Detection of random alterations to time-varying musical instrument spectra

Andrew Horner[a)]
Department of Computer Science, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong

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

Richard So[b)]
Department of Industrial Engineering and Engineering Management, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong

(Received 10 January 2003; accepted for publication 14 June 2004)

The time-varying spectra of eight musical instrument sounds were randomly altered by a time-invariant process to determine how detection of spectral alteration varies with degree of alteration, instrument, musical experience, and spectral variation. Sounds were resynthesized with centroids equalized to the original sounds, with frequencies harmonically flattened, and with average spectral error levels of 8%, 16%, 24%, 32%, and 48%. Listeners were asked to discriminate the randomly altered sounds from reference sounds resynthesized from the original data. For all eight instruments, discrimination was very good for the 32% and 48% error levels, moderate for the 16% and 24% error levels, and poor for the 8% error levels. When the error levels were 16%, 24%, and 32%, the scores of musically experienced listeners were found to be significantly better than the scores of listeners with no musical experience. Also, in this same error level range, discrimination was significantly affected by the instrument tested. For error levels of 16% and 24%, discrimination scores were significantly, but negatively correlated with measures of spectral incoherence and normalized centroid deviation on unaltered instrument spectra, suggesting that the presence of dynamic spectral variations tends to increase the difficulty of detecting spectral alterations. Correlation between discrimination and a measure of spectral irregularity was comparatively low.
 [DOI: 10.1121/1.1778741]

PACS numbers: 43.75.Cd, 43.66.Jh [SEM] Pages: 1800–1810

I. INTRODUCTION

It is common knowledge that musical instruments can be identified even when their spectra have been substantially altered. A trumpet is recognizable when performed in a vast cathedral or small bathroom, played through an advanced 3D-surround system or cheap PC speakers, or modified in various ways through a spectrum equalizer, even though all of these systems substantially modify the trumpet's source spectrum. What is less obvious is the amount of spectral change required for an audible change of timbre to occur. If the band-levels of a spectrum equalizer are set at random within a range of $\pm x$ dB, at what level of x would a listener begin to distinguish the modified signal from the input signal? Does discrimination vary from instrument to instrument? How much does time-varying spectral variation affect this process? For example, are spectral alterations of a relatively static organlike sound more easily detected than those that have more pronounced spectral variations? Answering these questions is the primary objectives of this paper.

Some related previous studies have considered the effect of simple spectral alterations of static spectra, speech, and audio signals. However, to our knowledge, no previous work has addressed the effect of random spectral variations of individual musical instrument sounds with time-variant spectra.

Plomp (1970) considered the correspondence between an error metric and discrimination data based on static musical instrument and vowel spectra. He concluded that differences in timbre can be predicted well from spectral differences. However, spectral differences between the sounds in that study were not systematically varied from small to large.

Toole and Olive (1988) reviewed previous research on the audibility of resonances and explored the effect of modifying the spectra of noise, music recordings, and speech by adding a single resonance to a unity gain signal path. They stated that "it is surprising just how much the ... signal ... can be modified without significantly altering perceived timbre." They concluded that resonances are more easily heard in noise input signals than in speech and music input signals.

The effect of transmission channel spectral-envelope distortion of vowel sounds in speech was investigated by Watkins (1991) and Watkins and Makin (1996). In the 1991 study, short-test vowel stimuli ranging on a continuum from

[a)]Address correspondence to either A. Horner at HKUST (electronic mail: horner@cs.ust.hk) or to J. Beauchamp at UIUC (electronic mail: jwbeauch@uiuc.edu).
[b)]Electronic mail: rhyso@ust.hk

''itch'' to ''etch'' were preceded by a ''carrier'' consisting of a four-word phrase. The carrier phrase was processed by a filter that could turn ''itch'' into ''etch'' in varying degrees. It was found that filtering the carrier phrase would cause subjects to compensate for the filtering and shift their boundary between itch and etch perception. In the 1996 study, spectral distortion was accomplished by increasing the contrast between high-amplitude and low-amplitude portions of the spectrum in such a way that the spectral slope was preserved. By manipulating spectral-envelope distortions of the carrier independently of the test sounds, the authors concluded that listeners compensate for distortion before perceptual features of the vowel are extracted. The principal consequence of this work for the current study is that the perception of a sound can be influenced by the sound that precedes it, suggesting that spectral variations might also have a significant influence on discrimination.

