Discrimination of musical instrument sounds resynthesized with simplified spectrotemporal parameters[a)]

Stephen McAdams[b)]
Laboratoire de Psychologie Expérimentale (CNRS), Université René Descartes, EPHE, 28 rue Serpente, F-75006 Paris, France and Institut de Recherche et de Coordination Acoustique/Musique (IRCAM/CNRS), I place Igor-Stravinsky, F-75004 Paris, France

James W. Beauchamp[b)]
School of Music and Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, 2136 Music Building, 1114 West Nevada Street, Urbana, Illinois 61801

Suzanna Meneguzzi
Laboratoire de Psychologie Expérimentale (CNRS), Université René Descartes, EPHE, 28 rue Serpente, F-75006 Paris, France and IRCAM, I place Igor-Stravinsky, F-75004 Paris, France

(Received 17 November 1997; revised 21 September 1998; accepted 23 September 1998)

The perceptual salience of several outstanding features of quasiharmonic, time-variant spectra was investigated in musical instrument sounds. Spectral analyses of sounds from seven musical instruments (clarinet, flute, oboe, trumpet, violin, harpsichord, and marimba) produced time-varying harmonic amplitude and frequency data. Six basic data simplifications and five combinations of them were applied to the reference tones: amplitude-variation smoothing, coherent variation of amplitudes over time, spectral-envelope smoothing, forced harmonic-frequency variation, frequency-variation smoothing, and harmonic-frequency flattening. Listeners were asked to discriminate sounds resynthesized with simplified data from reference sounds resynthesized with the full data. Averaged over the seven instruments, the discrimination was very good for spectral envelope smoothing and amplitude envelope coherence, but was moderate to poor in decreasing order for forced harmonic frequency variation, frequency variation smoothing, frequency flattening, and amplitude variation smoothing. Discrimination of combinations of simplifications was equivalent to that of the most potent constituent simplification. Objective measurements were made on the spectral data for harmonic amplitude, harmonic frequency, and spectral centroid changes resulting from simplifications. These measures were found to correlate well with discrimination results, indicating that listeners have access to a relatively fine-grained sensory representation of musical instrument sounds. [S0001-4966(99)00202-7]

PACS numbers: 43.66.Jh, 43.75.Wx [WJS]

INTRODUCTION

It has been traditional to view musical sounds in terms of a spectral model that describes them as a series of sinusoidal components, each having an amplitude and a frequency. Often, as is the case in this article, these sounds have frequencies which are harmonically related to a fundamental frequency, or at least approximately so. While many experiments on timbre have used fixed frequencies and fixed relative amplitudes (Miller and Carterette, 1975; Plomp, 1970; Preis, 1984; von Bismarck, 1974), analyses of musical instrument sounds reveal that these parameters have a great deal of variation, leading to the conjecture that these variations are responsible, in large part, for the uniqueness of the individual sounds.

For example, we can think of the amplitudes (A) and frequencies (f) varying over time (t) and having two parts, a smoothly or slowly moving part (1) and a more rapidly changing microvariation part (2):

$$A_k(t) = A1_k(t) + A2_k(t), \tag{1}$$

$$f_k(t) = f1_k(t) + f2_k(t), \tag{2}$$

where k refers to the harmonic number. Alternatively, since we consider only quasiharmonic sounds here, we can also break the frequency into two other parts:

$$f_k(t) = k\overline{f_0}(t) + \Delta fi_k(t), \tag{3}$$

where $\overline{f_0}$ is the fundamental frequency averaged over several harmonics and Δfi_k is an inharmonic frequency deviation, both varying over time.

Figure 1 gives a block diagram of a spectral representation model using the parameters of Eqs. (1) and (2), which is also an additive, sine-wave-synthesis model. The question to be explored in this article is: to what degree can these parameters be simplified, without making them discriminable, with respect to sounds containing the full amount of information? A given sound can be reconstituted with high quality from the full representation using time-varying additive synthesis. However, such a representation is quite data inten-

[a)]Portions of these results were presented at the 133rd meeting of the Acoustical Society of America (Beauchamp *et al.*, 1997).

[b)]Address correspondence to either S. McAdams at IRCAM (Electronic mail: smc@ircam.fr) or to J. Beauchamp at UIUC (Electronic mail: j-beauch@uiuc.edu).

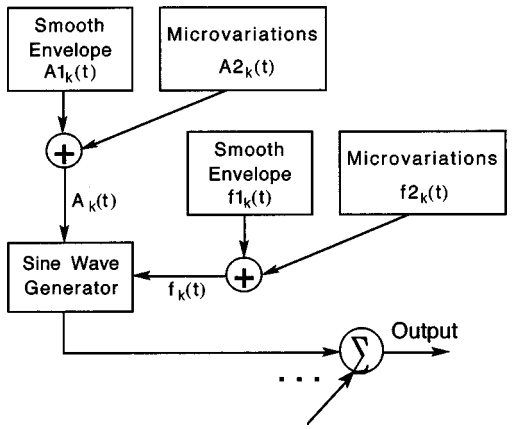

FIG. 1. Spectral-representation model using smooth and microvariation envelopes for amplitude and frequency. Each harmonic k is summed with the others to form the total output by additive synthesis.

sive. Any possibility of reducing the data would alleviate storage problems and accelerate the process of synthesis, which is particularly important for real-time sound synthesis. Also, one might hope that such simplifications would lead to the possibility of streamlined synthesis control using a few well-chosen, perceptually relevant parameters. Most important for us, however, is that knowledge about the sensitivity of human listeners to certain kinds of sound simplifications may provide clues for understanding the sensory representation of musical sounds. Specifically, this study is aimed at determining the relative perceptual importance of various spectrotemporal features which we have suspected are important for making timbral distinctions and for judging sound quality.

A few researchers have already addressed the problem of perceptually relevant data reduction using discrimination paradigms. Grey and Moorer (1977) used a rectangular-window, heterodyne-filter analysis algorithm and time-varying additive synthesis to prepare their stimuli based on 16 sounds from various bowed-string, woodwind, and brass instruments of duration 0.28 to 0.40 s. They asked their subjects (musical listeners) to discriminate between five versions of the sounds: (1) the digitized original analog tape recording, (2) a complete synthesis using all time-varying amplitude and frequency data resulting from the analysis stage, (3) a synthesis using a small number of line-segment approximations to the amplitude and frequency envelopes, (4) the same modification as version (3) with removal of low-amplitude initial portions of attack transients, and (5) the same modification as (3) with frequencies fixed in harmonic relation to the fundamental frequency (frequency-envelope flattening). Listeners heard four tones in two pairs and had to determine which pair contained a different tone. They were allowed to respond ''no difference heard.'' Discrimination scores were computed as the probability that the correct interval was chosen plus half the probability that a no difference response was given (ostensibly to simulate random guessing on those trials).

An important result was the low discrimination scores for comparisons of versions (2) and (3), which ranged from 0.48 to 0.80 (depending on the instrument), with an average of only 0.63. This indicated that microvariations in amplitude and frequency are usually of little importance, implying the possibility for significant data reduction. However, the authors gave no algorithm for fitting the straight lines to the data or criteria for error, but stated only that the number of segments varied between four and eight per parameter over each tone's duration. Also, since the tones were short and some segments were needed to fit attack transients, it is not clear how these results can be extrapolated for longer sounds. Discrimination rates between versions (3) and (4) and between (3) and (5) were similarly low, averaging 0.65 (range: 0.55 to 0.74) and 0.68 (range: 0.56 to 0.92), respectively. The results indicated that there were significant differences among the 16 instruments.

In general, discrimination rates for single simplifications were low, and relatively high rates (above 0.85) only occurred for multiple simplifications. For example, the average discrimination rate between versions (1) and (5), where three simplifications were combined, was 0.86. From our experience, these figures seem low. We can only conjecture that this was due to the short tones used, to noise on the analog tape used for stimulus presentation which may have masked some parameter variation details, and perhaps even to the experimental instructions which specifically oriented listeners toward differences in quality of articulation and playing style rather than toward any audible difference.

Charbonneau (1981) extended Grey and Moorer's study [based on their version (3) representation] by constructing instrumental sounds that maintained their global structure, while simplifying the microstructure of the amplitude and frequency envelopes of each harmonic partial. The first simplification was applied to the components' amplitude envelopes, each component having the same amplitude envelope (calculated as the average harmonic-amplitude envelope) scaled to preserve its original peak value and start- and end times. (This is similar to our amplitude-envelope coherence simplification; see Sec. I below.) The second simplification was similarly applied to the frequency envelopes, each having the same relative frequency variation as the fundamental, meaning that the sound remained perfectly harmonic throughout its duration (similar to our frequency-envelope coherence simplification; see Sec. I below). The third simplification resulted from fitting the start- and end-time data to fourth-order polynomials. Listeners were asked to evaluate the timbral differences between original [version (3)] and simplified sounds on a scale from 0 (no difference) to 5 (large difference). Results indicated that the amplitude-envelope simplification had the greatest effect. However, as for the Grey and Moorer study, the strength of the effect depended on the instrument.

Sandell and Martens (1995) used a different approach to data reduction. The harmonic time-frequency representation derived from a phase-vocoder analysis was treated as a data matrix that could be decomposed into a number of linearly recombinable principal components from either a temporal or a spectral perspective. The recombination of the appropriately weighted principal components can be used to regenerate the signal of a given instrument sound. These authors estimated the number of principal components necessary to

achieve a simplified sound that was not reliably discriminated from a sound reconstructed from the full (though down-sampled) analysis data. From these results, they could then compute the proportion of data reduction possible without compromising perceived sound quality. They achieved considerable data reduction for the three instruments tested, but the amount varied a great deal across instruments. One interpretation problem that often plagues perceptually oriented principal components analyses on acoustic data (see also, Repp, 1987) is that the perceptual nature and relevance of the individual components is most often difficult to conceive. For example, it is not clear that they could represent perceptual dimensions with clearly defined acoustic characteristics along which stimuli could be varied intuitively in sound synthesis.

This reservation notwithstanding, the results of these three studies demonstrate that timbre changes result from simplification of the signal representation. In fact, it is clear from the two earlier studies that the simplifications performed on temporal parameters, and specifically on time-varying functions of amplitude and frequency, influence to a greater or lesser degree the discrimination of musical sounds.

In the present study, we sought to determine precisely the extent to which simplified spectral parameters affect the perception of synthesized instrumental sounds, using tones of 2-s duration and without the use of straight-line approximations. We measured the discrimination of several kinds of simplifications for sounds produced by instruments of various families of resonators (air column, string, bar) and types of excitation (bowed, blown, struck). Two of the simplifications we chose (amplitude-envelope coherence and spectral-envelope smoothness) were derived from previous studies on timbre perception and corresponded to acoustic parameters that are highly correlated with perceptual dimensions revealed by multidimensional scaling techniques (Grey and Gordon, 1978; Iverson and Krumhansl, 1993; Krimphoff *et al.*, 1994; McAdams *et al.*, 1995). The other four simplifications related to the perceptibility of microvariations of amplitude and frequency over time, with much having been written about the latter (Brown, 1996; Dubnov and Rodet, 1997; McAdams, 1984; Sandell and Martens, 1995; Schumacher, 1992).

