Perceptual evaluations of synthesized musical instrument tones

John M. Grey, and James A. Moorer

Citation: The Journal of the Acoustical Society of America **62**, 454 (1977); doi: 10.1121/1.381508
View online: https://doi.org/10.1121/1.381508
View Table of Contents: https://asa.scitation.org/toc/jas/62/2
Published by the Acoustical Society of America

Perceptual evaluations of synthesized musical instrument tones

John M. Grey and James A. Moorer

Center for Computer Research in Music and Acoustics, Department of Music, Stanford University, Stanford, California 94305
(Received 1 May 1976; revised 21 April 1977)

An analysis-based synthesis technique for the computer generation of musical instrument tones was perpetually evaluated in terms of the discriminability of 16 original and resynthesized tones taken from a wide class of orchestral instruments having quasiharmonic series. The analysis technique used was the heterodyne filter—which produced a set of intermediate data for additive synthesis, consisting of time-varying amplitude and frequency functions for the set of partials of each tone. Three successive levels of data reduction were applied to this intermediate data, producing types of simplified signals that were also compared with the original and resynthesized tones. The results of this study, in which all combinations of signals were compared, demonstrated the perceptual closeness of the original and directly resynthesized tones. An orderly relationship was found between the form of data reduction incurred by the signals and their relative discriminability, measured by a modified AAAB discrimination procedure. Direct judgments of the relative sizes of the differences between the tones agreed with these results; multidimensional scaling of the latter data provided a visual display of the relationships among the different forms of tones and pointed out the importance of certain small details existing in the attack segments of tones. Of the three forms of simplification attempted with the tones, the most successful was a *line-segment approximation* to time-varying amplitude and frequency functions for the partials. The pronounced success of this modification strongly suggests that many of the microfluctuations usually found in the analyzed physical attributes of music instrument tones have little perceptual significance.

PACS numbers: 43.75.Wx, 43.66.Lj

INTRODUCTION

Research in the area of timbre perception is largely concerned with the relationships between the psychological and physical attributes of musical sounds. The use of analysis-based synthesis techniques for the computer generation of music instrument tones is an invaluable step in timbre research. Recently developed digital technique for the analysis of music instrument tones with quasiharmonic series[1-5] make available to the investigator a detailed description of the physical properties of a signal. (Note the use of the term quasiharmonic series is necessary in that the partials of actual musical signals have small deviations from perfect integer ratios in frequencies). Furthermore, the precision of digital synthesis allows the generation of a signal based upon the most minute details of the analysis. Used in conjunction, an analysis-based synthesis process produces a *resynthesized* musical tone whose physical properties are entirely described, being modeled on some *original* tone.

Clearly, a direct measure of the adequacy of an analysis technique employed to model music instrument tones is the perceived closeness of the resynthesized tone to the original. Among the several attempts to analyze and resynthesize musical tones[1-4] various degrees of success were reported in perceptually approximating the original signals. Generally, however, no rigorous perceptual measurements were taken.

Given a successful analysis-based synthesis technique, the investigator has a tool with which to examine the perceptual salience of various physical aspects of the signal. There exists the problem of the enormous complexity of the physical parameters given by analysis. It is therefore critical to determine the amount of detail in the specification of physical properties which is necessary for resynthesis. Detail that is not perceptually important may be omitted from the physical attributes of the resynthesized signal, thus producing a *data-reduced* tone.

The task of uncovering psychophysical relationships in the perception of signals should be greatly aided by the elimination of perceptually unessential physical properties; the process of data reduction thereby narrows in on the attributes of a signal that are relevant to its perception. Furthermore, the perceptual alteration of a tone with data reduction will reveal the significance of an eliminated physical property. The value of data reduction in formulating psychophysical relationships in timbre perception has been demonstrated by Risset,[6] where important attributes of brass instrument tones were uncovered.

In this paper we report the results of two related investigations: (1) an evaluation of the perceptual success of the analysis-based synthesis technique (see below) by comparing original and resynthesized tones; and, (2) an examination of the results for three successive levels of physical parameter simplification in the production of data-reduced tones. To perform these evaluations, we measured the ability of trained musicians to discriminate between the original, resynthesized, and data-reduced forms of the signals. We also had the listeners judge the relative sizes of heard differences between the signals. The various forms of 16 tones chosen from a wide class of orchestral instruments were evaluated in this way.

I. ANALYSIS AND SYNTHESIS TECHNIQUES

A heterodyne filter technique[5] was used to analyze the tones. We represent the signal as follows:

$$X(n) = \sum_{k=1}^{M} A_k(n) \sin\{nT[k\omega + 2\pi F_k(n)]\}, \quad (1)$$

where $X(n)$ is the signal at time nT, n is the sample number (time index), T is the time between consecutive samples, ω is the radian fundamental frequency of the note, k is the partial number, $A_k(n)$ is the amplitude of partial k at time nT, it is assumed to be *slowly* time-varying, M is the number of partials, and $F_k(n)$ is the frequency deviation of partial k at time nT. It is assumed to be *slowly* time varying.

The waveform is represented here by a sum of sinusoids with slowly time-varying amplitudes and frequencies. It is then the purpose of the analysis technique to estimate these amplitudes and frequencies independently for each partial. Equation (1) is both the *model* of the waveform and the *synthesis technique* for the class of nearly harmonic signals. As a model of a waveform, we say that $X(n)$ is a sampled-data function that can be decomposed into M partials. As a synthesis technique, we say that $X(n)$ represents the nth output sample of a waveform being synthesized, and that waveform can be computed as the sum of M sinusoids. Specifying values for the various numbers and functions involved (ω, M, T, $A_k(n)$, and $F_k(n)$), specifies entirely the waveform for all time.