A related area of research is *spectral profile analysis* (Green, 1988). Like random spectrum alteration, spectral profile analysis studies the ability of listeners to discriminate an original stimulus from a spectrally altered version of the stimulus. However, there are significant differences between profile analysis and the approach of the current study. For one, spectral profile analysis usually has only considered static spectra with log-spaced rather than harmonic-spaced components. Also, the reference spectra are usually flat, which is quite different from the spectra of acoustic instruments, which include frequency roll-off. One exception to this is the study by Versfeld and Houtsma (1991) that examined the effect of slope in downward- and upward-ramp-shaped spectra. They found that the detection of spectral changes in a sound is strongly dependent on the frequency spacing of the components, concluding that harmonic spectra would give quite different results than log-spaced spectra. Some researchers have also used ''perturbed'' or ''jagged'' spectra (Kidd *et al.*, 1991). Lentz and Richards (1998) noted that spectral jaggedness increases the threshold of alteration detection.

Another basic difference from our approach is that profile analysis studies have usually only attempted to determine the threshold of discriminating an increase in the amplitude of a single spectral component, or at most a few components in a band. Bernstein *et al.* (1987) observed that thresholds for amplitude changes to a single component cannot be used for accurate prediction of thresholds for changes to multiple components. Hall *et al.* (1984) and Buus (1985) investigated single tone masking by spectra consisting of slowly varying noise or sinusoids. They found that the masking threshold for a single tone was lowered (or ''released'') when masking spectral components were comodulated, meaning that they were amplitude modulated in-phase by a common low-frequency signal. However, Mendoza *et al.* (1996) found that as time-envelopes of narrow-band noise maskers become more incoherent (less correlated), the amount of masking release decreases (i.e., the masking increases).

It is difficult to generalize profile analysis results to the discrimination of broadband alterations to time-varying (dynamic) harmonic spectra, although the above results suggest that changes to multiple components of acoustic spectra are more difficult to detect than are single components. Also, comodulation-masking-release results suggest that alterations to coherent spectra are easier to detect than are alterations to incoherent spectra.

In the present study, we sought to determine discrimination scores for sustained musical instrument sounds with different levels of time-invariant random spectrum alteration. Noting that discrimination varied significantly from instrument to instrument, we then correlated the discrimination scores with spectral incoherence, spectral centroid deviation, and spectral irregularity of the original sound spectra, to determine whether these different measures of spectral variation are significantly related to discrimination. We present the techniques used for analysis and synthesis of the stimuli and then discuss the discrimination experiment. We then present the results in terms of the effect on the discrimination of the instrument, musical experience, and measured spectral variations of the original stimuli.

II. STIMULUS PREPARATION

Eight sustained musical instrument sounds (six of them also used by McAdams *et al.*, 1999) were selected as prototype signals for stimulus preparation. Each entire sound was used as a stimulus, including the attack, sustain, and release. The sounds were first subjected to time-variant spectrum analysis using a computer-based phase vocoder method (Beauchamp, 1993). This phase vocoder is different from most in that it allows a fixed analysis frequency (f_a) to be tuned to an estimated fundamental frequency of the input signal. The analysis method yields frequency deviations between harmonics of the analysis frequency and the corresponding frequencies of the input signal, which are assumed to be approximately harmonic relative to the fundamental. The harmonic frequency deviations are assumed to be within $\pm 2\%$ of the corresponding harmonic of the analysis frequency.

A. Signal representation

Each sound stimulus was represented as a sum of sinusoids with time-varying amplitudes and frequencies:

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

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 (approximately 311 Hz), $\Delta f_k(\tau)$ is the kth harmonic's frequency deviation, so that $f_k(\tau)=kf_a+\Delta f_k(\tau)$ is the estimated frequency of the kth harmonic, and $\theta_k(0)$ is the initial phase of the kth harmonic.

The parameters used for resynthesis in this study are f_a and $A_k(t)$. The frequency deviations, $\Delta f_k(t)$, were set to zero in order to restrict listener attention to the amplitude data. Although the $A_k(t)$ were only stored as samples occurring every $T_o=1/(2f_a)$ s, the synthetic signals were gener-

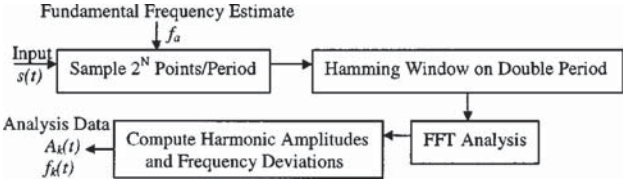

FIG. 1. Method for time-varying spectral analysis that yields the amplitude and frequency deviations for each harmonic, k. The estimated frequency for harmonic k is given by $f_k(t)=kf_a+\Delta f_k(t)$, where f_a is the analysis fundamental frequency.

ated at a much higher data rate (sample frequency 22 050 or 44 100 Hz) by using linear interpolation between these values.