In addition, various combinations of these simplifications were applied to the sounds in groups of two, three, or four. We hypothesized that accumulated simplification along several perceptual dimensions would increase discrimination. Below, we present the technique used for analyzing and synthesizing the stimuli, followed by a description of the discrimination experiment. The results are then discussed in terms of musical synthesis and the perceptual representation of musical signals.

I. ANALYSIS/SYNTHESIS METHOD

Seven prototype musical instrument sounds were selected for analysis using a computer-based phase-vocoder method (Beauchamp, 1993). This phase vocoder is different than most in that it allows tuning of the fixed analysis frequency (f_a) to coincide with the estimated fundamental frequency of the input signal. The analysis method yields the frequency deviations between harmonics of this analysis frequency and the corresponding frequency components of the input signal, which are assumed to be at least approximately harmonic relative to its fundamental.

A. Signal representation

For each sound, an analysis frequency was chosen that minimized the average of the harmonic frequency deviations. Thus, a time-varying representation was achieved for each sound according to the formula

$$s(t)=\sum_{k=1}^{K} A_k(t)\cos\left(2\pi\int_0^t (kf_a+\Delta f_k(t)dt)+\theta_k(0)\right), \tag{4}$$

where

$s(t)=$ sound signal,
$t=$ time in s,
$k=$ harmonic number,
$K=$ number of harmonics,
$A_k(t)$ is the amplitude of the kth harmonic at time t,
$f_a=$ analysis frequency,
$\Delta f_k(t)$ is the kth harmonic's frequency deviation, such that $kf_a+\Delta f_k(t)$ is the exact frequency of the kth harmonic, and
$\theta_k(0)$ is the initial phase of the kth harmonic.

The parameters used for synthesis that were simplified in this study are $A_k(t)$ and $\Delta f_k(t)$. No attempt was made to simplify $\theta_k(0)$. Although $A_k(t)$ and $\Delta f_k(t)$ were only stored as samples occurring every $1/(2f_a)$ s, the signal was approximated with reasonable accuracy at a much higher resolution (sample frequency 22 050 or 44 100 Hz) by using linear interpolation between these values. Synthesis was accomplished by additive (or Fourier) synthesis of the harmonic sine waves.

B. Prototype sounds

Sounds of the instruments clarinet, flute, harpsichord, marimba, oboe, trumpet, and violin were selected in order to have one representative from each of several families of instruments whose tones are at least approximately harmonic. Five of the sounds were taken from the McGill University Master Samples recordings, one from Prosonus (oboe), and one (trumpet) had been recorded at the UIUC School of Music. An attempt was made to select sounds that were of high quality, that represented the instruments well, and that had fundamental frequencies close to 311.1 Hz (E-flat 4), a note within the normal playing range of these instruments.[1] Since synthesis was accomplished by an additive method based on Eq. (1), it was easy to alter the stimuli's fundamental frequencies (f_a) to be exactly 311.1 Hz. Table I gives some basic characteristics of the prototype sound signals.

C. Analysis method

The phase vocoder method used for analysis consists of the following steps:

TABLE I. Data for the seven instrument sounds used in the study. For McGill source recordings, the numbers indicate volume:track-index. For Prosonus recordings, they indicate woodwinds volume:band-index. Attack (t_1) is the time in the original sound at which the attack was estimated to end. Decay (t_2) is the time in the original sound at which the decay was estimated to begin. The marimba and harpsichord, being impulsively excited instruments, have no sustain portions. The marimba, being shorter than the target 2-s duration, was stretched rather than shortened, and so the attack and decay values were not used.

	Source of original recording	Original fundamental frequency (Hz)	Original duration of sound, t_L (s)	Number of harmonics used in analysis, K	Attack, t_1 (s)	Decay, t_2 (s)
Clarinet (Cl)	McGill (2:10-14)	311.4	3.81	70	0.05	3.50
Flute (Fl)	McGill (9:86-04)	311.0	2.31	70	0.25	2.10
Harpsichord (Hc)	McGill (11:95-06)	311.1	2.97	70	0.04	2.97
Marimba (Mb)	McGill (3:04-23)	312.2	1.83	70	...	...
Oboe (Ob)	Prosonus (W1:04-04)	311.8	2.98	30	0.15	2.20
Trumpet (Tp)	UIUC	350.0	2.43	31	0.32	1.30
Violin (Vn)	McGill (9:63-03)	311.1	4.94	66	0.22	4.10

(1) Band-limited interpolation of the input signal to produce a power-of-two number of samples per analysis period ($1/f_a$), which is the lowest possible to exceed the number of original samples in this time interval.
(2) Segmentation of the input signal into contiguous frames whose lengths are equal to twice the analysis period ($2/f_a$) and which overlap by half an analysis period ($f_a/2$).
(3) Multiplication of each signal frame by a Hamming window function whose length is two analysis periods ($2/f_a$).
(4) Fast Fourier transform (FFT) of the resulting product to produce real and imaginary components at frequencies 0, $f_a/2$, f_a, $3f_a/2$,..., $f_s/2-f_a$, where f_s is the sampling frequency. Components which are not positive integer multiples of f_a are discarded.
(5) Right-triangle solution of each retained real and imaginary part to give the amplitude and phase of each harmonic.
(6) Computation of the frequency deviation for each harmonic by a trigonometric identity which essentially gives the difference in phase between frames for each harmonic.
(7) Storage of the harmonic-and frequency-deviation data in an ''analysis file.'' The number of harmonics stored is less than $f_s/(2f_a)$. The analysis file for each sound is the basis for further sound processing.

Further details of this procedure are discussed by Beauchamp (1993).

The analysis system may be viewed as a set of contiguous bandpass filters which have identical bandwidths (f_a) and are centered at the harmonics of the analysis frequency (f_a). The basic assumption is that the signal consists of harmonic sine waves which line up with the filters such that each filter outputs one of the sine waves. The analysis gives the amplitude and frequency of each sine wave. When the sine waves are summed, the signal is almost perfectly reconstructed. In fact, the sine-wave sum can be viewed as that created by processing the input signal by the sum of the bandpass-filter characteristics. It can be shown that this sum is flat within ± 1 dB over the range $[f_a/2, f_s/2]$. Figure 2 shows a block diagram of the basic analysis/ synthesis system and Fig. 3 shows a typical set of amplitude and frequency data.

D. Types of simplification

Spectral simplifications were performed on the analysis data, after which the sounds were synthesized by the additive method. In order that sound duration would not be a factor in the study, most of the sounds were shortened to a 2-s duration by resampling the analysis data. Rather than resampling at a uniform rate, the sounds were resampled to preserve their attack and decay portions and shorten their interior portions while retaining their microstructural variations in amplitude and frequency. This was done by first observing the sound's rms amplitude given by

$$A_{\mathrm{rms}}(t) = \sqrt{\sum_{k=1}^{K} A_k^2(t)}, \tag{5}$$

and then identifying by eye the time intervals corresponding to the attack and decay as $(0,t_1)$ and (t_2,t_L) (see Table I for chosen values of t_1 and t_2), where t_L is the original sound duration. The marimba was an exception to this procedure, since its original duration was 1.83 s. The data for this instrument were simply stretched to obtain a duration of 2 s,

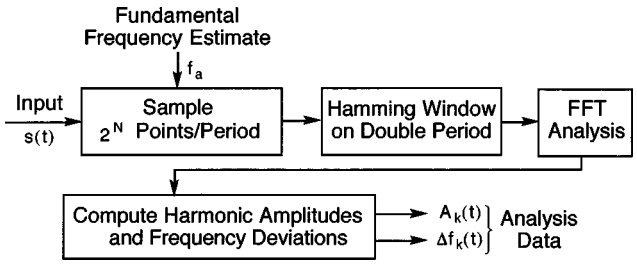

FIG. 2. Method for time-varying spectral analysis that yields the amplitude and frequency deviations for each harmonic k. The exact frequency for harmonic k is given by $f_k = k\, f_a + \Delta f_k(t)$, where f_a is the analysis fundamental frequency.

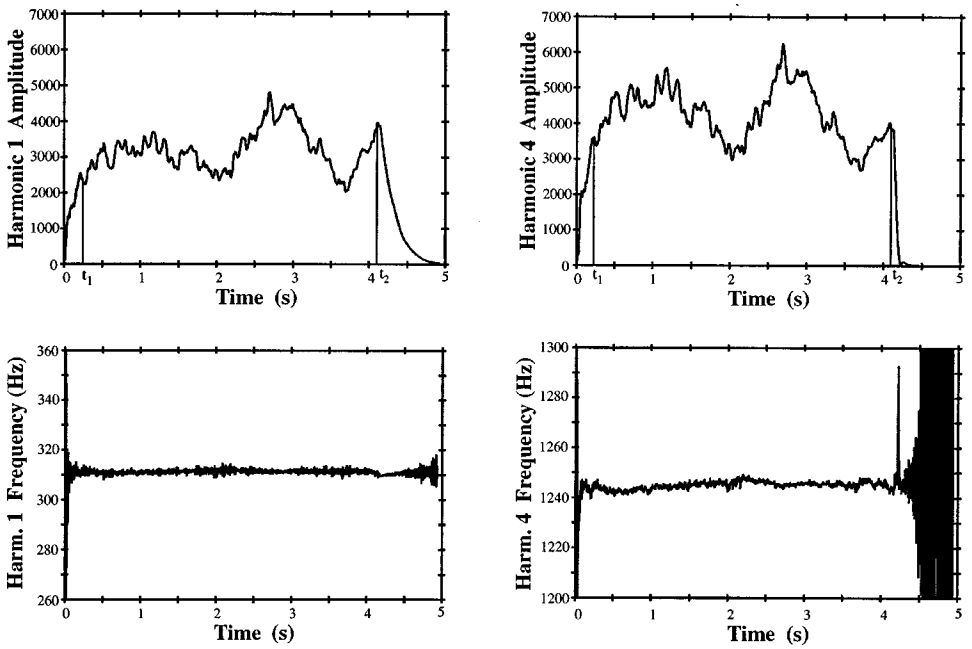

FIG. 3. Example spectral-analysis data for original violin tone (left column: first harmonic; right column: fourth harmonic; upper row: amplitude envelopes; lower row: frequency envelopes). Note the difficulty in reliably estimating the frequency of harmonic 4 when its amplitude approaches zero. Attack (t_1) and decay (t_2) boundaries are indicated.

and no notable degradation of the musical quality of the original was noted by the authors.

Second, for each harmonic amplitude and frequency deviation, the time intervals (t_1, t_1+2) and (t_2-2, t_2) were cross-faded using a cubic function to minimize any discontinuities. Thus, between the times t_1 and t_2, the sound was transformed from what it sounded like in the region of time t_1 to what it sounded like in the region of time t_2 over a period of $2-(t_1+t_L-t_2)$ s. This gave each sound a total duration of 2 s. In order for this method to work properly, we assumed that each sound had a microstructure which was statistically uniform over the interval (t_1, t_2). Since the sounds selected had no vibrato, this assumption seemed to be valid, and the resulting synthesized sounds were judged by the authors to be free of artifacts. Details on the duration-shortening algorithm are given in Appendix A. Figure 4 shows a set of data corresponding to Fig. 3 after application of the duration-shortening algorithm. Note that t_1 and t_2 are indicated in Fig. 3.

