functions in the discrete domain. This is done because the ultimate realization of these processes is on a digital computer. Inaccuracies can result from doing the mathematics in the continuous case and assuming it can be converted to the discrete domain merely by sampling. The conversion to the discrete domain must be done with some care. Since discrete mathematics and the science of digital signal processing have become so well developed, there is little reason to define our processes in the continuous domain, only to subsequently realize them in the discrete domain.

We have borrowed the notation of numerical analysis for functions sampled at equally-spaced intervals by placing the sample number as a subscript. This is done partly for notational convenience and partly because in the computer, sampled functions are represented by matrices. A method is sought for determining the functions $A_{n\alpha}$ and $\theta_{n\alpha}$ of a tone from a musical instrument, so that we can then synthesize an approximation to the waveform $F_\alpha$ from those functions by use of equation (A1). this equation uses sinusoids at time-varying phase angles rather than as sine and cosine quadrature components because it is more efficient to synthesize the waveform in this form.

Let us turn to the problem of determining the functions $A_{n\alpha}$ and $\theta_{n\alpha}$ of a music instrument

tone. To aid the analysis, we must assume the frequencies of the partial tones, $\omega_n$, are nearly harmonically related. That is, there is some frequency, $\omega$, such that $\omega_n$ is approximately $n\omega$. We shall call this frequency $\omega$ the fundamental frequency of the tone.

Basically, the method is as follows: First, compute the following two summations at each point in time $\alpha$.

$$(A2)\quad a_{n\alpha} = \sum_{i=\alpha}^{\alpha+N-1} F_i \sin(n\omega_0 ih+\phi_0)$$

$$(A3)\quad b_{n\alpha} = \sum_{i=\alpha}^{\alpha+N-1} F_i \cos(n\omega_0 ih+\phi_0)$$

The initial phase angle, $\phi_0$ is included for generality. It will be shown below that the method is independent of the initial phase angle. From these, we calculate two more sequences:

$$(A4)\quad A_{n\alpha} = (a_{n\alpha}^2+b_{n\alpha}^2)^{1/2}$$

$$(A5)\quad \theta_{n\alpha} = \operatorname{atan}(a_{n\alpha}/b_{n\alpha})$$

The summations are taken to be over one period of a sinusoid of frequency $\omega_0$, that is, $Nh\omega_0 = 2\pi$. This places somewhat of a restriction on the frequency of analysis, $\omega_0$, because in the discrete domain, the period, N, is restricted to integral values. This has not proved a problem in our experience. If the partial tones are nearly harmonically related, if the amplitude and phase functions of the tone vary slowly with time, and if $\omega_0$ is nearly equal to the fundamental frequency of the tone, then $A_{n\alpha}$ and $\theta_{n\alpha}$, as computed by equations (A2) through (A5), will indeed be approximations to the actual amplitudes and phases of the partials of the tone under analysis.

To review, we do the computations indicated in equations (A2), (A3), (A4), and (A5) for each of the partials of a tone, over the entire time interval spanned by the tone. The output $A_{n\alpha}$ and $\theta_{n\alpha}$ may then be used in equation (A1) to synthesize a tone based on the analysis.

Now let us compute the response of the heterodyne filter as defined by equations (A2), (A3), (A4), and (A5) to a sinusoid of constant amplitude and phase. We do this by substituting for $F_i$ in equations (A2) and (A3) the function $\sin(\omega ih)$. We may compute the summations without error by use of the summation calculus (Hamming). Using the fact that $Nh\omega_\theta = 2\pi$, we may calculate $A_{n\alpha}$ and $\theta_{n\alpha}$ explicitly.

$$(A6)\quad A_{n\alpha} = \frac{1}{4N^2}\sin^2(\omega Nh/2)\left\{\frac{1}{\sin^2[(\omega+n\omega_\theta)h/2]} + \frac{1}{\sin^2[(\omega-n\omega_\theta)h/2]} + \frac{2\cos[n\omega_\theta h-2\phi_\theta]}{\sin[(\omega+n\omega_\theta)h/2]\,\sin[(\omega-n\omega_\theta)h/2]}\right\}$$

The expression for $\theta_{n\alpha}$ is a bit long and is thus not included here. As we consider the limit as $\omega$ approaches $n\omega_\theta$, we find great simplification of the results. Let us define $\Delta\omega$ as $(\omega-n\omega_\theta)$.