B. Prototype instrument sounds

Sounds of a bassoon, clarinet, flute, horn, oboe, saxophone, trumpet, and violin were used to represent the several wind and the bowed string instrument families. Five of the sounds were taken from the McGill University Master Samples recordings, two from Prosonus (bassoon and oboe), and one (trumpet) had been recorded at the University of Illinois Urbana–Champaign School of Music. Except for the bassoon and horn, these sounds were also used by McAdams *et al.* (1999), who give more details about their characteristics.

C. Analysis method

The phase vocoder method used for analysis consists of the following steps (see Fig. 1):

(1) Band-limited interpolation of the input signal to produce the lowest possible power-of-two number of samples per analysis period ($T_a=1/f_a$), which exceeds 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 ($2T_a$) and which overlap by half an analysis period ($T_o=T_a/2$, where T_o is the time between frames). This amount of overlap minimally satisfies the sample-rate/bandwidth criterion (Allen, 1977).
(3) Multiplication of each signal frame by a Hamming window function whose length is two analysis periods ($2T_a$). When f_a is tuned to the input signal's fundamental frequency, the Hamming window provides separation of neighboring harmonics by at least 42 dB.
(4) Fast Fourier transform (FFT) of the resulting product to produce real and imaginary components (a_k and b_k) at frequencies 0, $f_a/2$, f_a, $3f_a/2$,...,$f_s/2-f_a$, where f_s is the sampling frequency. Components that are not positive integer multiples of f_a are discarded.
(5) Conversion of each retained real and imaginary component pair to give the amplitude and phase of each harmonic [$A_k=\sqrt{a_k^2+b_k^2}$; $\theta_k=\text{atan}\,2(b_k,a_k)$].
(6) Computation of the frequency deviation for each harmonic by a trigonometric identity, which essentially gives the difference in phase between adjacent frames for each harmonic. The phase difference divided by $2\pi T_o$ gives the frequency deviation.
(7) Storage of the harmonic amplitude 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.

For $f_s=44\,100$ Hz and $f_a=311.1$ Hz (Eb4), the maximum number of harmonics that can be analyzed is 70. For $f_s=22\,050$ Hz, this reduces to 35. Because harmonic amplitudes were judged (by visual inspection of spectra) to be insignificant beyond $K=35$ for the bassoon, oboe, and trumpet sounds, $f_s=22\,050$ Hz was used for these. The other sounds were sampled at 44 100 Hz.

Further details on the analysis procedure are discussed by Beauchamp (1993).

The analysis system may be viewed as a set of contiguous bandpass filters that 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 whose frequencies line up with the filter band centers, so that the output of each filter is one of the sine waves. The analysis gives the amplitude and frequency of each sine wave. When the filter outputs are summed, the signal is almost perfectly reconstructed. In fact, the output signal can be viewed as that created by processing the input signal by the sum of the bandpass-filter characteristics. For the Hamming window, it can be shown that this sum varies only by 1.4 dB over the range $[f_a/2,f_s/2]$, with maxima occurring at the harmonic frequencies and minima at the halfway points. Figure 1 shows a block diagram of the basic analysis/synthesis system.

D. Duration and loudness equalization

Sound duration is a potential factor in discrimination. For example, a sound lasting 5 s might well be easier to discriminate than a sound lasting 0.3 s. In order that duration would not be a factor in the study, the sounds were standardized to a 2-s duration by interpolating the analysis data. Next, the eight duration-equalized prototype sounds were compared, and amplitude multipliers were determined such that the sounds were judged to have equal loudness. McAdams *et al.* (1999) give more details on these equalizations. It should be mentioned, however, that these equalizations were not essential for the present study, since each discrimination pair was always derived from a single prototype sound.

E. Frequency flattening

In order that they would not be factors in this study, frequency variations and inharmonicity were eliminated from the sounds by setting each harmonic's frequency equal to the product of its harmonic number (k) and the fixed analysis frequency (f_a), resulting in flat, equally spaced frequency envelopes; i.e.,

$$f_k=kf_a. \tag{2}$$

Frequency flattening was previously shown to have an effect on discrimination by Gray and Moorer (1977), Charbonneau (1981), and McAdams *et al.* (1999). For the five sustain instruments they tested, McAdams *et al.* found that frequency

flattening resulted in discrimination scores in the range of 48% to 82% and that discrimination is weakly correlated (34% to 57%) with the amount of frequency deviation in the original sound. In the current study, frequency flattening was employed to focus listeners' attention on harmonic amplitude variations and also to avoid frequency fluctuations which may influence subjects' judgments, especially considering that audible artifacts can result from these fluctuations when they are amplified by altered harmonic amplitudes. The frequency-flattened sounds then served as the reference sounds for this study, and their corresponding time-varying harmonic amplitude spectra are henceforth referred to as the *analysis data*.