Finally, the seven duration-equalized prototype sounds were compared, and amplitude multipliers were determined such that the sounds were judged by the authors to have equal loudness. When the sounds were synthesized with the shortened duration, the amplitude multipliers, and a synthesis fundamental frequency of 311.1 Hz, they were judged to be equal in loudness, pitch, and duration. (It should be mentioned, however, that this equalization was not central for the present study, since each discrimination pair was always derived from a single prototype sound.) The equalized sounds then served as the reference sounds for this study, and their corresponding data sets are henceforth referred to as the analysis data.

Six primary simplifications of the analysis data were performed prior to synthesis. Each of these simplifications constitutes a major reduction in the amount of data used for synthesis.

1. Amplitude-envelope smoothness (AS)

The objective of this operation was to remove microvariations or noise in the harmonic amplitude over time, as these had been shown to be perceptually salient in previous work by Charbonneau (1981). These envelopes $A_k(t)$ were processed by a second-order recursive digital filter having a Butterworth response and a smoothing cutoff frequency of 10 Hz. This essentially removed all microdetail in the amplitude envelopes. However, we did not smooth the attack portions of the envelopes $(0 \leqslant t \leqslant t_1)$ since we only wished to determine the importance of microdetail in the amplitude envelopes thereafter. Smoothing the attack portions would have slowed the attacks, unintentionally affecting discrimination of the simplified sounds from their corresponding reference sounds. In order to avoid discontinuity, the attack portion of each amplitude envelope was cross-faded into the subsequent smoothed portion over a few frame points corresponding to the delay of the filter. In this way, the attack portions were essentially unaltered by the smoothing operation (see Table I for t_1 values).

2. Amplitude-envelope coherence (AC) (spectral envelope fixing)

The objective was to test the effect of eliminating *spectral flux* (defined as the change in shape of a spectral envelope over time) without changing the rms amplitude envelope or the average spectrum. Spectral flux has been found to be an important perceptual dimension of timbre (Grey, 1977; Krumhansl, 1989; Krimphoff *et al.*, 1994). To eliminate spectral flux, the amplitude envelope $A_k(t)$ for each har-

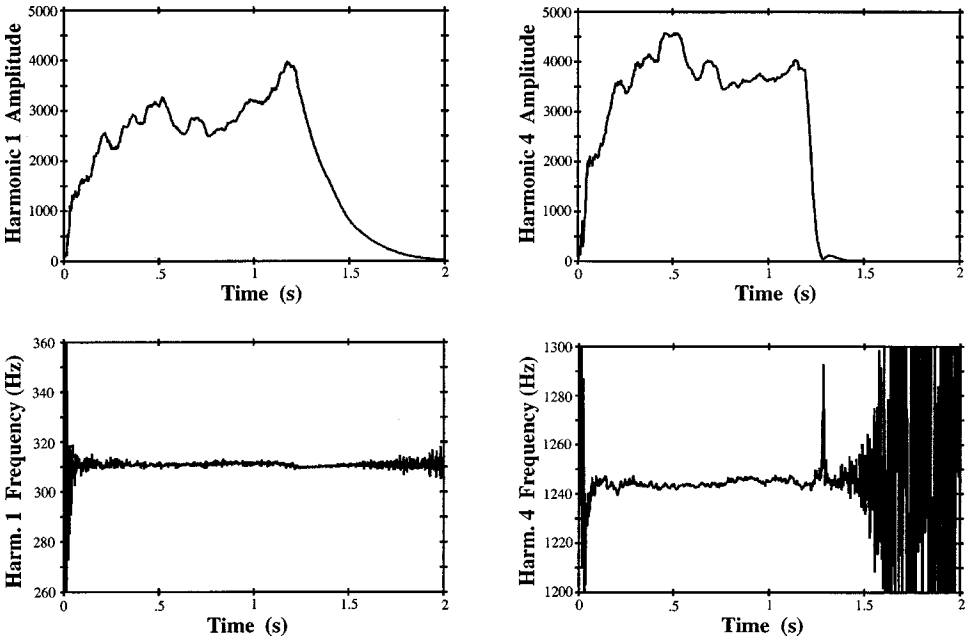

FIG. 4. Example spectral-analysis data for violin tone with duration reduced to 2 s (left column: first harmonic; right column: fourth harmonic; upper row: amplitude envelopes; lower row: frequency envelopes).

monic k was replaced by a function which was proportional to the rms envelope and the average amplitude of the harmonic. Thus, the harmonic-amplitude ratios $[A_2(t)/A_1(t)$, etc.] were fixed during the course of the sound. In addition, the amplitudes were scaled in order to preserve the rms envelope under this transformation. The formula for this transformation is:

$$A_k(t) \leftarrow \frac{\overline{A_k} A_{\mathrm{rms}}(t)}{\sqrt{\Sigma_{k=1}^{K} \overline{A_k}^2}}, \tag{6}$$

where $\overline{A_k}$ signifies the time average of the kth harmonic amplitude over the sound's duration and $\leftarrow$ signifies the replacement operation. Note that with this transformation, all amplitude envelopes of all harmonics have the same shape, albeit with different scale factors.

3. Spectral envelope smoothness (SS)

The question to be answered here is whether jaggedness or irregularity in the shape of a spectrum is perceptually important. For example, the clarinet has a characteristically ''jagged'' up-and-down spectral envelope due to weak energy in the low-order, even harmonics. A smoothing of this spectral envelope would give it more of a low-pass form. Spectral-envelope smoothness was found by Krimphoff *et al.* (1994) to correspond to the third dimension of Krumhansl's (1989) 3D space. To test this, the time-varying spectra were smoothed with respect to frequency. To accomplish this, at each time frame each harmonic amplitude was replaced by the average of itself and its two neighbors (except for endpoint harmonics number 1 and K, where averages of themselves and their neighbors were used)

$$A_1(t) \leftarrow \frac{A_1(t)+A_2(t)}{2}, \tag{7a}$$

$$A_k(t) \leftarrow \frac{A_{k-1}(t)+A_k(t)+A_{k+1}(t)}{3}, \quad k \notin \{1,K\}, \tag{7b}$$

$$A_K(t) \leftarrow \frac{A_{K-1}(t)+A_K(t)}{2}. \tag{7c}$$

This smoothing algorithm is not unique and may not be optimal, but it is perhaps the simplest one can imagine. According to this algorithm, the smoothest possible spectrum is one that follows a straight-line curve (i.e., $A_k = a + b \cdot k$), since such a spectral envelope would not be altered by this transformation.

Figure 5 compares the time-varying amplitude spectrum of a reference sound with those obtained after increasing amplitude-envelope smoothness, amplitude-envelope coherence, and spectral-envelope smoothness algorithms have been applied. The effect of these operations on the reference time-varying spectrum is readily apparent.

4. Frequency envelope smoothness (FS)

We wished to test the auditory importance of frequency microvariations in a parallel fashion to that of amplitude microvariations. Therefore, the envelopes $\Delta f_k(t)$ were processed similarly to the $A_k(t)$ envelopes in amplitude-envelope smoothing described above, except that smoothing was done over the entire sound's duration, including the attack phase. This operation did not grossly affect the frequency variation during the attack, as amplitude-envelope smoothing would have affected amplitude variation during that period had it included the attack.

5. Frequency envelope coherence (FC) (harmonic frequency tracking)

Here, we wanted to test the discriminability of inharmonicity among a sound's partials, even if it sometimes occurs

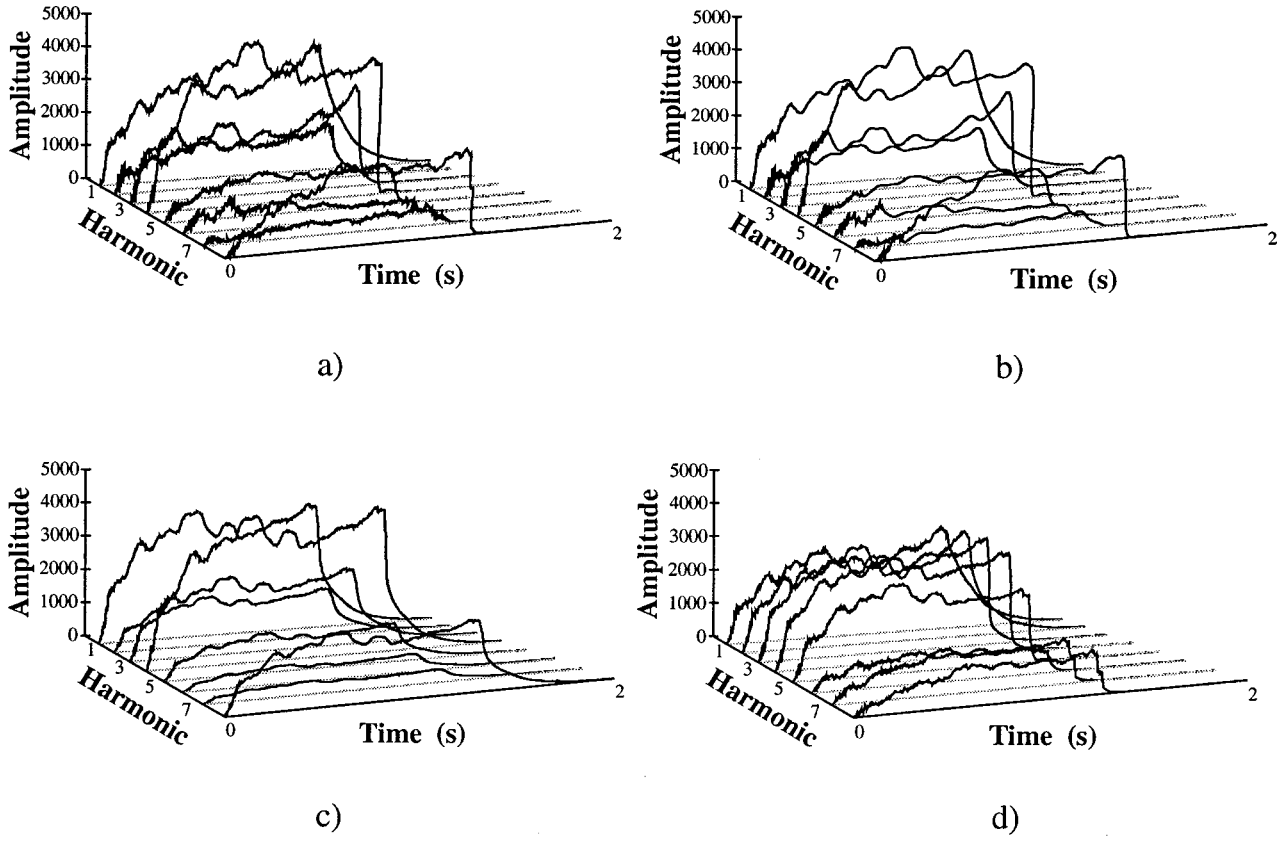

FIG. 5. Simplifications of amplitude envelopes for harmonics 1 to 8: (a) full violin-tone analysis data (reference sound) (b) after amplitude-envelope smoothing, (c) after rms envelope substitution (amplitude-envelope coherence), (d) after spectral-envelope smoothing.

only momentarily. Analogously to the amplitude-envelope coherence case, all frequency envelopes over time are tied together in a perfect harmonic relation. First, an average temporal-frequency contour was computed on the envelope for the first five harmonics, and then the individual harmonic contours were set equal to this contour multiplied by their respective harmonic numbers.

$$f_K(t) \leftarrow k\overline{f_0(t)}, \tag{8}$$

where $\overline{f_0(t)}$ is defined by

$$\overline{f_0(t)} = \frac{\Sigma_{k=1}^{5} A_k(t)\,(1/k)\,f_k(t)}{\Sigma_{k=1}^{5} A_k(t)}. \tag{9}$$

With this method, the strongest harmonics among the first five receive the ''highest votes'' for determining the average fundamental frequency of the sound. The measured frequency of the first harmonic could have been used instead of $\overline{f_0}$. However, it is possible that the first harmonic may be weak in amplitude, which with phase-vocoder analysis would result in a poorly defined frequency envelope (Moorer, 1978). This method obviates that problem.