$$(A7)\quad \lim_{\omega\to\omega_\theta} A_{n\alpha} = \frac{1}{4N^2}\{0 + N^2 + 0\} = 1/4$$

$$(A8)\quad \lim_{\omega\to\omega_\theta} a_{n\alpha}/b_{n\alpha} = \frac{\sin\{2\omega_\theta h[(N-1)/2+\alpha]\} + N\sin\{\Delta\omega h[(N-1)/2+\alpha]\}}{\cos\{2\omega_\theta h[(N-1)/2+\alpha]\} + N\cos\{\Delta\omega h[(N-1)/2+\alpha]\}}$$

If $N>>1$ then (A8) reduces greatly.

$$(A9)\quad \lim_{\omega\to\omega_\theta} a_{n\alpha}/b_{n\alpha} = \tan\{\Delta\omega h[(N-1)/2+\alpha]\}$$

Thus we see that in the limit, $A_{n\alpha}$ approaches one quarter of the amplitude of the input sinusoid and $\theta_{n\alpha}$ is a term related to the difference in frequency of the input sinusoid with the analysis frequency. With instruments with partials whose frequencies deviate from the ideal, this provides a dynamic estimation of those frequencies.

Notice that neither $A_{n\alpha}$ nor $\theta_{n\alpha}$ are functions of $\phi_0$, the initial phase angle. Under the assumptions stated, this process is independent of initial phase. Notice also that $A_{n\alpha}$ is no longer a function of $\alpha$, the time parameter, but $\theta_{n\alpha}$ is. Fortunately, $\theta_{n\alpha}$ is a linear function of $\alpha$ whose slope is simply $\Delta\omega h$.

Figure A2 shows a plot of $A_{n\alpha}$ as indicated in equation (A1) for a range of values of $\omega$. In this case, $\omega_0$ is $2\pi(125\ Hz)$ and n is 4. We see that there is a zero of transmission at all integral multiples of $\omega_0$ except the $n^{th}$ multiple.

This technique is useful as long as the amplitudes and phases of the partials of the input waveform change slowly with time. If the frequencies of the partials deviate from integral multiples of the fundamental by too great an amount, further error may be introduced.

One must remember that the heterodyne filter as defined here is a *nonlinear* filter. Although we derived the response to a sinusoid, superposition does not hold in such a filter. The calculation of $a_{n\alpha}$ and $b_{n\alpha}$, however, is quite linear and superposition again applies. The general statements made above still hold, but only because of the special nature of the input signal.

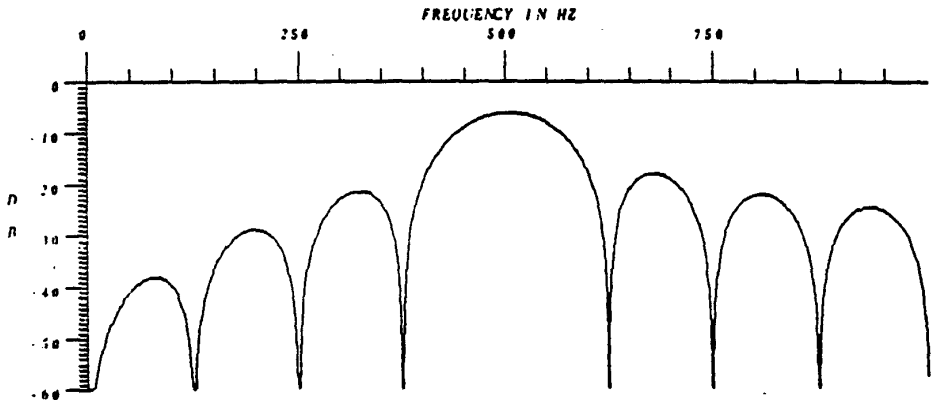



FIGURE A2· Log magnitude frequency response of the heterodyne filter [Moorer, 1973] for a base frequency of 125 Hz and n=4. There is a zero of transmission at every multiple of 125 Hz except the fourth one.

## Appendix B: Graphical Techniques

As aids to the researcher, we have designed several different methods for displaying the results of analysis. The output of the heterodyne filter can, of course, be displayed as a number of isolated amplitude and frequency functions, covering the individual components, as is shown in Figure A1 for the first eight harmonics of a tone. The total duration of the tone is about 300 milliseconds and its fundamental frequency is approximately 310 Hz. Seventeen harmonics were actually analyzed for this tone, but we present the isolated plots for only the first eight of these in Figure A1. Three more pages of such plots would be necessary to display the remaining harmonics. It is clear that there is much information in these plots, and furthermore, that the researcher needs more convenient ways of looking at the data.