We can think of the analysis technique as a bank of analysis devices, each one of which extracts one partial from the waveform and detects its frequency and amplitude as a function of time. The technique we used to estimate these frequency and amplitude functions for the current study is as follows:

$$a_k(n) = \sum_{r=n}^{n+N-1} X(r) \sin(rk\omega_0 T), \quad (2)$$

$$b_k(n) = \sum_{r=n}^{n+N-1} X(r) \cos(rk\omega_0 T), \quad (3)$$

$$\hat{A}_k(n) = [a_k^2(n) + b_k^2(n)]^{1/2}, \quad (4)$$

$$\hat{\theta}_k(n) = \operatorname{atan}\left(\frac{a_k(n)}{b_k(n)}\right), \quad (5)$$

$$\hat{F}_k(n) = \frac{1}{2\pi} \frac{d\hat{\theta}_k(n)}{dt}, \quad (6)$$

where ω_0 is the angular frequency of analysis, k is the partial number, $X(n)$ is the input waveform at time nT, and N is the integer nearest the number of samples in one period of the input waveform. Note that ω_0 should be set to $2\pi/NT$. The hats on the amplitude, angle, and frequency variables (A, θ, F) indicate that these are *estimates* of the actual values and are not necessarily equal to the exact values that may have generated the tone.

In Eq. (6), we show a time derivative. Since the function we are differentiating is a sampled-data function, we must mention how this differentiation is to be accomplished. In this study, this was done by performing linear regressions on the phase curves taking the slope of these regressions to be the derivative. Forming the derivative in this manner is similar to a band-limited differentiation in that additional smoothing of the curves is accomplished.

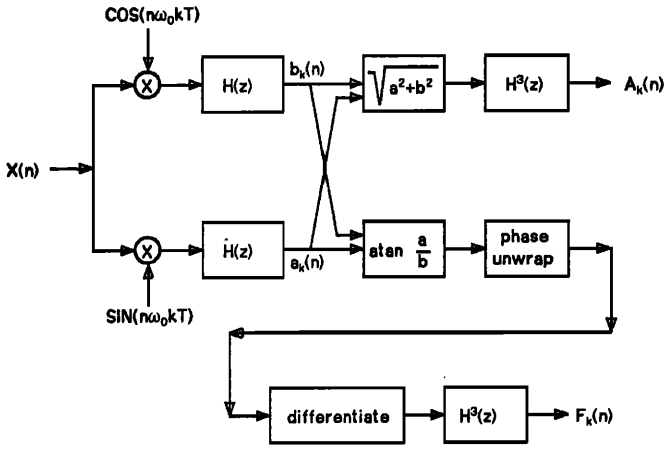

FIG. 1. Example of heterodyne filter analysis for a single partial. This procedure is repeated with successive values of k to get each partial separately.

Note that the fundamental frequency of the tone must be known, and it must be nearly constant over the interval of analysis. We rely on the summation over one period as in Eqs. (2) and (3) to place a zero of transmission at the harmonically related frequencies other than that of the partial under analysis. A zero of transmission is a frequency where the sum shown in Eqs. (2) and (3) is exactly zero, where the input signal is exactly annihilated. If the fundamental is chosen incorrectly, then the frequencies of the zeros will not necessarily correspond to the frequencies of adjacent harmonically related partials. This means that the results will not correspond to a single partial, but to some kind of nonlinear combination of partials. Our goal with this technique is to extract each partial separately.

Empirical studies have been done by Moorer[5] by analyzing synthetic signals of known harmonic content. These studies seem to indicate that the estimation method can tolerate no more than a 2% deviation (error) in the fundamental frequency of the input signal. This shows that this method is only useful for *isolated* musical tones with *no vibrato*. This constrained the kinds of tones that were selected for this study.

The functions $A_k(n)$ and $F_k(n)$ are filtered by averaging over one period. The Z transform of such a filter is $(1/N)(1-Z^{-N})/(1-Z^{-1})$. This filter places a zero of transmission at all harmonic frequencies except the one under analysis in addition to providing a roughly 6-dB/octave rolloff. This method is shown in Fig. 1 for the kth partial of a quasiperiodic signal. The input signal is first multiplied by a sine and cosine at the nearest harmonic frequency of the partial in question. The resulting products are then averaged over one period of the fundamental frequency of the signal. The resulting sequences are then in phase quadrature. We can get the amplitude of the partial by taking the square root of the sum of the squares of the two sequences. We can get the frequency of the partial by first getting the angle between the two sequences as computed by the arctangent function. This produces the *principal value* of

the phase angle. We must then regenerate the phase angle as a continuous function by "unwrapping" the principal value. Methods for doing this have been described by Schafer[7] and Moorer.[5] We can then differentiate the phase angle to get the frequency deviation. This deviation is the difference between the actual frequency of the partial and $k\omega_0$, the analysis frequency for that harmonic. These amplitude and frequency functions are then averaged over one period three more times to further cancel out the effects of adjacent partials on the analysis. Any "leakage" from adjacent partials will produce ripple on the estimated functions.