6. Frequency envelope flatness (FF)

This simplification tested listeners' abilities to discriminate the combination of no frequency variations and no inharmonicity, as after this operation is performed, neither are present in the synthesized sounds. Indeed, there is no frequency envelope, as each harmonic's frequency is set equal to the product of its harmonic number (k) and the fixed analysis frequency (f_a). This operation had previously been found to have an effect on discrimination by Grey and Moorer (1977) and Charbonneau (1981).

Figure 6 shows a reference set of harmonic-frequency envelopes in comparison to those which have been simplified by frequency-envelope smoothing, frequency-envelope coherence, and frequency-envelope flattening.

Each simplification is accompanied by a certain amount of data reduction. Formulas for data reduction are given in Appendix B.

II. EXPERIMENTAL METHOD

A. Subjects

The 20 subjects were aged 19 to 35 years and reported no hearing problems. They included ten musicians (six males, four females) and ten nonmusicians (four males, six females). Musicians were defined as being professionals, semiprofessionals, or having at least 6 years of practice on an instrument and playing it daily. Nonmusicians were defined as having practiced an instrument for not more than 2 to 3 years in their childhood or adolescence, and no longer playing. The subjects were paid for their participation with the exception of three who were members of the auditory-perception team at IRCAM.

B. Stimuli

The seven instruments chosen belong to the air column (air reed, single reed, lip reed, double reed), string (bowed,

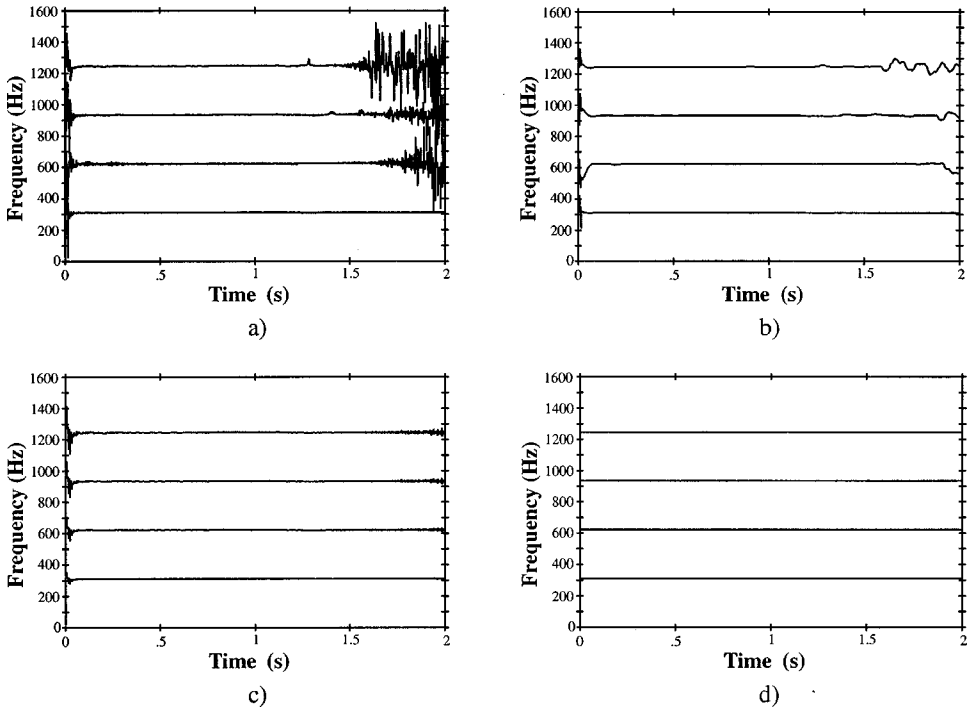

FIG. 6. Simplifications of frequency envelopes for harmonics 1 to 4: (a) full violin-tone analysis data (reference sound), (b) after frequency-envelope smoothing, (c) after average frequency-envelope substitution (frequency-envelope coherence), (d) after replacement by fixed harmonics (frequency-envelope flattening).

plucked), and bar (struck) families: clarinet, flute, harpsichord, marimba, oboe, trumpet, and violin. Each was analyzed and synthesized with the reference sound-analysis data (before modification). In no case could the original recorded sound be discriminated from the full synthesis when presented in an AA–AB discrimination paradigm at better than 64%.[2] The sounds were stored in 16-bit integer format on hard disk. All ''reference'' sounds (full synthesis) were equalized for fundamental frequency (311.1 Hz or E-flat 4) and for duration (2 s) (see Sec. I D for a description of the technique for equalizing duration in synthesis). They were also equalized for loudness in an adjustment procedure by the authors. The different kinds of simplifications and their combinations that were applied to the stimuli are illustrated graphically in Fig. 7. Six simplifications concerned a single parameter, three concerned two parameters, and one each concerned three and four parameters.[3] The 11 simplified sounds for each instrument were synthesized with the method described above on a NeXT computer. They were equalized for loudness within each instrument in an adjustment procedure by the authors.

C. Procedure

A two-alternative forced-choice (2AFC) discrimination paradigm was used. The listener heard two pairs of sounds (AA–AB) and had to decide if the first or second pair contained two different sounds. The dependent variable was the d' measure of sensitivity to the physical difference derived from signal-detection theory using a 2AFC model (Green and Swets, 1974; Macmillan and Creelman, 1991). The trial structure could be one of AA–AB, AB–AA, BB–BA, or BA–BB, where A represents the reference sound and B one of the 11 simplifications. This paradigm has the advantage of presenting to the listener both a ''same'' pair and a ''different'' pair between which the different one must be detected. All four combinations were presented for each simplification and for each instrument. The two 2-s sounds of each pair were separated by a 500-ms silence, and the two pairs were separated by a 1-s silence. On each trial, a button labeled (in French) ''The first pair was different: key 1'' appeared on the left of the computer screen and a button labeled ''The second pair was different: key 2'' appeared on the right. The computer would not accept a response until all four sounds in a trial had been played. This was indicated by a dimming of the labels on the buttons during sound presentation.

For each instrument, a block of 44 trials was presented to the subjects (four trial structures$\times$11 simplifications). Each block was presented twice in succession, and performance for each simplification was computed on eight trials for each subject. Seven pairs of blocks were presented corresponding to the seven instruments. The total duration of the experiment was about two h and 20 min. For 13 subjects, the experiment was divided into two sessions performed on different days, with four instruments on one day and three instruments on the other. For seven other subjects, it was performed in one day with several pauses between instruments.

The experiment was controlled by the PSIEXP interactive program (Smith, 1995) running on a NeXT computer. Subjects were seated in a Soluna S1 double-walled sound-isolation booth facing a window through which the computer screen was visible. Sounds were converted through NeXT digital-to-analog converters, amplified through a Canford power amplifier, and then presented through AKG K1000

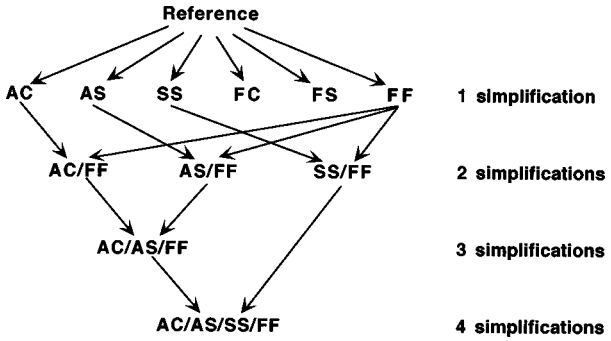

FIG. 7. Schema illustrating the accumulation of stimulus simplifications. For key, see Table II.

open-air headphones at a level of approximately 70 dB SPL as measured with a Bruel & Kajer 2209 sound-level meter fitted with a flat-plate coupler.

At the beginning of the experiment, the subject read instructions and asked any necessary questions of the experimenter. Five or six practice trials (chosen at random from the instrument being tested) were presented in the presence of the experimenter before the first block for each instrument. Then the two experimental blocks for that instrument were presented. The order of presentation of the 44 trials was random within each block, and the order of presentation of the instruments was randomized for each subject.

III. RESULTS

Discrimination rates were computed for each simplification of each instrument's reference sound across the four trial structures and two repetitions for each subject. The means across both groups of subjects for the 11 simplifications on seven instrument sounds are given in Table II and plotted in Fig. 8. Accumulated simplifications involving amplitude-envelope coherence (AC), amplitude-envelope smoothness (AS), and spectral-envelope smoothness (SS) are joined by lines to visualize the effect of accumulation. In general, spectral-envelope smoothness and amplitude-envelope coherence simplifications were the most easily discriminated, followed by coherence (FC) and flatness (FF) of frequency envelopes, and finally amplitude-(AS) and frequency-(FS) envelope smoothness. With one exception, the accumulation of simplifications improved discrimination, attaining nearly perfect discrimination for all instruments. The pattern of discrimination differences across simplification types is very different for each instrument, suggesting that the acoustic structure of each sound is affected differentially by these simplifications.

To evaluate the different factors included in this experiment, several statistical analyses were performed. The dependent variable in these analyses was the d' index of sensitivity (derived from proportion-correct discrimination rates in Table A.5.2 from Macmillan and Creelman, 1991). A global mixed analysis of variance (ANOVA) was performed on between-subjects factor musical training (2) and within-subjects factors instrument (7) and simplification (11). Mixed ANOVAs on musical training and simplification were also performed for the data of each instrument individually.

TABLE II. Results of discriminating six basic simplifications and five combinations of simplifications compared to the reference sounds (complete resynthesis of the originals after frequency, duration, and loudness matching). Key: AC=amplitude-envelope coherence, AS=amplitude-envelope smoothness, SS=spectral-envelope smoothness, FC=frequency-envelope coherence, FS=frequency-envelope smoothness, FF=frequency-envelope flatness, Cl=clarinet, Fl=flute, Hc=harpsichord, Mb=marimba, Ob=oboe, Tp=trumpet, Vn=violin.