To obtain a more easily-grasped picture of the relationships between all harmonics of a tone, it has been found most informative to view the entire set of harmonics together. One method designed for this is a three-dimensional perspective plot of Amplitude x Frequency x Time. Figure B1 illustrates such a plot of the amplitudes of all seventeen partials of this same tone. The fundamental appears as the backmost function in the picture, while the highest harmonic is represented as the frontmost function. This form of display allows us to more readily discover relationships among the harmonics. The perspective plot can be spatially rotated on-line by the computer, so that the observer is able to see the three-dimensional representation from any angle. This has been very helpful in getting a more comprehensive understanding of the behavior of the partials of a tone as a function of time.

Another form of display revealing the evolution of the partials of a tone as a function of time is the sequential line-spectrum plot. Here, we make use of animation techniques to present successive moments in the tone, presenting a plot of the amplitudes of all the harmonics at each moment in time. One plot is shown in Figure B2, taken from the middle of the above tone. This strictly on-line display presents such two-dimensional frequency by amplitude plots of the partials for successive instants in time, and the viewer can follow the amplitude changes for the partials from the beginning to the end of the tone.

A fourth way of examining the output of the heterodyne filter, inspired by the conventional speech spectrograph, is given in Figure B3. The particular advantage of this form of display is that it presents both frequency and amplitude information at once in a concise plot, allowing us to view relationships between the two as functions of time. Here, the thickness of each bar is proportional to the log of the amplitude of that harmonic. The vertical position represents its instantaneous frequency, as determined from the phase drift of the harmonic. The utility of this display is its representation of the phase information with respect to amplitudes.

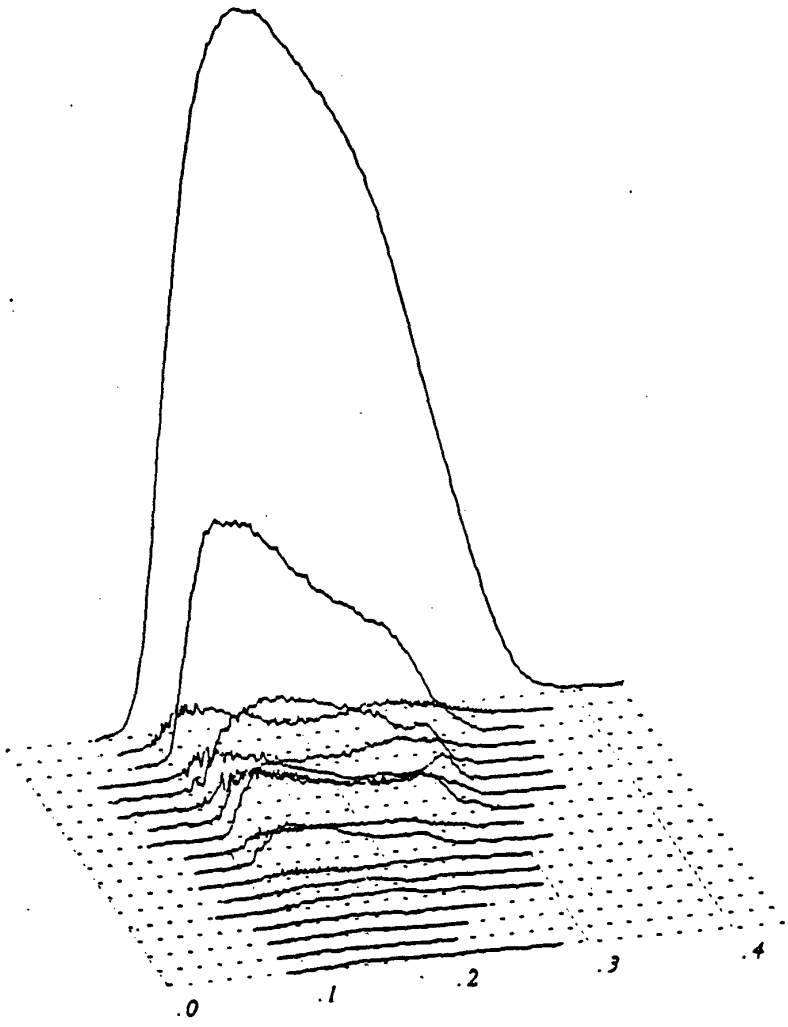



FIGURE - B1: Amplitude x Frequency x Time perspective plot of the time-variant amplitude functions of an analyzed tone. The X-axis = time, with 1/10 second divisions: the Y-axis = amplitude; the Z-axis = frequency. Individual harmonics are plotted, starting with the fundamental in the background and ending with the highest harmonic in the foreground. Note that the frequency domain as plotted is only an idealized presentation of the set of harmonic frequencies, and does not reflect any deviations in frequencies with time if present.