To actually use this technique in practice, we perform the following procedure. We first digitize an isolated tone from a musical instrument. We then determine its fundamental frequency as accurately as possible in any of a number of ways (see, for instance, Moorer[8]). Using this information, we extract each partial separately, using the process shown in Fig. 1 which is partially specified by Eqs. (2)–(6). This process must be applied M times, once for each partial that is present in the original waveform. In practice, we determine the value of M by applying the harmonic extractor to higher and higher partials until the output is no longer sensible, or until the amplitude detected is sufficiently small. An ALGOL program to do the computational portion of this procedure is given by Moorer.[5]

We shall not attempt here to extensively compare this method of analysis to other similar techniques.[1–4,9] The principle difference with Beauchamp's method is in how often the formulas (2)–(6) are evaluated. We evaluate these equations for every point in the sampled signal, rather than only "a few times per period" as in Beauchamp. The reason is that since Eqs. (4) and (5) specify a *nonlinear* transformation, this has the effect of increasing the bandwidth of the amplitude and phase functions, when they are considered sampled-data signals themselves. This increase in bandwidth means that a high sampling rate must be used to represent them. If the amplitude and frequency curves are not evaluated at the same rate as the original signal, then aliasing will occur. This aliasing is most noticeable in the transient regions of the tones, because that is where the frequency becomes non-band-limited. Quite realistic syntheses can be made using a much lower data rate, but we were concerned with making the synthetic tones as close to the original tones for the purposes of the study. For this reason, we chose to use high data rates on the analysis data.

Processed in this way for perceptual measurements were 16 notes of short duration, taken from the brass, string, and woodwind families of instruments.

II. STIMULI

The stimuli were derived from 16 instrumental notes played near E♭ above middle C, approximately 311 Hz whose durations ranged between 280–400 msec. Note that the control for uniformity in fundamental frequency resulted in taking tones from outside the middle ranges for several instruments; for instance, E♭ above middle C is towards the upper range of the bassoon and trombone while it falls into the lower range of the flute. The note is not a particularly unusual or difficult one on any of the instruments, however.

The tones were performed and recorded in an IAC acoustical isolation chamber having inside dimensions of 8.5×9 ft; the acoustical environment had very little reverberation. The recording was done on a Revox tape recorder, half-track monaural at 7.5 in. per sec, using Scotch 206 $\frac{1}{4}$ in. low-noise tape. They were then digitized by an Analogic 14-bit analogue-to-digital converter at a rate of 25 600 samples per second and the high-order 12 bits were stored in digital form for computer analysis or playback. The 16 instrumental tones consisted of: two oboes (different instruments and players), English horn, bassoon, E♭ clarinet, bass clarinet, flute, two alto saxophones (one instrument, played at p and mf), soprano saxophone, trumpet, French horn, muted trombone, and three celli (one instrument, played normally, muted sul tasto, and sul ponticello). These 16 tones were chosen to represent a wide variety of instrumental sounds.

The tones were digitally analyzed with a heterodyne filter technique, described above, producing a set of time-varying amplitude and frequency functions for each partial of the instrumental tone. Digital additive synthesis was used to produce a synthetic tone, where a set of partials, each controlled in amplitude and frequency through time, were added together. The time-varying amplitude and frequency functions may be considered as arrays of parameter values at successive, equally-spaced time intervals. For synthesis, a control function consisted of a temporal interpolation between these sampled parameter values.

Each of the 16 instrument notes appeared in at least four of the five following conditions: original tone, complex resynthesized tone, line-segment approximation, cut-attack approximation, and constant-frequencies approximation. These conditions are defined as follows: (1) *original* tone: the digitized waveform of a recorded instrument tone; (2) *complex resynthesized* tone: a tone generated by the analysis-based synthesis technique described above, using the complete results of the heterodyne analysis [see Figs. 2(a) and 2(b)]; (3) *line-segment* approximation: a tone produced by reducing the complex data from analysis to time-varying functions consisting of four to eight line segments, where there was an attempt to model original activity in the attack including initial, low-amplitude inharmonicity or noise—note that modeling was done in terms of simple line segments also [see Figs. 3(a) and 3(b)]; (4) *cut-attack* approximation: a further reduction from the line-segment approximation which excluded any clearly delineated initial segments of low-amplitude inharmonicity in the attack, the duration of which varied from 25 to 60 msec depending upon the tone—only 9 of the 16 tones were found to have such clearly delineated segments for elimination, so the procedure was applied to just those tones [see Figs. 4(a) and 4(b)]; and, (5) *constant frequencies* approximation: a tone which substituted constant frequencies for the time-variant frequency functions, but used the line-