	Instrument							
Simplification	Cl	Fl	Hc	Mb	Ob	Tp	Vn	Mean
AC	0.81	0.96	0.97	0.97	0.75	0.98	0.95	0.91
AS	0.56	0.80	0.79	0.59	0.54	0.73	0.59	0.66
SS	0.98	0.97	0.96	0.99	0.99	0.82	0.99	0.96
FC	0.69	0.72	0.93	0.50	0.53	0.77	0.70	0.69
FS	0.56	0.59	0.84	0.67	0.72	0.81	0.69	0.70
FF	0.70	0.72	0.91	0.62	0.48	0.82	0.73	0.71
AC/FF	0.86	0.98	0.97	0.96	0.94	0.99	0.99	0.95
AS/FF	0.69	0.94	0.92	0.65	0.81	0.86	0.71	0.80
SS/FF	1.00	0.99	0.97	0.98	0.98	0.89	0.98	0.97
AC/AS/FF	0.87	0.98	0.98	0.97	0.86	1.00	0.98	0.95
AC/AS/SS/FF	0.99	0.98	0.98	0.98	0.99	0.97	1.00	0.99

For the data within each instrument, Tukey–Kramer HSDs (''honestly significant differences'') were computed to determine the critical difference between condition means at a significance level of 0.05. This technique allows a robust comparison among all means of a data set by the simultaneous construction of confidence intervals for all pairs (Ott, 1993). Finally, in order to determine which simplifications were reliably different from chance performance, single-sample *t*-tests were performed against a hypothetical mean of 0.50 with probabilities being corrected for multiple tests with the Bonferroni adjustment.

A. Effects of musical training

Musicians discriminated simplifications from reference sounds slightly better overall than nonmusicians (86.8% vs 82.2%) by 3.0% to 7.1% across instruments [$F(1,18) = 8.05$, $p < 0.05$].[4] There was no interaction of this factor with other factors in the global analysis. In the individual ANOVAs, there were significant main effects of musical training for four of the seven instruments [flute: $F(1,18) = 5.01$, $p < 0.05$; marimba: $F(1,18) = 9.76$, $p < 0.01$; oboe: $F(1,18) = 6.99$, $p < 0.05$; violin: $F(1,18) = 5.70$, $p < 0.05$], and there were significant musical training by simplification interactions for two instruments [clarinet: $F(10,180) = 2.93$, $p < 0.05$; violin: $F(10,180) = 2.55$, $p < 0.05$]. So overall, there was a small effect of musical training that was globally reliable and present in the majority of instruments but which varied differently across simplification conditions in only two of the instruments. Given the small size of the effect, we will not consider it any further.

B. Effects of instrument

In the global ANOVA, there were highly significant effects of instrument [$F(6,108) = 28.80$, $p < 0.0001$], simplification [$F(10,180) = 237.97$, $p < 0.0001$], and their interaction [$F(60,1080) = 9.87$, $p < 0.0001$]. This strong interaction revealed very large differences in the effects of a given sim-

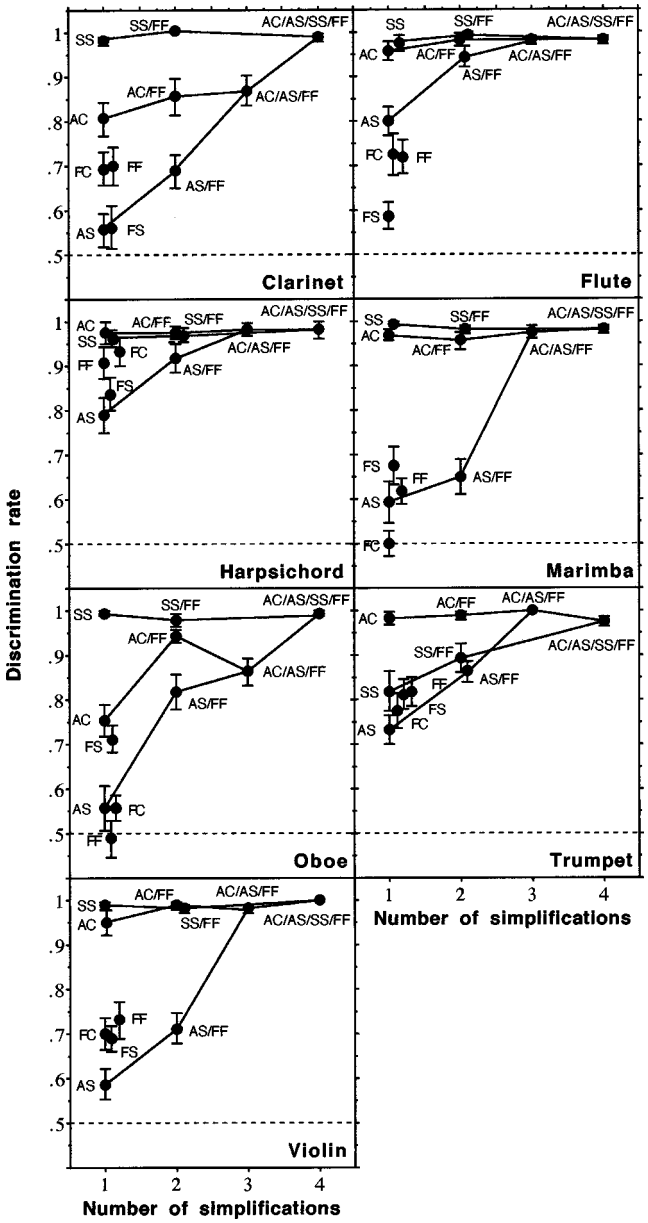

FIG. 8. Discrimination rates as a function of the number of simplifications performed on sounds from seven instruments. The letter codes refer to the simplification types (see Table II caption). Simplifications involving AS, AC, and SS are connected to visualize the effect of their accumulation. The vertical bars represent ±1 standard error of the mean. Chance performance is at 0.5. Some points have been displaced laterally for visibility.

plification across instruments. We will therefore only consider differences among simplifications within the individual ANOVAs for each instrument.

C. Effects of the simplifications and their accumulation

The main effect of simplification was highly significant ($p<0.0001$) for all seven instruments [clarinet: $F(10,180)=40.14$; flute: $F(10,180)=41.14$; harpsichord: $F(10,180)=11.54$; marimba: $F(10,180)=71.82$; oboe: $F(10,180)=43.40$; trumpet: $F(10,180)=22.05$; violin: $F(10,180)=81.65$], indicating a large variation in discriminability of the different types of simplification. Single-sample *t*-tests adjusted for multiple tests indicated that only nine of the 42 single simplifications were *not* discriminated above chance. These include AS and FS for the clarinet, FS for the flute, AS and AC for the marimba, AS, FC, and FF for the oboe, and AS for the violin. Note that no single simplification is ''successful'' (i.e., indistinguishable from the reference sound) for all seven instruments. However, amplitude-envelope smoothness was only reliably discriminated from the reference in flute, harpsichord, and trumpet. In order to evaluate the significance of the differences among simplifications, a clustering organization is projected onto the mean data in Fig. 9 in which means that are smaller than the critical Tukey–Kramer HSD for that instrument are enclosed in a bounded region. The critical differences are listed in Table III. In general, simplifications involving amplitude-envelope coherence (AC) and spectral-envelope smoothness (SS) are found in the highest cluster, showing near-perfect discrimination for most instruments, although AC is less well discriminated in the clarinet and oboe, and SS is less well discriminated in the trumpet.

As a general rule, the discrimination of a multiple simplification was roughly equal to the discrimination of the constituent simplification which had the highest discrimination rate. For example, take the clarinet sound. Discrimination was near chance for AS, around 70% for FF, about 80% for AC, and nearly perfect for SS. Accumulating AS and FF gave a rate no different from that for FF. Similarly, AC/FF and AC/AS/FF had rates no different from that of AC, while SS/FF and AC/AS/SS/FF were not different from SS alone. This rule held for 32 of the 35 multiple-simplification conditions. Thus, there were only three cases where the accumulation of two simplifications was better discriminated than either of the constituent simplifications: AS/FF was better than AS and FF for the flute, and AC/FF and AS/FF were better than their constituents for the oboe. There was only one case where an accumulated simplification resulted in a *decrease* in discrimination performance: AC/AS/FF was discriminated worse than AC/FF for the oboe, suggesting that the addition of the amplitude-envelope smoothness reduced the effect of amplitude-envelope coherence and/or frequency-envelope flatness. Taken together, these results suggest that it is generally sufficient to examine the individual effects of a single, ''most-potent'' simplification for each instrument to explain the behavior of their combinations. In order to compare across instruments, the discrimination rates for the six single simplifications are shown for each instrument in Fig. 10.

IV. MEASUREMENTS OF SPECTRAL DIFFERENCES BETWEEN REFERENCE AND SIMPLIFIED SOUNDS

A. Amplitude and frequency errors

The effect of the simplifications on sounds was directly measured from the analysis file data by computing normalized rms differences between reference and simplified sounds. Accordingly, for the amplitude simplifications, we measured the relative difference between reference (*Ar*) and

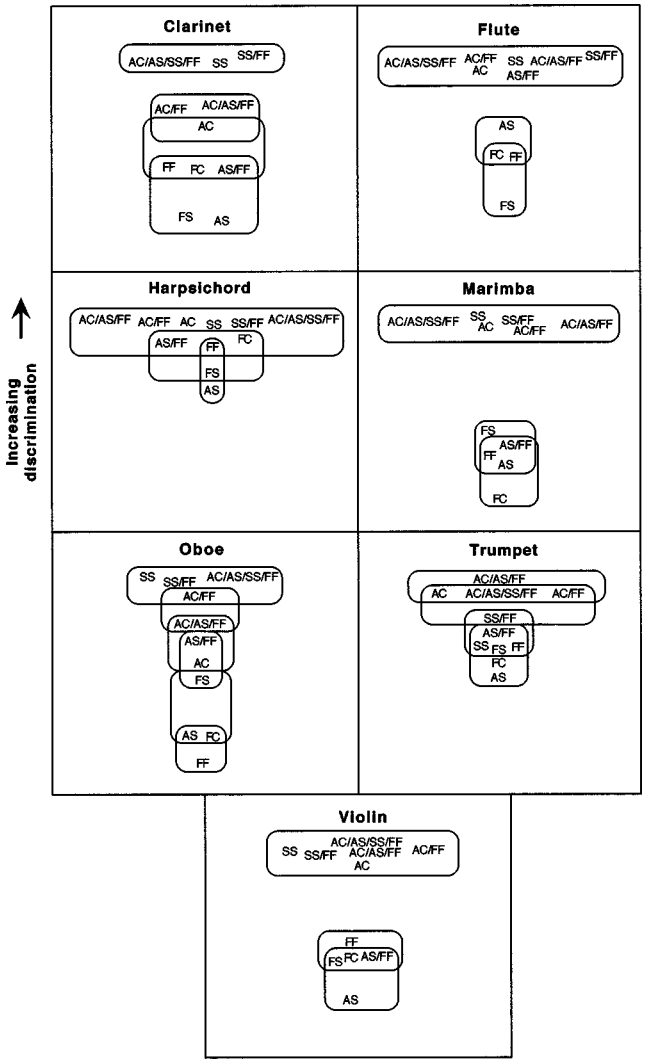

FIG. 9. Schematic representation of significant differences between means as revealed by Tukey–Kramer HSD tests. Discrimination performance is organized along the vertical dimension within each panel, as in Fig. 8. Simplifications with means whose differences are not bigger than the critical difference (see Table III) are enclosed within a bounded region. In the oboe data, for example, FF is not significantly different from AS and FC but is different from FS. However, AS and FC are not significantly different from FS.

simplified (*As*) time-varying amplitude spectra (which are assumed to represent sounds having the same mean frequencies and same duration) using

$$\mathrm{ERR}_{\mathrm{amp}}=\frac{1}{I}\sum_{i=1}^{I}\left(\frac{\Sigma_{k=1}^{K}(As_k(i)-Ar_k(i))^2}{\Sigma_{k=1}^{K}As_k(i)\cdot Ar_k(i)}\right)^{1/2}, \tag{10}$$

TABLE III. Critical Tukey–Kramer differences for the mean discriminations of simplifications computed across both groups of subjects.