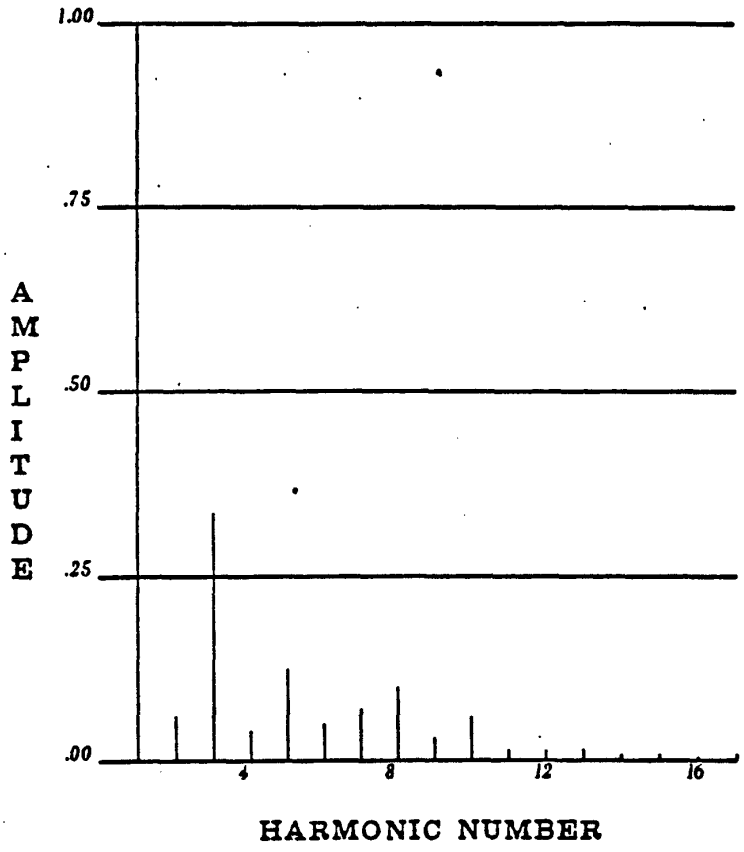



FIGURE B2. Line spectrum plot of the peak amplitudes for the harmonics of an analyzed tone. The X-axis = Harmonic number and the Y-axis = Amplitude (relative to the first harmonic). Individual harmonics are plotted starting with the fundamental on the left and ending with the highest harmonic on the right. Every 4th harmonic is numbered. A simulated movie may be produced by the computer using this type of display per 'frame' of the animation, so that the observer is able to see the individual harmonics grow and decay in amplitude with time.