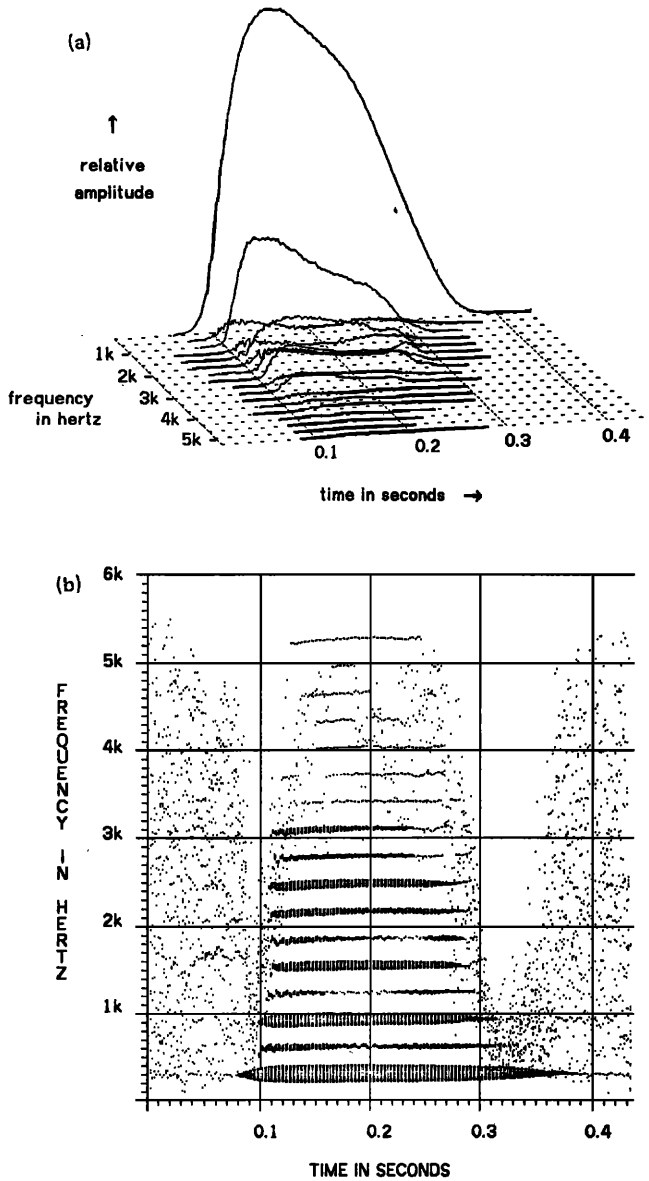

FIG. 2. (a) Time-varying amplitude functions derived from heterodyne analysis for bass clarinet tone, shown in three-dimensional perspective plot. The vertical axis is amplitude, the horizontal axis is time, and the depth axis is frequency, in which the functions for the partials are plotted in order of their frequencies—the fundamental appearing the furthermost behind. These functions were used for the complex re-synthesized tone (see text). (b) Spectrographic representation of heterodyne analysis for bass clarinet tone, plotted with frequency on the ordinate and time on the abscissa. Sampling of sound pressure levels presented at 3.6 msec intervals, as bars around the frequencies of the energy. Width of bars decreases from 200 Hz to a point as relative level drops from 0 to −40 dB compared to peak level of energy. Relative levels below −40 dB are also plotted as points.

segment approximations to the amplitude functions [see Figs. 5(a) and 5(b)].

All stimuli (including the original tones) were recorded on tape after being played back from the computer through a 12-bit digital-to-analogue converter at a rate of 25 000 samples per second. The recording was done on a Sony 854-4 tape recorder, quarter-track monaural at 7.5 in. per sec, on Maxell U50 $\frac{1}{4}$ in. low-noise tape. In the synthesized conditions, 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. However, enhanced discriminability was still expected between the original tone and the synthesized tones due to the interaction of the noise in the signal with the analogue-to-digital converter, which was not accurately modeled just by adding the tap hiss into the synthetic signals. (Note that the preferable technique would bypass tape altogether and directly record the signals into the computer through an analogue-to-digital converter; but this requires a recording studio linked directly with the computer.)

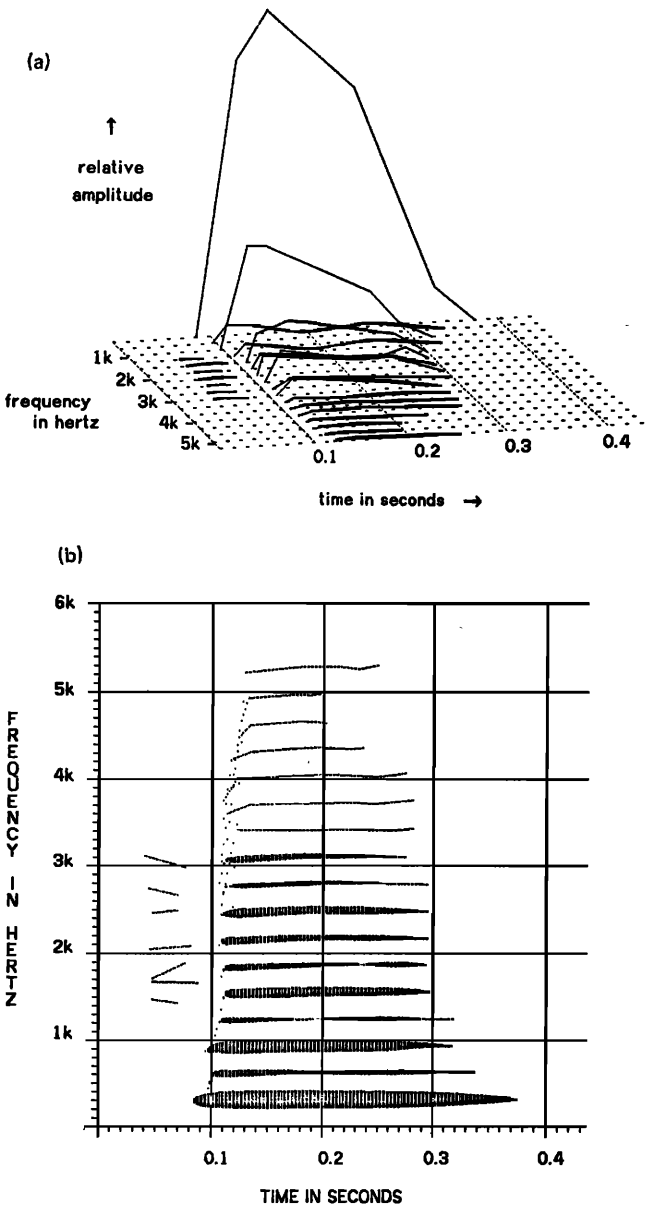

FIG. 3. (a) Time-varying amplitude functions, as in Fig. 1(a). These functions were derived from the complex analyzed functions by using six to eight line-segment approximations, and were used for the line-segment approximation in the synthesis of a data-reduced tone (see text). (b) Spectrographic plot, as in Fig. 1(b). Shown here is the line-segment approximation used for the synthesis of a data-reduced tone.