Instrument	Critical difference
Clarinet	0.217
Flute	0.196
Harpsichord	0.226
Marimba	0.187
Oboe	0.207
Trumpet	0.215
Violin	0.170

where i is the number of the analysis time frame and I is the total number of frames. $\mathrm{ERR}_{\mathrm{amp}}$ can vary between 0 and 1. In our set of sounds, it varied between about 0.01 and 0.58. With this formula, the error at any instant relative to the amplitude at that instant is computed. Due to the amplitude product in the denominator, Eq. (10) accentuates low-amplitude portions, giving them the same weight as high-amplitude portions. It is assumed here that proportional-amplitude errors are more relevant than absolute-amplitude errors. The normalized squared errors are accumulated over harmonics and are then averaged over time. One could argue that this might be improved by first accumulating amplitudes by critical bands before averaging, but this would complicate the calculation considerably and would not guarantee any improved result.

In a similar manner, for the frequency simplifications, we measured the difference between reference (*fr*) and simplified (*fs*) series of time-varying frequency data using

$$\mathrm{ERR}_{\mathrm{freq}}=\frac{1}{If_a}\sum_{i=1}^{I}\left(\frac{\sum_{k=1}^{K}\left(\frac{(Ar_k(i)(fs_k(i)-fr_k(i))}{k}\right)^2}{\Sigma_{k=1}^{K}Ar_k^2(i)}\right)^{1/2}. \tag{11}$$

Frequency differences are divided by the harmonic number k, because we assume that they are intrinsically amplified linearly with k. The frequency difference for each harmonic k is weighted by its amplitude, giving greater votes to higher-amplitude harmonics. This is beneficial because lower-amplitude harmonics tend to have more oscillation in their frequency data, which is an artifact of the analysis process and not representative of the sound itself (Moorer, 1978). Besides averaging over time, we normalize by the average fundamental frequency (f_a), so that the results are presented as a proportion of the fundamental. The values of $\mathrm{ERR}_{\mathrm{freq}}$ in our set of sounds were very low (between 0.0009 and 0.0134).

The amplitude and frequency-error results for the six basic simplifications for the seven instruments are shown in Tables IV and V, respectively. The mean d' scores are plotted in Fig. 11 as a function of the logarithm of the error values for the amplitude (a) and frequency (b) simplifications. Although there is some dispersion in the plot, the overall relations between listener-obtained discrimination scores and the objective measurements are clear. For most cases, larger errors predict higher sensitivity. If discrimination scores are expressed in terms of d', $\log(\mathrm{ERR}_{\mathrm{amp}})$ explains 77% of the variance in discrimination performance for single-amplitude simplifications. The amount of variance explained increases to 88% if the outlying point due to the AC condition for the marimba is removed. Note that the various amplitude simplifications are quite different overall in their discriminability (AS<AC<SS).

The picture is quite different for the frequency simplifications. First, the data are much more scattered, indicating that $\mathrm{ERR}_{\mathrm{freq}}$ explains much less variance than did $\mathrm{ERR}_{\mathrm{amp}}$ for corresponding conditions; explained variance in d' by $\log(\mathrm{ERR}_{\mathrm{freq}})$ is only 34% but increases dramatically to 57% when the outlying point due to the FF condition for the oboe

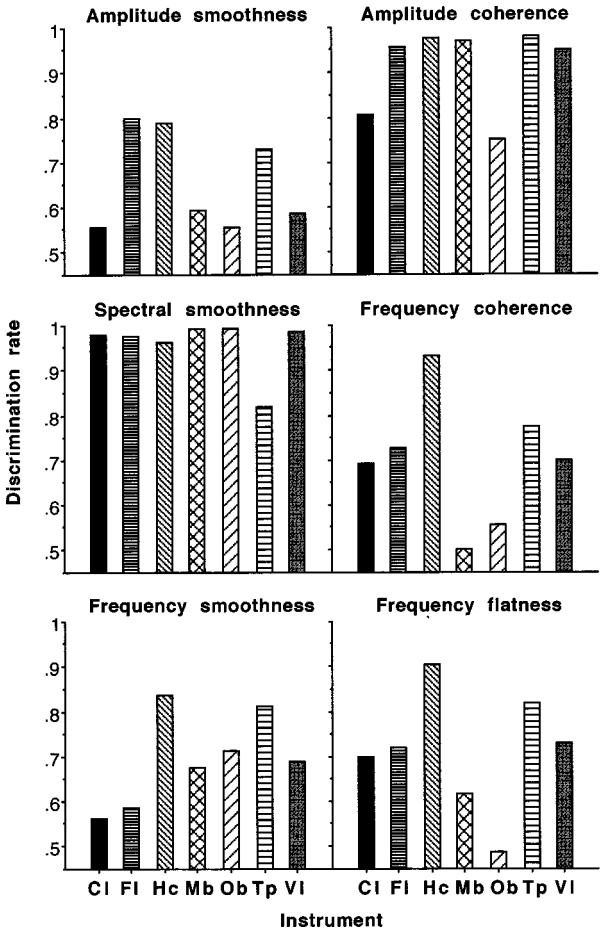

FIG. 10. Discrimination rates for the seven different instrument sounds having been simplified in six ways (see the text for a complete description). For instrument key, see Table II.

is removed. Second, there is a much greater overlap between the conditions indicating that there is a less systematic effect of the simplification condition and that each simplification type affects the various instruments to very different degrees.

B. Effect of spectral-amplitude changes on centroid

Since the centroid of the spectrum has been shown to be strongly correlated with one of the most prominent dimensions of multidimensional-scaling representations of timbral differences (Grey and Gordon, 1978; Iverson and Krumhansl, 1993; Kendall and Carterette, 1996; Krimphoff *et al.*, 1994; Krumhansl, 1989; Wessel, 1979), one might conjecture that a listener's ability to detect spectral-amplitude modifications is due to detection of attendant centroid changes rather than to the modifications themselves. Although in synthesized tones spectral centroid can be controlled independently of other spectral-amplitude modifications, they are not necessarily separable in musical instrument tones. Nonetheless, we have found them to be statistically independent to a substantial degree in a number of our stimuli.

We define time-varying normalized spectral centroid (SC) to be

TABLE IV. Relative spectral differences between reference and simplified spectra for basic and accumulated amplitude simplifications. The values represent ERR_{amp} [Eq. (10)]. Note that the values for basic simplifications and those simplifications accumulated with the FF simplification would be identical, since the FF operation has no effect on the amplitudes. For key, see Table II caption.

	Instrument						
Simplification	Cl	Fl	Hc	Mb	Ob	Tp	Vn
AC	0.100	0.164	0.204	0.033	0.122	0.280	0.350
AS	0.017	0.024	0.035	0.016	0.020	0.024	0.015
SS	0.565	0.324	0.258	0.505	0.377	0.143	0.401
AC/AS/FF	0.101	0.165	0.207	0.035	0.124	0.280	0.350
AC/AS/SS/FF	0.578	0.342	0.282	0.508	0.418	0.299	0.511

$$\mathrm{SC}(i)=\frac{\Sigma_{k=1}^{K} kA_k(i)}{\Sigma_{k=1}^{K} A_k(i)}. \tag{12}$$

To test the degree to which an amplitude simplification affects the centroid, we calculate the rms-amplitude-weighted mean centroid change based on the centroids of the simplified (SCs) and reference (SCr) spectra:

$$\overline{\Delta \mathrm{SC}}=\frac{\sum_{i=1}^{I}\left|\frac{\mathrm{SC}s(i)}{\mathrm{SC}r(i)}-1\right|\cdot A_{\mathrm{rms}}(i)}{\Sigma_{i=1}^{I} A_{\mathrm{rms}}(i)}. \tag{13}$$

This quantity is zero if there is no difference in centroid and it is unbounded, although for our simplifications $\overline{\Delta \mathrm{SC}}$ attained a maximum value of 0.3.

Of course, the amplitude-envelope coherence (AC) simplification may result in a large centroid change for tones with a great deal of spectral flux, since it was designed to eliminate any centroid change during the course of a sound. However, centroid effects, some quite sizable, also occur for AS and SS operations, although the changes induced by AS are generally much less than those due to the other two amplitude simplifications. Table VI gives a list of the average relative centroid changes for the three amplitude simplifications. Mean discrimination data (d') are plotted as a function of $\overline{\Delta \mathrm{SC}}$ in Fig. 11(c). Note that these averages are based on magnitudes of the SC changes. Further inspection of Table VI reveals that for the instruments tested, centroid increases in stimuli with spectral-envelope smoothness are always positive, whereas for the other two simplifications, the change in centroid can go in either direction—even during the sounds. The logarithm of the mean centroid change ex-

TABLE V. Relative spectral differences between reference and simplified spectra for basic frequency simplifications. The values represent ERR_{freq} [Eq. (11)]. Note the values for FF would override all accumulations of this operation with other simplifications. For key, see Table II caption.

	Instrument						
Simplication	Cl	Fl	Hc	Mb	Ob	Tp	Vn
FC	0.001	0.002	0.010	0.001	0.003	0.003	0.003
FS	0.001	0.004	0.013	0.003	0.008	0.004	0.003
FF	0.002	0.005	0.013	0.003	0.010	0.007	0.004

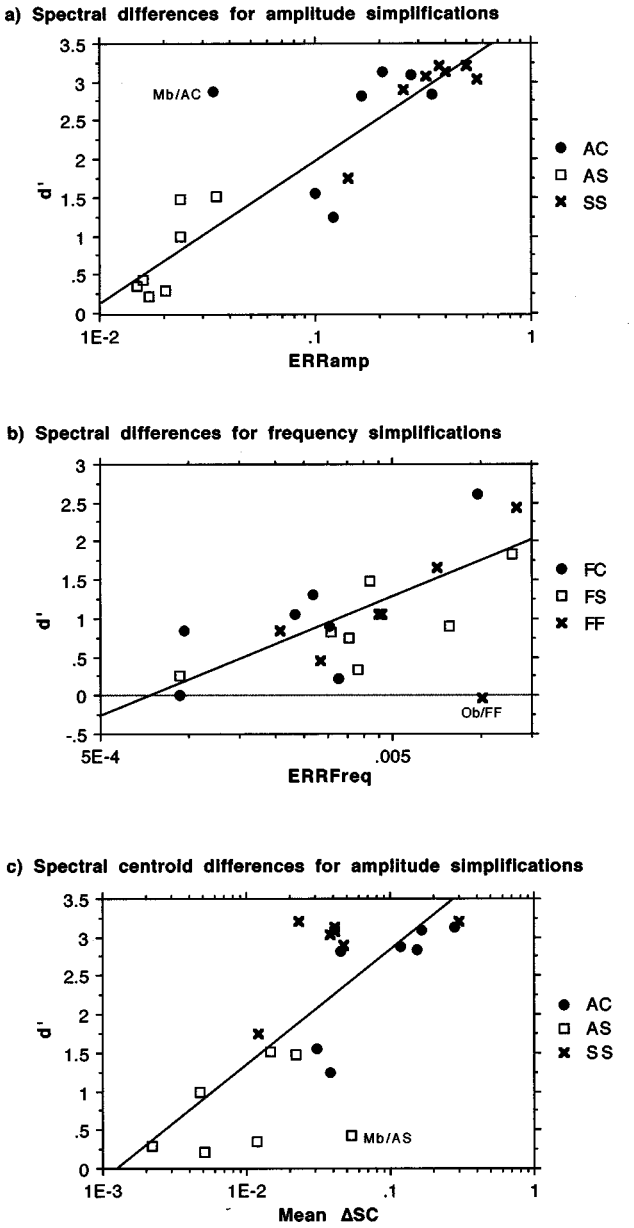

FIG. 11. Discrimination scores (d') plotted as functions of the logarithms of three objective measures [ERR_{amp} (a), ERR_{freq} (b), and $\overline{\Delta SC}$ (c)] for the single amplitude and frequency-envelope simplifications indicated for each panel. The linear regression lines were computed without one outlying point indicated in each panel: Mb/AC in (a), Ob/FF in (b), and Mb/AS in (c), since the removal of these single points resulted in dramatic increases in the correlation coefficients in each case (see the text for complete descriptions of the objective measures).

plains 54% of the variance in d', but this value increases to 65% when the outlying point due to the marimba in the AS condition is removed.