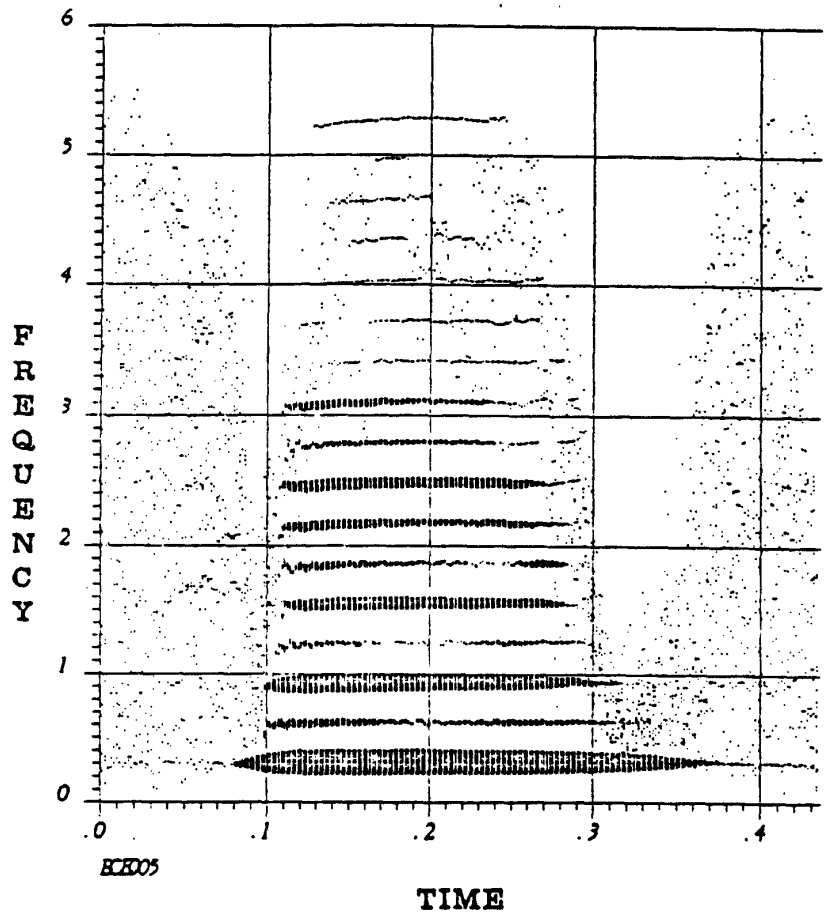



FIGURE B3: Spectrographic plot of an analyzed tone. The X-axis = time, with 1/10 second divisions, and the Y-axis = frequency, in KHz. Individual harmonics are shown starting with the fundamental on the bottom and ending with the highest harmonic on top. The width of the vertical bars plotted along each harmonic gives the intensity at that time. The minimal width, a point, is 40+ dB down from the maximum intensity of a single harmonic (in this case the fundamental). The frequency of a harmonic at any point in time is given by the center of the bar representing it. Various display parameters may be set by the observer: dB range down to the point display, maximal width of the bar, detail present in the time domain, etc.

*Interactive Function Editors*

To assist in the simplification and manipulation of the analyzed amplitude and frequency functions used for synthesis, programs have been developed that enable the researcher to easily specify changes to the computer and see the immediate results. Synthesis can be performed and the result heard in a matter of one or two minutes. One graphically-based editor deals with the two functions for a single harmonic at a time. Harmonics are displayed in order, and the amplitude and/or frequency functions may be altered in a number of ways, including smoothing, small number of line segment approximation and modulating; before and after versions may be stored, and array arithmetic performed, e.g. subtract the smoothed function from the complex version and save the results for later analysis of the micro-fluctuations. In addition, the commands given to the program may be saved and automatically applied to other harmonics or other notes, performing the same set of complex operations, and displaying the results as they are applied.

In order to assist the researcher in evaluating the importance of observed properties in an analyzed tone which has been vastly simplified, graphical techniques have been implemented to allow three-dimensional function editing, with the Amplitude x Frequency x Time perspective displays of tones. This editor is most useful in performing manipulations on a set of functions which are composed of small numbers of line segments. It allows the user to view the set of functions in three dimensions and rotate them into any position. The functions can be altered by attaching a cursor to any breakpoints of the line segments. The cursor can be moved either by light pen or teletype commands, and the function is altered appropriately. The resulting sound will be calculated upon exiting from the editor, and a playing program can be called to play the sound and compare it with the original sound and any other versions desired.

## Appendix C: Multidimensional Scaling Techniques

*Spatial Model for Perceptual Relationships*

It has been found most useful to employ a spatial model to represent the judged relationships between sets of stimuli, such as auditory signals. Computer-based multidimensional scaling algorithms have been developed for the reduction of the very complex data obtained from the subjective evaluations of perceived relationships between all members in a set of stimuli. The results are displayed in a form which is much more easily comprehended and interpreted by the investigator: that of a *geometric* configuration of points which represent the individual stimuli. The structure of the subjective evaluations of the set of stimuli is then mapped into an n-dimensional geometric space, where the distances between the points are determined by the measure of the *psychological* distances between all pairs of stimuli.

The psychological measures of distance which can be mapped into a spatial model include the *confusability* or the judged *similarity* of stimuli. One experimental procedure involves the identification of individual signals, perhaps presented under different conditions, and the result is a square confusion matrix of stimulus by response. A transform of this matrix yields the relative psychological distances, directly related to confusability, of all points in the data matrix. Another experimental procedure involves the rating of relative similarity for all pairs of stimuli in the set. The similarity ratings also can be placed in a square matrix, where each entry point is for one particular of all the possible ordered pairs of stimuli. The psychological distance in this case is inversely related to the similarity of stimuli, and is directly related to their dissimilarity.

The common characteristic of the scaling programs we find useful is their generation of an empirically-based representation of the relationships between the stimuli, rather than some theoretically imposed, *a priori* representation. We proceed from the *perceptual* data and will compare the representation of this data to the known physical attributes of the stimuli. The uncovering of psychophysical relationships, that is, making correlations between the subjective, psychological judgments of the stimuli and their physical properties, is essentially a matter of *interpreting* the representation of the perceptual data in terms of the known physical data for the stimuli. Note the significant difference of this more exploratory attitude from the traditional experimental approach, which begins with some *a priori* notions of the nature of the results and constructs the experiment and analyzes the data around these notions. The multidimensional scaling approach is useful when the stimuli are inherently so complex that we have no *a priori* notions and are willing to rely totally on the empirical method of analysis. We are concerned with both the dimensionality and the general properties of the space. Correlations with physical parameters are sought. Various programs exist which are useful in assessing the correspondence of data structures, so that it would prove fruitful to formally represent the physical data.