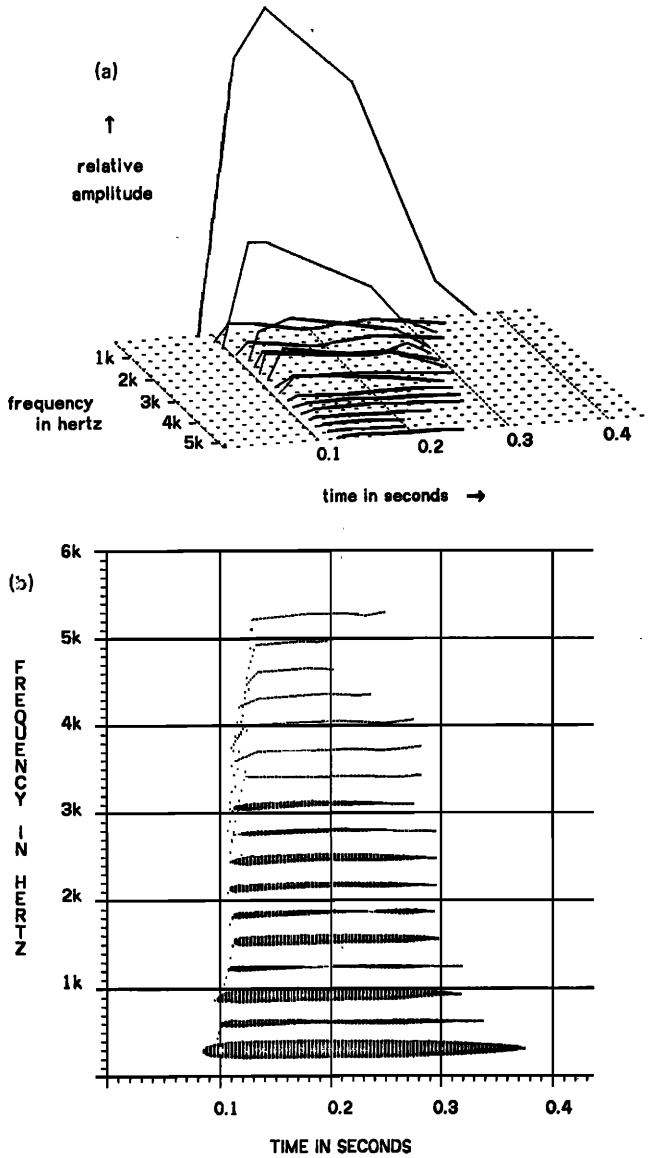

FIG. 4. (a) Time-varying amplitude functions, as in Fig. 1(a). These functions were derived from the line-segment approximations by deleting the initial low-amplitude content of the upper partials, and were used for the cut-attack approximation in the synthesis of a data-reduced tone (see text). (b) Spectrographic plot, as in Fig. 1(b). Shown here is the cut-attack approximation used for the synthesis of a data-reduced tone.

TABLE I. Octave-band sound pressure levels for E♭ Clarinet using B & K sound-level meter in the experimental laboratory. Data presented are the sound pressure level differences between signal +noise and noise alone measured in a given octave band.

Frequency Hz	Octave-band sound level dB
125	17
250	50
500	44
1000	47
2000	59
4000	52
8000	31

Playback was accomplished by means of a Dynaco PAT-4 preamplifier and Dynaco Stereo 120 amplifier connected to an Altec 604-E loudspeaker, and was done in a moderately reverberant room approximately 12×12 ft in dimensions. Playback level was measured at a distance of 3 ft from the speaker with a B & K sound level meter for the E♭ clarient tone and the data are given in Table I. This level was generally a comfortable, moderate level for all listeners.

III. LISTENERS AND PROCEDURE

Sixteen listeners at Stanford University were employed for this experiment. Listeners were musically sophis-