V. DISCUSSION

The discrimination data show that the sensitivity of listeners to simplifications of musical instrument spectra depends on the parameter modified and the instrument sound processed, while being relatively unaffected by the musical training of the listeners. The interaction between type of simplification and instrument is most likely due to a combination of the perceptual salience of the parameter simplified and the strength of that parameter in the particular sound. From Fig. 10 it is quite obvious that spectral-envelope smoothness and amplitude-envelope coherence are the most discriminable simplifications. However, spectral-envelope smoothness causes a smaller perceptual change for the trumpet than the other instruments, which is not unexpected since its spectrum is quite smooth to begin with. For this latter instrument, amplitude-envelope coherence is the most discriminable simplification, due to the strong degree of spectral flux present in brass tones (Grey, 1977). Further, the amplitude-envelope coherence simplification is much less discriminable for the clarinet and oboe because their spectra do not undergo as much spectral flux as most other instruments (Grey, 1977).

The other simplifications result in lesser discrimination scores, either because these involve parameters of lesser perceptual salience or because the parameters have insufficient strength to result in higher scores. Again, scores depend on the instrument tested. Amplitude-envelope smoothness seems to be most important for the flute, trumpet, and harpsichord, the former two because of their relatively large temporal variations and the latter because of its effect on the decay curves.

Objective measures of average spectral ($\overline{\Delta SC}$) and spectrotemporal change (ERR_{amp},ERR_{freq}) were developed in an attempt to quantify the acoustic cues that give rise to the discrimination performance. Figure 11 clearly shows that the three amplitude simplifications (AC, AS, and SS) have different effects on changes in amplitude-envelope structure and consequently that they are discriminated to differing degrees as well. Spectral-envelope smoothness is almost always the most easily detected, followed by amplitude-envelope coherence, and finally by amplitude-envelope

TABLE VI. Average relative-magnitude change of centroid [$\overline{\Delta SC}$, Eq. (13)] for the three basic amplitude simplifications and two accumulations of those simplifications. Note that centroids would not be appreciably affected by the FF simplification since the frequency variations were less than 1% during the significant amplitude portions of the sounds. A minus sign indicates that, on average, the centroid for the simplified sound decreased compared to the reference sound. For key, see Table II caption.

	Instrument						
Simplification	Cl	Fl	Hc	Mb	Ob	Tp	Vn
AC	0.031	0.045	0.279(−)	0.118(−)	0.038	0.166	0.155
AS	0.005(−)	0.023	0.015(−)	0.054(−)	0.002(−)	0.005	0.012(−)
SS	0.039	0.041	0.047	0.299	0.023	0.012	0.042
AC/AS/FF	0.031	0.046	0.282(−)	0.117(−)	0.038	0.166	0.156
AC/AS/SS/FF	0.058	0.061	0.284(−)	0.359	0.038	0.168	0.156

smoothness, which gives performance not far from chance for four of the seven instrument sounds.

Similarly, note that for most sounds the effect of amplitude-envelope smoothness on spectral centroid change (mean=0.016) is much less than that of amplitude-envelope coherence (0.119), with the flute being the only exception (0.023 vs 0.045, respectively). Indeed, the flute has a comparatively high discrimination score for amplitude-envelope smoothness (0.80), whereas the marimba with amplitude smoothness has a moderately high spectral-centroid change (0.054) and a relatively low discrimination rate (0.59). Also, surprisingly, the flute has a high discrimination score for the amplitude-coherence simplification (0.96), even though its change of centroid (0.041) is slightly less than the amplitude-based simplifications (overall mean for the three, 0.045). On the other hand, we see that in comparison to the amplitude-envelope coherence, spectral-envelope smoothness causes moderate to high centroid changes (0.02–0.30), with the trumpet (0.01) being the obvious exception as mentioned above. The marimba exhibits a large relative centroid change (0.30), but this is true because the marimba's sound is dominated by its fundamental. In this case, spectral-envelope smoothness makes a profound change by introducing a second harmonic which was originally nonexistent. Note, however, that spectral-envelope smoothness will have an effect on any jagged spectrum, regardless of whether the spectrum is changing or not, whereas amplitude-envelope coherence only affects sounds with time-varying spectra. Since spectral-envelope smoothness inherently affects the centroid, we cannot tell whether discrimination is due to this effect or directly to the change of spectral-envelope fine structure, but it is probably due to a combination of these effects.

All of the amplitude simplifications produce changes in both $\mathrm{ERR}_{\mathrm{amp}}$ and $\overline{\Delta \mathrm{SC}}$ measures. Further, these two objective measures partially explain the variance in discrimination performance and yet are not strongly correlated between them ($r=0.61$). This suggests that they may both contribute to the discrimination of amplitude-related changes in the spectrotemporal morphology of the simplified instrument sounds. To test this idea, the logarithms of both parameters were selected as independent variables in a stepwise regression across single-amplitude simplifications with d' as the dependent variable. This technique tests the independent contribution of each parameter, which only enters the regression if its contribution is statistically significant (F-to-enter=4, in our case). Both parameters successfully enter into the regression. The final result is given by the following linear regression equation, by which 83% of the variance in the data is explained:

$$d' = 4.34 + 1.35 \cdot \log(\mathrm{ERR}_{\mathrm{amp}}) + 0.64 \cdot \log(\overline{\Delta \mathrm{SC}}). \qquad (14)$$

It becomes clear from this analysis that there are at least two perceptual cues contributing to discrimination performance in these sound simplifications.

The striking thing about the frequency-related simplifications is their relative weakness in creating discriminable differences in the stimuli. This result may be due primarily to the fact that, in normal instrument sounds without vibrato, the amount of frequency variation is relatively small. Indeed, as can be gleaned from Table V, the largest change in frequency variation created by flattening or smoothing the frequency envelopes is for the harpsichord and is on the order of 1.3%; the next largest is on the order of 1% produced by frequency flattening of the oboe. It is perhaps surprising, therefore, that so much has been written in the literature about the importance of frequency microvariations in the creation of naturalness in synthetic sounds (Dubnov and Rodet, 1997; McAdams, 1984; Sandell and Martens, 1995; Schumacher, 1992). Nonetheless, there are certainly classes of musical sounds where pitch contour plays an important role in musical expressiveness, such as vibrato and portamento, particularly in vocal and bowed-string sounds.

The effect of combining amplitude-related and frequency-related cues for the accumulated simplifications is less clear in the data, however. In a stepwise regression of the entire set of d' scores on all three objective measures, only $\mathrm{ERR}_{\mathrm{amp}}$ entered significantly into the regression and then explained only 63% of the total variance. So, while individual cues seem to explain a large portion of the variance for the basic simplifications, their combined use in judging accumulated simplifications remains uncertain. This may be due in part to the judgment strategy discussed above, namely that listeners respond to the most salient parameter in an accumulation of parameters. In the discrimination data, there are only four out of 35 cases where an accumulation gives higher discrimination scores than the best of its component simplifications: AC/FF is better than either component for the oboe, and AS/FF is better than either component for clarinet, flute, and oboe. If our objective measures are truly indicative of the perceptual cues being used by listeners to perform the task, they should have a similar pattern to the discrimination data with accumulations having the same or slightly higher values than their constituent simplifications. Globally this is the case (see Tables IV, V, and VI): AC/AS/FF is approximately equal to AC for all seven instruments in terms of both $\mathrm{ERR}_{\mathrm{amp}}$ and $\overline{\Delta \mathrm{SC}}$, and AC/AS/SS/FF is approximately equal to or slightly higher than SS or AC in all seven instruments in terms of $\mathrm{ERR}_{\mathrm{amp}}$, although it is quite a bit higher for clarinet, flute, and marimba in terms of $\overline{\Delta \mathrm{SC}}$. The combination of the psychophysical data and the objective measurements would thus seem to globally support the most-potent cue judgment strategy.

VI. CONCLUSIONS

The results of this study point very strongly to (1) spectral-envelope shape (jagged vs smooth) and (2) spectral flux (time variation of the normalized spectrum) as being the most salient physical parameters that we have studied related to timbre discrimination, followed in order by (3) the presence of frequency variation, (4) frequency incoherence (inharmonicity), (5) frequency microvariation, and (6) amplitude microvariation. Simplifications (reductions or eliminations) of these parameters give rise to changes in the spectrotemporal morphology of an instrument sound's sensory representation, to which both musician and nonmusician

listeners are very sensitive. This sensitivity is only slightly greater in musicians than in nonmusicians. The level of discrimination resulting from the modifications was globally greater for the amplitude simplifications than for the frequency simplifications, with the exception of amplitude smoothing. Thus, musical-sound synthesis should pay particular attention to spectral-envelope fine-structure and spectral flux if a high degree of audio quality is to be ensured.

Objective measures were defined that predict a great deal of the discrimination performance. These measures are related to changes in the amplitude envelopes and the spectral centroid for amplitude simplifications, and to changes in the frequency envelopes for frequency simplifications. Since discrimination can be predicted by physical measurement of differences in the time-varying spectra, it appears that the importance of these parameters is in direct proportion to the extent to which they actually vary in musical sounds, as we have shown with the strong interaction between simplification type and musical instrument. Further work is needed to examine the relative perceptual sensitivity of listeners to these different physical factors. We have also shown that if several parameters are varied simultaneously, listeners appear to use the most salient one, and their discrimination performance can, for the most part, be predicted on the basis of it alone. While it is likely that this acute sensitivity to the fine-grained spectral and temporal structure of the musical sounds exists across the entire range of pitch, dynamics, and articulation possible on each instrument, further research will be needed to determine the relative importance of the different objective parameters in different regions of an instrument's musical ''space.''

ACKNOWLEDGMENTS

We would like to thank Sophie Savel for running subjects in the control experiment, Bennett Smith for the wonders of PSIEXP, as well as John Hajda and two anonymous reviewers for helpful comments.