*Multidimensional Scaling Algorithms*

We will briefly describe the two types of multidimensional scaling, abbreviated MDS, found most useful in our research. One is a nonmetric MDS algorithm and the other is a metric individual differences scaling. These programs both attempt to represent input data matrices in the form of a configuration of points located in an n-dimensional geometric space, where n, the number of dimensions, is specifiable by the user. The points correspond to the stimuli, whose *psychological* distances are given in the input matrices. The coordinates of the points are obtained by an iterative computational algorithm which optimizes the correspondence of interpoint distances in the spatial representation to the measured psychological distances between the stimuli.

We first describe nonmetric MDS, a technique which originated with Shepard [1962a, 1962b] and was further developed by Kruskal [1964a, 1964b]. The program which we used, Kruskal's M-D-SCAL 5M with stress formula 2 and the primary approach to ties, was implemented on the PDP-10 computer by Phipps Arabie. For the nonmetric MDS, the optimal spatial representation of a subjective response matrix is one in which the rank order of the values of psychological distance are the same as the rank order of the interpoint distances in the n-dimensional geometric configuration. A monotonic function maps psychological values into distances in the spatial representation, the spatial distances being adjusted in a series of iterations to best fit the psychological distance structure by this criterion of monotonicity.

The individual differences scaling, which we shall refer to by the name INDSCAL, as termed by Carroll and Chang [1970], is a metric MDS algorithm developed to utilize the individual differences in sets of response matrices for analysis. The program that we use was implemented by Phipps Arabie for the PDP-10, and is the INDSCAL program described by Carroll and Chang [1970]. It generates an n-dimensional representation for the complete set of matrices by means of a three-way analysis - the basic two-way response matrix times the individuals. It analyzes the variations in the set of individual data matrices to uniquely determine a rotation for the axes in the space. Also produced by this analysis is a representation of weightings which account for individual response variations along the spatial dimensions of the idealized group stimulus configuration. The weightings are themselves mapped into an n-dimensional spatial configuration which can be used to assess the relationships between individual subjects or experimental conditions. See Carroll and Wish [1973] for an application of the method in the representation of confusion matrices for speech signals.

The use of both types of MDS is enhanced with clustering analysis programs which also deal with psychological distance matrices, but do not necessarily assume a dimensional space for the distances. One such program was developed by Johnson [1967], and it produces a tree-structure which represents the hierarchical clustering relationships of stimuli in the matrix as inferred from their psychological distances. This algorithm is the type that we have implemented for

clustering analysis. In our use of it, clustering proceeds from the set of individual stimulus points. The closest pair of stimuli are merged into a cluster, and treated then as a single point whose distance from any other point in the matrix will be the largest distance between any point in the cluster and the other point in question. This is the so-called diameter method. Shepard [1972] demonstrates the beneficial use of this analysis in conjunction with nonmetric MDS for the analysis of confusion matrices for speech signals.

# VI. BIBLIOGRAPHY

Ainsworth, W. A. Duration as a Cue in the Recognition of Synthetic Vowels. JASA 51, 648-651 (1972).

American Standard Acoustical Terminology. S1.1-1960. American Standards Association, Inc., New York (1960).

Backhaus, H. Ueber die Bedeutung der Ausgleichsvorgange in der Austrik. Zeit. Tech. Physik 13, 31-46 (1932).

Bartholomew, W. T. Acoustics of Music. Prentice-Hall, Inc., New Jersey (1945).

Beauchamp, J. W. A Computer System for Time-Variant Harmonic Analysis and Synthesis of Musical Tones. in Music by Computers, H. von Foerster & J. W. Beauchamp, ed. Wiley, New York (1969).

Berger, K. W. Some Factors in the Recognition of Timbre. JASA 36, 1888-1891 (1964).

v. Bismarck, G. Timbre of Steady Sounds: A Factorial Investigation of its Verbal Attributes. Acoustica 30, 146-159 (1974).

v. Bismarck, G. Sharpness as an Attribute of the Timbre of Steady Sounds.. Acoustica 30, 159-172 (1974).

Boring, E. G. Sensation and Perception in the History of Experimental Psychology. Appleton-Century Co., Inc., New York (1942).

Burns, E. M. and Ward, W. D. Categorical Perception of Musical Intervals. abstract, JASA 55 (1974).