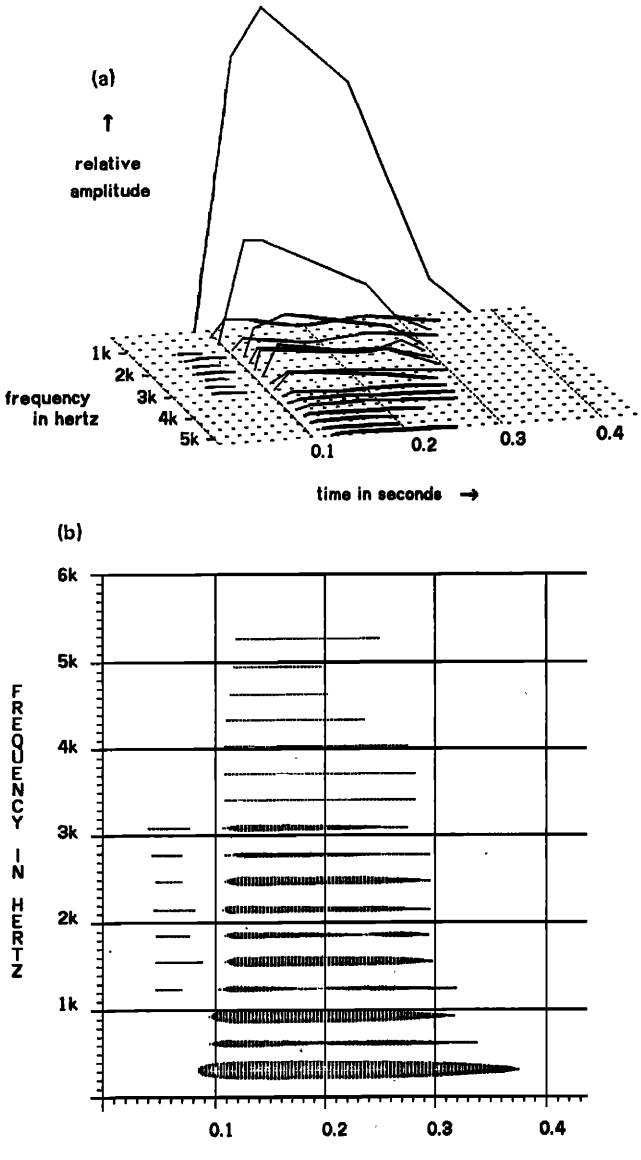

FIG. 5. (a) Time-varying amplitude functions, as in Fig. 1(a). These functions were derived from the line-segment approximations by setting the frequencies of the partials to be constants chosen by frequency averaging. These functions were used for the constant-frequencies approximation in the synthesis of a data-reduced tone (see text). (b) Spectrographic plot, as in Fig. 1(b). Shown here is the constant-frequencies approximation used for the synthesis of a data-reduced tone.

ticated, some actively involved in advanced 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 ft 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.

All possible pairwise combinations of the applicable conditions per instrument note were used to compose the trials, so there were ten basic pairwise trial-combinations for the nine instrument notes having five tonal conditions, and six basic pairwise trial combinations for the other seven instrument notes having only four 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 nonpractice trials in the two sessions.

The trials were structured in an AAAB discrimination paradigm. On each trial four tones were played: Three were *identical* and the fourth was *different*, and the position of the *different* tone was randomly selected. The first and last two tones were separated in times of onset by 1 sec, while the second and third tones were separated by 2 sec, 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 msec. There was 2 sec of silence between the warning knock and the first tone of the trial; 4 sec of silence followed each trial before the warning knock for the next trial, in which listeners could make their responses.

Listeners were asked to discriminate which tone of the four notes per trial was in any way different than the others; the instructions called upon listeners to rate how *equally* a sequence of four 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 four-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, on a scale of one to ten. 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. In the event that no difference was heard, listeners were told to indicate this by a mark on their answer sheet.

Note that the instructions did not call for listeners to judge which note had a different *timbre* of the four note sequence, but instead asked for the judgment of difference in terms of the perceived quality of articulation or playing style. Generally, the different forms of a note did not appear to change in terms of instrumental source, but rather exhibited small qualitative alterations that resembled differences that may occur in playing the same note on an instrument with slightly different styles. Although these differentiations do properly fall into the domain of timbre discrimination, we felt that the musical education of our listeners called for the use of the more specific terminology employed here—rather than the word "timbre" which could have given rise to ambiguities and misunderstanding.

IV. RESULTS

The collected data were averaged over the four repeated measurements per trial combination for each listener. The discrimination score was computed by adding the number of times the different pair was correctly identified to half the number of times that "no differences" were heard as explicitly marked by listeners. The latter was a simulation of random guessing for those trials. The sum was divided by four to produce a mean discrimination score. In that the distance estimates were to be used as measurements of the relative sizes of heard differences between tonal conditions, the distance scores were computed from correct trials only. The averaged distance estimation equaled the sum of the estimates on the correct trials, divided by the number of correct trials in that trial combination.

One and two-way analyses of variance revealed no significant differences in discrimination scores for the four possible positions of the "different" tone in the AAAB paradigm. The overall amount of trials in which "no differences" were reported averaged 25% of the cases. Individuals varied considerably in the number of times they reported "no differences," the standard deviation being 16%. Again, no significant differences were found in the occurrences of "no differences" reported among the four possible position of the "different" tone in the AAAB task.

The mean subject scores were 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 per tone appear in Fig. 6 and the mean distance estimates per note are given in Fig. 7. Figure 8 presents the overall means for discrimination scores and distance estimates for the collection of tones taken as a whole. The correlation between the discrimination scores and distance estimations per trial-combination was 0.92. This is presented graphically in Fig. 9, in which the ordinate represents the distance estimates and the abscissa represents the discrimination scores. An expected negative correlation was found between the discrimination scores and the percentages of "no differences" reported per trial-combination, computed to be −0.89.

In order to explore and interpret relationships among the various tonal conditions for the 16 instrumental timbres, the distance estimates were considered to indicate the relative sizes of the subjective differences heard between the tones. These measures are appropriate for analysis with multidimensional scaling techniques, which

	.727					.594			
	.742	.586				.617	.656		
TRP	.867	.844	.766		BSN	.836	.859	.688	
	.898	.797	.617	.570		--	--	--	--
	.531					.500			
	.648	.563				.891	.680		
FHRN	.703	.688	.602		BCLR	.875	.828	.758	
	.750	.531	.547	.602		.914	.828	.742	.859
	.672					.727			
	.609	.531				.750	.570		
TRB	.727	.672	.648		EHRN	.867	.648	.625	
	--	--	--	--		--	--	--	--
	.781					.641			
	.906	.805				.883	.805		
CEL1	.914	.828	.633		SAX1	.930	.898	.727	
	--	--	--	--		.938	.742	.703	.805
	.680					.742			
	.906	.742				.898	.695		
CEL2	.938	.820	.680		SAX2	.984	.844	.656	
	--	--	--	--		.961	.883	.711	.766
	.797					.828			
	.836	.570				.930	.578		
CEL3	.906	.641	.563		SSAX	.999	.891	.922	
	--	--	--	--		.883	.742	.570	.930
	.617					.672			
	.641	.477				.859	.563		
OBO1	.664	.633	.641		FLT	.930	.758	.711	
	.820	.734	.719	.859		--	--	--	--
	.648					.648			
	.625	.602				.617	.602		
OBO2	.719	.703	.641		ECLR	.844	.742	.648	
	.617	.547	.578	.734		.820	.789	.664	.711