APPENDIX A

We can write the reduced-duration harmonic amplitude envelope as

$$A_k(t) \leftarrow \begin{cases} A_k(t), & 0 \leqslant t < t_1, \\ (1-\alpha(x)) \cdot A_k(t) + \alpha(x) \cdot A_k(t+t_2-t_3), & \\ \quad t_1 \leqslant t < t_3, & \\ A_k(t+t_L-2), & t_3 \leqslant t \leqslant 2, \end{cases} \tag{A1}$$

where

$$t_3 = 2-(t_L-t_2), \quad \alpha(x) = 3x^2 - 2x^3, \quad x = \frac{(t-t_1)}{(t_3-t_1)},$$

and t_L is the duration of the original sound.

Note that $\alpha(x)$ is a cubic spline with the following properties:

(1) The derivative of $\alpha(x)$ with respect to x is zero at $x=0$ and $x=1$,
(2) $\alpha(0)=0$,
(3) $\alpha(1)=1$.

The same method obviously applies to the harmonic-frequency envelopes.

APPENDIX B

Since the data rate for each harmonic amplitude or frequency envelope is originally $2f_a$, the overall data rate for K harmonic amplitude and frequency envelopes is $4K \cdot f_a$.

Amplitude-envelope smoothing (AS) and frequency-envelope smoothing (FS) essentially reduce the data rate for each harmonic envelope from $2f_a$ to $2f_c$, where f_c is the filter cutoff frequency. If only amplitude-envelope smoothing were applied, the data rate for K harmonics would be reduced to $2K \cdot f_c + 2K \cdot f_a = 2K \cdot (f_c + f_a)$. In our case, since $f_a = 311$ Hz and $f_c = 10$ Hz, the data reduction factor would be $4.311/[2 \cdot (10+311)] = 1.94$. The same result would apply if only frequency-envelope smoothing were applied. On the other hand, if both were applied the new total data rate would be $4K \cdot f_c$, and the data reduction factor would be f_a/f_c. In our case, this is $311/10 = 31.1$, i.e., there is only a substantial overall data reduction if both amplitude- and frequency-envelope smoothing are applied.

Spectral-envelope smoothing does not reduce the data rate very much, at least not with the current definition of smoothing. Since the order of the smoothing function is 2, the reduction is approximately a factor of 2.

Amplitude- (AC) and frequency coherence (FC) simplifications essentially replace multiple envelopes by single envelopes. If one of these simplifications were applied, the data rate falls from $4K \cdot f_a$ to $2K \cdot f_a + 2f_a = 2(K+1) \cdot f_a$. So, the data reduction factor would be approximately 2. If both were applied, the data reduction factor would be exactly K, the number of harmonics. In our case, this varies from 30 to 70, depending on the instrument.

The data rate for flattened frequency envelopes is zero. So, if frequency flattening (FF) is applied, the data rate goes from $4K \cdot f_a$ to $2K \cdot f_a$, a factor of 2 reduction.

Data rates after combinations of data simplifications can be calculated from the individuals. For example, if AC and FF are combined, the data rate becomes $2f_a$. For AS and FF, it would be $2K \cdot f_c$. For SS and FF, it is $K \cdot f_a$. For AC, AS, and FF, it is $2f_c$. For AC, AS, SS, and FF, it is just f_c. The corresponding data-reduction factors are for AC/FF, $2K$; for AS/FF, $2f_a/f_c$; for SS/FF, 4; for AC/AS/FF, $2K \cdot f_a/f_c$; for AC/AS/SS/FF; $4K \cdot f_a/f_c$.

[1] We were unable to find a trumpet tone of suitable quality recorded at E-flat 4, so we used a tone recorded by author J.B. which was within a whole tone of that pitch, *F*4. When resynthesized at E-flat 4, it sounded perfectly natural to all of the authors.

[2] A control experiment designed to test discrimination of the digitized recordings and the fully analyzed–resynthesized sounds was conducted with six listeners. Each subject performed 40 trials for each instrument using the paradigm described in Sec. II C. The discrimination rates for oboe, clarinet, flute, harpsichord, marimba, trumpet, and violin were 0.62, 0.54, 0.59, 0.64, 0.53, 0.57, and 0.53, respectively. Chance performance would be at 0.50 in this two-alternative forced-choice paradigm. Discrimination rates above

0.55 are significantly different from chance ($p<0.05$) by an exact binomial test for 240 trials (6 listeners×40 trials). Therefore, the full reconstructions were discriminated from the original recordings in four of the seven cases (oboe, harpsichord, marimba, and trumpet), although their discrimination rates are still quite low. That the discrimination scores are as high as 64% is actually quite surprising, given that the authors could not discriminate them informally. The subjects must have been operating from extremely subtle cues which only became obvious upon repeated listening. Two possible cues are phase differences and low-frequency noise differences between the original and synthetic cases. Improvements in the analysis/synthesis algorithms should be able to close this gap in the future. However, for this study we are confident that the vast majority of the spectrotemporal features survived the analysis/synthesis process intact. Moreover, the duration-shortened resynthesized sounds were created to be the reference sounds, so our conclusions, which relate to these sounds, are not affected by this finding.

[3]Audible artifacts were sometimes noticeable with the AC and SS simplifications. Amplitude-envelope coherence created two types of artifacts: a ''shh'' noise at the end of the sound (clarinet) or a kind of general muting of the sound (harpsichord, marimba, and trumpet). In fact, in rendering the amplitude envelopes coherent, one increases the amplitude of weak harmonics that start later and end earlier in parts of the signal that are near the noise floor. The frequency estimation is not very precise for these temporal regions of such harmonics (Moorer, 1978), and by increasing their amplitude, the existing imprecise fluctuations in frequency become clearly audible. The spectral-envelope smoothing creates a kind of gargling sound in the flute, marimba, and trumpet. This simplification also increases the level of weak harmonics whose representations have been corrupted by stronger neighbors due to channel leakage, thus amplifying fluctuations due to imprecise estimation of their frequency. These two kinds of artifacts disappear or are notably reduced when combined with frequency coherence or frequency flatness, since the frequency fluctuations producing the artifacts are then reduced or eliminated.

[4]Probabilities are corrected where necessary by the Geisser–Greenhouse epsilon (Geisser and Greenhouse, 1958), which is a conservative adjustment to account for inherent correlations among repeated measures.

Beauchamp, J. W. (**1993**). ''Unix workstation software for analysis, graphics, modifications, and synthesis of musical sounds,'' *94th Convention of the Audio Engineering Society*, Berlin (Audio Engineering Society, New York), Preprint 3479 (L-I-7).

Beauchamp, J. W., McAdams, S., and Meneguzzi, S. (**1997**). ''Perceptual effects of simplifying musical instrument sound time-frequency representations,'' J. Acoust. Soc. Am. **101**, 3167.

Brown, J. C. (**1996**). ''Frequency ratios of spectral components of musical sounds,'' J. Acoust. Soc. Am. **99**, 1210–1218.

Charbonneau, G. R. (**1981**). ''Timbre and the perceptual effects of three types of data reduction,'' Comput. Music J. **5(2)**, 10–19.

Dubnov, S., and Rodet, X. (**1997**). ''Statistical modeling of sound aperiodicities,'' in *Proceedings of the International Computer Music Conference*, 1997, Thessaloniki (International Computer Music Association, San Francisco), pp. 43–50.

Geisser, S., and Greenhouse, S. W. (**1958**). ''An extension of Box's results on the use of the *F* distribution in multivariate analysis,'' Ann. Math. Stat. **29**, 885–891.

Green, D. M., and Swets, J. A. (**1974**). *Signal Detection Theory and Psychophysics* (Krieger, Huntington, NY).

Grey, J. M. (**1977**). ''Multidimensional perceptual scaling of musical timbres,'' J. Acoust. Soc. Am. **61**, 1270–1277.

Grey, J. M., and Gordon, J. W. (**1978**). ''Perceptual effects of spectral modifications on musical timbres,'' J. Acoust. Soc. Am. **63**, 1493–1500.

Grey, J. M., and Moorer, J. A. (**1977**). ''Perceptual evaluations of synthesized musical instrument tones,'' J. Acoust. Soc. Am. **62**, 454–462.

Iverson, P., and Krumhansl, C. L. (**1993**). ''Isolating the dynamic attributes of musical timbre,'' J. Acoust. Soc. Am. **94**, 2595–2603.

Kendall, R. A., and Carterette, E. C. (**1996**). ''Difference thresholds for timbre related to spectral centroid,'' in *Proceedings of the 4th International Conference on Music Perception and Cognition, Montreal*, pp. 91–95 (Faculty of Music, McGill University).

Krimphoff, J., McAdams, S., and Winsberg, S. (**1994**). ''Caractérisation du timbre des sons complexes. II. Analyses acoustiques et quantification psychophysique,'' Journal de Physique **4**(C5), 625–628.

Krumhansl, C. L. (**1989**). ''Why is musical timbre so hard to understand?,'' in *Structure and Perception of Electroacoustic Sound and Music*, edited by S. Nielzén and O. Olsson (Excerpta Medica, Amsterdam), pp. 43–53.

Macmillan, N. A., and Creelman, C. D. (**1991**). *Detection Theory: A User's Guide* (Cambridge U.P., Cambridge).

McAdams, S. (**1984**). ''Spectral Fusion, Spectral Parsing, and the Formation of Auditory Images,'' unpublished Ph.D. dissertation, Stanford University, Stanford, CA, App. B.

McAdams, S., Winsberg, S., Donnadieu, S., De Soete, G., and Krimphoff, J. (**1995**). ''Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes,'' Psychol. Res. **58**, 177–192.

Miller, J. R., and Carterette, E. C. (**1975**). ''Perceptual space for musical structures,'' J. Acoust. Soc. Am. **58**, 711–720.

Moorer, J. A. (**1978**). ''The use of phase vocoder in computer music applications,'' J. Audio Eng. Soc. **24**, 717–727.

Ott, R. L. (**1993**). *An Introduction to Statistical Methods and Data Analysis* (Duxbury, Belmont, CA).

Plomp, R. (**1970**). ''Timbre as a multidimensional attribute of complex tones,'' in *Frequency Analysis and Periodicity Detection in Hearing*, edited by R. Plomp and G. F. Smoorenburg (Sijthoff, Leiden), pp. 397–414.

Preis, A. (**1984**). ''An attempt to describe the parameter determining the timbre of steady-state harmonic complex tones,'' Acustica **55**, 1–13.

Repp, B. H. (**1987**). ''The sound of two hands clapping: An exploratory study,'' J. Acoust. Soc. Am. **81**, 1100–1109.

Sandell, G. J., and Martens, W. L. (**1995**). ''Perceptual evaluation of principal-component-based synthesis of musical timbres,'' J. Audio Eng. Soc. **43**, 1013–1028.

Schumacher, R. T. (**1992**). ''Analysis of aperiodicities in nearly periodic waveforms,'' J. Acoust. Soc. Am. **91**, 438–451.

Smith, B. K. (**1995**). ''PSIEXP. An environment for psychoacoustic experimentation using the IRCAM Musical Workstation'', in *Program of the SMPC95: Society for Music Perception and Cognition, Berkeley, CA*, edited by D. L. Wessel (University of California, Berkeley).

von Bismarck, G. (**1974**). ''Sharpness as an attribute of the timbre of steady sounds,'' Acustica **30**, 159–172.

Wessel, D. L. (**1979**). ''Timbre space as a musical control structure,'' Comput. Music J. **3(2)**, 45–52.