Carroll, J. D., and Chang, J. J. Analysis of Individual Differences in Multidimensional Scaling Via an N-Way Generalization of "Eckart-Young" Decomposition. Psychometrika 35, 283-319 (1970).

Carroll, J. D. and Wish, M. Models and Methods for Three-Way Multidimensional Scaling. in Contemporary Developments in Mathematical Psychology, R. C. Atkinson, D. H. Krantz, R. D. Luce and P. Suppes, eds. W. H. Freeman, San Francisco (1973).

Chowning, J. M. The Synthesis of Complex Audio Spectra by Means of Frequency Modulation. J. Audio Eng. Soc. 21, 526-534 (1973).

Clark, M., Robertson, P., and Luce, D. A. A Preliminary Experiment on the Perceptual Basis for Musical Instrument Families. J. Audio Eng. Soc. 12, 199-203 (1964).

Clark, M., and Milner, P. Dependencies of Timbre on the Tonal Loudness Produced by Musical Instruments. J. Audio Eng. Soc. 12, 28-31 (1964).

Cutting, J. E. and Rosner, B. S. Categories and Boundaries in Speech and Music. Perception & Psychophysics 16(3), 564-570 (1974).

Ekman, G. Two Methods for the Analysis of Perceptual Dimensionality. Perc. Mot. Skills 20, 557-572 (1965).

Fletcher, H. Loudness, Pitch and the Timbre of Musical Tones and their Relation for the Intensity, the Frequency and the Overtone Structure. JASA 6, 59-69 (1934).

Freedman, M. D. Analysis of Musical Instrument Tones. JASA 41, 793-806 (1967).

Freedman, M. D. A Method for Analysing Musical Tones. J. Audio Eng. Soc. 16, 419-425 (Oct. 1968).

Gjaevenes, K., and Rimstad, E. R. The Influence of Rise Time on Loudness. JASA 51, 1233-1239 (1972).

Helmholtz, H. L. F. On the Sensations of Tone as a Physiological Basis for the Theory of Music. A. J. Ellis, trans. Dover, New York (1954).

Johnson, S. C. Hierarchic Clustering Schemes. Psychometrika 32, 241-254 (1967).

Jost; E. Akustische und Psychomettische Untersuchungen an Klavinettenklangen, Arno Volk Verlag, Koln (1967). reviewed in Webster, et. al. (1970).

Jusczyk, P. W., Cutting, J. W. and Rosner, B. S. Categorical Perception of Non-Speech Sounds in the Two-Month-Old Infant. unpublished report, University of Pennsylvania, Philadelphia (1974).

Kruskal, J. B. Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis. Psychometrika 29, 1-27 (1964a).

Kruskal, J. B. Nonmetric Multidimensional Scaling: A Numerical Method. Psychometrika 29, 115-129 (1964b).

Lane, H. The Motor Theory of Speech Perception: A Critical Review. Psych. Review 74, 275-309 (1965).

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. Perception of the Speech Code. Psych. Review 74, 431-461 (1967).

Lichte, W. H. Attributes of Complex Tones. Journ. Exp. Psych. 28, 455-480 (1941).

Lichte, W. H. and Gray, R. F. The Influence of the Overtone Structure on the Pitch of Complex Tones. Journ. Exp. Psych. 49, 431 (1955).

Licklider, J. C. R. Basic Correlates of the Auditory Stimulus. in Handbook of Experimental Psychology, S. S. Stevens, ed. Wiley, New York (1951).

Luce, D. A. Physical Correlates of Nonpercussive Musical Instrument Tones. PhD thesis. MIT (1963).

Luce, D. A., and Clark, M. Duration of Attack Transients of Nonpercussive Orchestral Instruments. J. Audio Eng. Soc. 13, 194-199 (1965).

Mathews, M. V. The Technology of Computer Music. MIT Press, Mass. (1969).

Mathews, M. V., and Kohout, J. Electronic Simulation of Violin Resonances. JASA 53, 1620-1626 (1973).

Moorer, J. A. The Heterodyne Method of Analysis of Transient Waveforms. Artificial Intelligence Laboratory, Stanford University (1973).

Moorer, J. A. On the Segmentation and Analysis of Continuous Musical Sound by Digital Computer. Ph.D. Thesis, Stanford University (1974).

Nordenstreng, K. The Perception of Complex Sounds: Semantic Differential Attributes of Speech and Music. in Contemporary Research in Psychology of Perception, J. Jarvinen, ed. WSOY, Helsinki (1969).

Ohm, G. S. Ueber die Definition des Tones, Nebst Daran Geknupfter Theorie der Sirene und Ahnlicher Tonbildender Vorrichtungen. Ann. Phys. Chem. 59, 513-565 (1843).