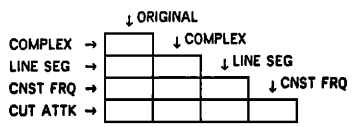

FIG. 6. Mean discrimination scores for instrumental tones compared under five experimental conditions: (1) original tone in digitized waveform; (2) complex tone re-synthesized; (3) line-segment approximation; (4) cut-attack approximation; and (5) constant frequencies approximation. Note that seven tones were not tested for all five conditions (see text). The form of the entries for each matrix is given at the bottom of the figure; one matrix is provided for each of the 16 instrumental tones.

produce a geometric "map" best representing the relative sizes of subjective differences observed among a stimulus set.

The distance estimates were then treated with a multidimensional scaling algorithm which takes individual differences in input matrices into account during the computation.[10] A three-way analysis was performed upon 16 lower-half matrices, each representing the comparison of tonal condition pairs for one of the 16 individual instrumental timbres. The resulting "map" represented the overall relationship of the five tonal conditions for the 16 instrumental timbres. A one-dimensional solution was obtained for the common four conditions for all 16 instrumental notes. Consistent with the ordering of the degrees of modification in the four conditions, the ordering along the dimension was *original* tone, *complex resynthesis*, *line segment* approximation, and *constant frequencies* approximation.

Somewhat more informative was the solution in two dimensions obtained for all five conditions common to nine of the instrumental notes. The stimulus solution is given in Fig. 10. 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.

A closer examination of the matrices show that the 16 instrumental notes did not behave in exactly the same way with respect to the set of modification conditions imposed upon them. There is a wide range of variance between the 16 instrumental notes in those trials that paired the *original* tone with another 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,

	2.75					2.48			
	3.71	2.28				2.00	2.37		
TRP	5.57	4.01	3.15		BSN	4.27	3.40	3.13	
	4.69	3.15	2.72	1.84		--	--	--	--
	1.70					1.25			
	2.58	2.05				4.24	2.95		
FHRN	3.30	2.83	1.84		BCLR	4.58	4.48	3.31	
	2.74	1.59	1.44	2.10		4.66	4.27	3.27	4.29
	2.46					3.11			
	2.48	1.60				3.78	2.12		
TRB	3.51	2.76	2.73		EHRN	4.51	1.91	1.95	
	--	--	--	--		--	--	--	--
	3.73					1.96			
	4.32	3.25				4.97	3.83		
CEL1	5.25	3.77	2.43		SAX1	5.85	4.78	2.91	
	--	--	--	--		4.93	3.46	2.89	3.17
	3.53					3.79			
	5.23	4.49				5.63	3.23		
CEL2	6.60	4.60	3.23		SAX2	6.94	3.82	3.25	
	--	--	--	--		6.31	4.92	3.12	3.17
	3.30					4.57			
	4.80	2.27				5.52	1.78		
CEL3	3.94	2.59	2.45		SSAX	7.51	5.59	4.51	
	--	--	--	--		5.13	3.07	2.85	4.70
	2.02					3.24			
	2.05	1.52				4.72	2.33		
OBO1	2.64	2.05	2.51		FLT	5.49	3.26	2.67	
	4.02	3.03	3.44	4.13		--	--	--	--
	2.09					2.66			
	2.13	1.66				2.37	2.81		
OBO2	3.01	2.49	2.47		ECLR	4.04	2.64	3.06	
	2.41	2.24	3.16	2.90		4.24	3.95	2.65	3.58

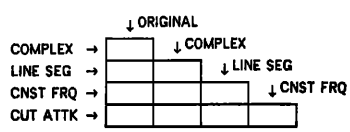

FIG. 7. Mean distance estimations for instrumental tones compared under five experimental conditions, as presented in Fig. 6.

.675

.772 .627

.856 .769 .682

.845 .733 .650 .760

DISCRIMINATION

2.79

3.78 2.53

4.81 3.44 2.85

4.35 3.30 2.84 3.32

DISTANCE

FIG. 8. Overall discrimination scores and distance estimates for the conditions compared, averaged over all 16 stimulus tones. The form of the entries for the matrices is as presented in Fig. 6 and 7.

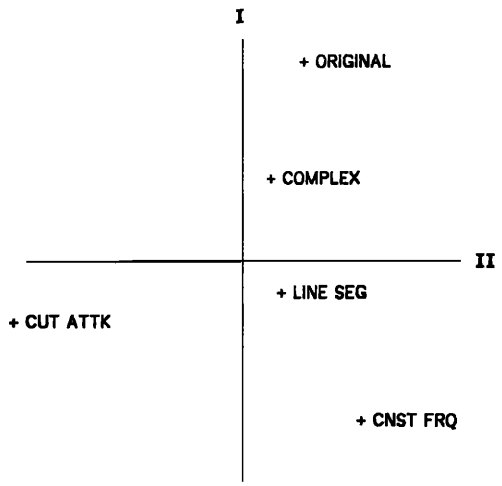

FIG. 10. Two-dimensional scaling solution for distance estimates obtained from multidimensional sealing. The five conditions are labeled on the plot (see text for interpretation).