Pisoni, D. B. and Lazarus, J. H. Categorical and Noncategorical Modes of Speech Perception along the Voicing Continuum. JASA 55(2), 328-333 (1974).

Plomp, R. The Ear as a Frequency Analyzer. JASA, 36, 1628-1636 (1964).

Plomp, R. Timbre as a Multidimensional Attribute of Complex Tones. in Frequency Analysis and Periodicity Detection in Hearing. R. Plomp and G. F. Smoorenburg, ed. A. W. Sijthoff, Leiden (1970).

Plomp, R., and Bouman, M. A. Relation Between Hearing Threshold and Duration for Tone Pulses. JASA 31, 749 (1959).

Plomp, R., and Levelt, W. J. M. Tonal Consonance and Critical Bandwidth. JASA 38, 548 (1965).

Plomp, R. and Mimpen, A. M. The Ear as a Frequency Analyzer, II. JASA 43, 764-767 (1968).

Plomp, R. and Steeneken, H. J. M. Effects of Phase on the Timbre of Complex Tones. JASA 46, 409-421 (1969).

Plomp, R. and Steeneken, H. J. M. Pitch vs. Timbre. Seventh Int. Congr. on Acoustics, Budapest (1971).

Risset, J. C. Computer Study of Trumpet Tones. Bell Telephone Labs, Murray Hill, New Jersey (1966).

Ritsma, R. J. Frequencies Dominant in the Perception of the Pitch of Complex Sounds. JASA 42, 191-198 (1967).

Roederer, J. G. Introduction to the Physics and Psychophysics of Music. Springer-Verlag, New York (1973).

Saldanha, E. L., and Corso, J. F. Timbre Cues for the Recognition of Musical Instruments. JASA 36, 2021-2026 (1964).

Shepard, R. N. The Analysis of Proximities: Multidimensional Scaling with an Unknown Distance Function. I. Psychometrika 27, 125-140 (1962a).

Shepard, R. N. The Analysis of Proximities: Multidimensional Scaling with an Unknown Distance Function. II. Psychometrika 27, 125-140 (1962b).

Shepard, R. N. Psychological Representation of Speech Sounds. in Human Communication, E. E. Davis and P. B. Denes, eds. McGraw-Hill, New York (1972).

Slawson, A. W. Vowel Quality and Musical Timbre as Functions of Spectrum Envelope and Fundamental Frequency. JASA 43, 87-101 (1968).

Solomon, L. N. Semantic Approach to the Perception of Complex Sounds. JASA 30, 421-425 (1958).

Stevens, S. S., and Davis, H. Hearing - Its Psychology and Physiology. Wiley, New York (1938).

Strong, W., and Clark, M. Synthesis of Wind-Instrument Tones. JASA 41, 39-52 (1967a).

Strong, W., and Clark, M. Perturbations of Synthetic Orchestral Wind-Instrument Tones. JASA 41, 277-285 (1967b).

Studdert-Kennedy, M., Liberman, A. M., Harris, K. S. and Cooper, F. S. Motor Theory of Speech Perception: A Reply to Lane's Critical Review. Psych. Review 77, 234-249 (1970).

Taylor, C. A. The Physics of Musical Sounds. American Elsevier Pub. Co., Inc., New York (1965).

Webster, J. C., Carpenter, A., and Woodhead, M. M. Identifying Meaningless Tonal Complexes. JASA 44, 606-609 (1968a).

Webster, J. C., Carpenter, A., and Woodhead, M. M. Identifying Meaningless Tonal Complexes II. Journ. Aud. Res. 8, 251-260 (1968b).

Webster, J. C., Carpenter, A., and Woodhead, M. M. Perceptual Constancy in Complex Sound Identification. Brit. J. Psych. 61, 481-489 (1970).

Wedin, L., and Goude, G. Dimension Analysis of the Perception of Instrumental Timbre. Scand. Journ. Psych. 13, 228-240 (1972).

Wessel, D. L. report to Psychometric Society Meeting, San Diego (1973).

Wessel, D. L. report to C.M.E., University of Calif., San Diego (1974).

Winckel, F. Music, Sound and Sensation: A Modern Exposition. Dover, New York (1967).

Young, R. W. Musical Acoustics. in McGraw-Hill Encyclopedia of Science and Technology. McGraw-Hill, New York (1960).

Zwicker, E., Flottorp, G., and Stevens, S. S. Critical Bandwidth in Loudness Summation. JASA 29, 548-557 (1957).

Zwicker, E. and Scharf, B. A Model of Loudness Summation. Psych. Review 72(1), 3-26 (1965).