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. Found in the matrices 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 high 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.

V. CONCLUSIONS

We have reported the perceptual discrimination and subjective distance estimations between the original, resynthesized, and data-reduced tones for a wide class of tones. Of 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 on the tones in their discrimination.

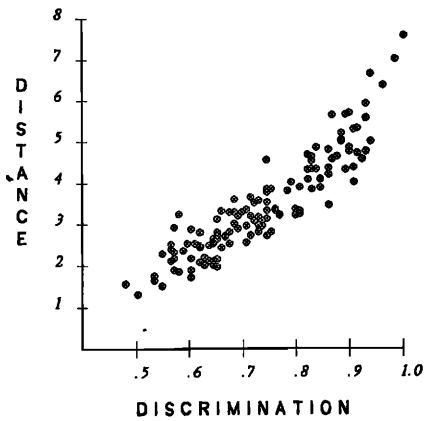

FIG. 9. Mean discrimination scores (on the abscissa) plotted against mean distance estimates (on the ordinate), given for each condition of each stimulus tone. Pearson product-moment correlation was 0.92 (see text).

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.[11–13] 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 various 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 compared to differences that occur in musical situations. The large percentage of "no differences" reported corresponds to this subjective report. Tones were heard to be lifelike, but having slightly different articulations or styles of playing. It generally was *not* the case that the original was heard to be natural while the synthetic versions were not.

The discrimination scores for many of the *original* tones versus their *resyntheses* were better than expected. 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-varying noise of the *original* signal in the other conditions. No rigorous conclusion can be made with these data. Ideally, however, the analysis–synthesis loop should be perfectly transparent, both mathematically and perceptually. There now exist newer methods which can deliver this transparency that were not available for the current study.[14]

Differences between the *complex resynthesis* and the *line segment* approximation conditions were harder to distinguish, yielding performance close to chance for over half of the instrumental cases. The finding suggests that the highly complex microstructure in the time-varying amplitude and frequency functions given by the analysis is not absolutely essential to the timbre 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[6] with approximating the trumpet tone now to all families of orchestral instruments having quasiharmonic series. 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 resynthesis* and the *line segment* approximation suggests 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, timbre differences will be heard with this degree of modification.

Future uses of simplified synthesis schemes in timbre perception research are suggested by the investigations of Strong and Clark[15,16] on the interactions of various timbre cues. This was done by altering the physical parameters that controlled synthesis and by directly exchanging the cues from one sort of instrumental tone with those of a different instrumental source.

ACKNOWLEDGMENTS

This work was supported in part by NEA Grant C50-31-282 and by NSF Contracts BNS 75-17715 and DCR 75-00694. We would like to thank John Chowning, Loren Rush, Max Mathews, and Dave Wessel for their encouragement and advice in this investigation, and Elizabeth Dreisbach for running the experimental sessions. Also, we are indebted to the Stanford Artificial Intelligence Laboratory, whose computer facilities were used to perform this research.

[1]D. A. Luce, "Physical correlates of nonpercussive musical instrument tones," Ph.D. dissertation (M.I.T., 1963) (unpublished).
[2]M. D. Freedman, "Analysis of musical instrument tones," J. Acoust. Soc. Am. 41, 793–806 (1967).
[3]M. D. Freedman, "A method for analyzing musical tones," J. Audio Eng. Soc. 16, 419–425 (1968).
[4]J. W. Beauchamp, "A computer system for time-variant harmonic analysis and synthesis of musical tones," in *Music by computers*, edited by H. von Foerster and J. W. Beauchamp (Wiley, New York, 1969).
[5]J. A. Moorer, "On the segmentation and analysis of continuous musical sound by digital computer," Ph.D. dissertation Stanford University Department of Music Technical Report, STAN-M-3 (1975).
[6]J. C. Risset, "Computer study of trumpet tones," Bell Telephone Labs, Murray Hill, NJ (1966) (unpublished).
[7]R. W. Schafer, "Echo removal by discrete generalized linear filtering," Ph.D. dissertation (M.I.T., 1969) (unpublished).
[8]J. A. Moorer, "The optimum comb method of pitch period analysis of continuous digtized speech," IEEE Trans. Acoust. Speech Signal Process. ASSP-22, 303–338 (1974).
[9]J. W. Beauchamp, "Analysis and synthesis of cornet tones using nonlinear interharmonic relationships," J. Aud. Eng. Soc. 23, 778–795 (1975).
[10]J. D. Carroll and J. J. Chang, "Analysis of individual differences in multidimensional scaling via an *N-way* generalization of 'Eckart–Young' decomposition," Psychometrika 35, 238–319 (1970).
[11]E. L. Saldanha and J. F. Corso, "Timbre Cues for the Recognition of Musical Instruments," J. Acoust. Soc. Am. 36, 2021–2026 (1964).
[12]K. W. Berger, "Some factors in the recognition of timbre," J. Acoust. Soc. Am. 36, 1888–1891 (1964).
[13]L. Wedin and G. Goude, "Dimension analysis of the perception of instrumental timbre," Scand. J. Psychol. 13, 228–240 (1972).
[14]J. A. Moorer, "The use of the phase vocoder in computer music applications," Audio Eng. Soc. preprint number 1146 (E-1), 55th convention (1976).
[15]W. Strong and M. Clark, "Synthesis of wind-instrument tones," J. Acoust. Soc. Am. 41, 39–52 (1967).
[16]W. Strong and M. Clark, "Perturbations of synthetic orchestral wind-instrument Tones," J. Acoust. Soc. Am. 41, 277–285 (1967).