F. Random spectrum alteration

Time-invariant random alteration was performed on the analysis data, after which the sounds were regenerated by additive synthesis. Randomly altered harmonic amplitudes are indicated with the prime symbol, e.g., $A'_k(t)$. Alteration was accomplished by multiplying each harmonic amplitude by a randomly selected scalar, r_k:

$$A'_k(t) = r_k A_k(t). \tag{3}$$

The $\{r_k\}$ were selected uniformly in the range $[1-2\varepsilon, 1+2\varepsilon]$, where ε is referred to as the *error level*. Equation (3) describes a linear stationary process, as the value of each r_k remains the same throughout the sound. The goal of this method is to perturb each harmonic amplitude so that a desired overall spectral error is achieved, without changing the spectral centroid or rise time.

The *relative-amplitude spectral error* is defined as

$$\varepsilon' = \frac{1}{N}\sum_{n=0}^{N-1} \sqrt{\frac{\Sigma_{k=1}^{K}(A_k(t_n) - A'_k(t_n))^2}{\Sigma_{k=1}^{K} A_k^2(t_n)}}, \tag{4}$$

where n is the analysis frame number, $t_n = nT_o$ is the frame time, and N is the number of frames used in the calculation. For this study, $N=20$, where 10 values are taken from the perceptually important attack and 10 values are taken from the remainder of the sound, both equally spaced in time. The attack time is defined as the time from the beginning of the sound to the sound's maximum rms amplitude. Equation (4) is similar to the spectral distance measure used by McAdams *et al.* (1999), except that the normalization is slightly different and fewer amplitude values are used.

Under the definitions of Eqs. (3) and (4), the relative-amplitude spectral error usually varies between 0 and 1. Note that Eq. (4) computes the average of the spectral error at each instant relative to the rms amplitude at that instant. Due to the rms amplitude in the formula's denominator, low-amplitude portions of the sound are given the same weight as high-amplitude portions. (It is assumed that proportional-amplitude errors are more relevant than absolute-amplitude errors.) The normalized-squared errors are then accumulated over the harmonics and averaged over time. One could argue that the error metric might be perceptually improved by first accumulating amplitudes by critical bands before averaging, but we have investigated this possibility and, for a similar set of sound signals, have found no improvement in terms of correlation with discrimination results.

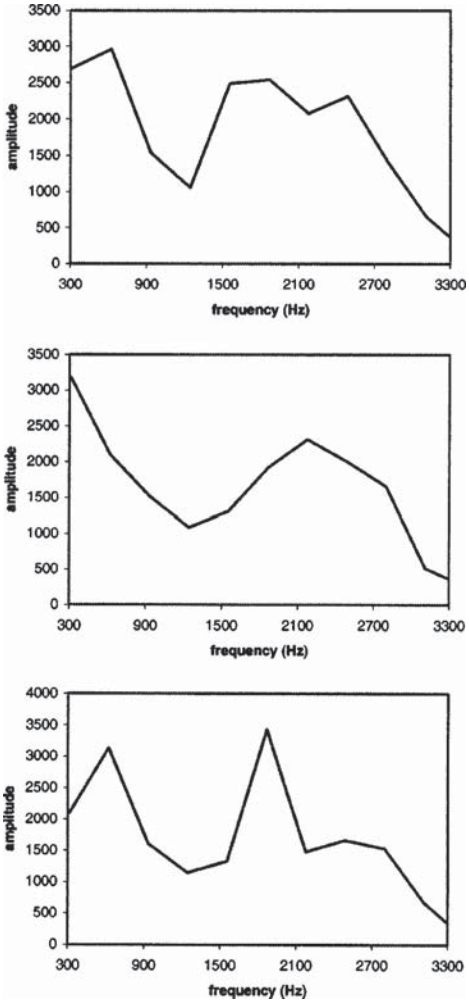

FIG. 2. Original trumpet sound spectral envelope (left) and two spectral envelopes (center and right) generated by random spectrum alteration of the original sound resulting in approximately 24% relative-amplitude spectral error.

Since the $\{r_k\}$ were selected uniformly in the range $[1-2\varepsilon, 1+2\varepsilon]$, the relative-amplitude spectral error was expected to be approximately ε. However, the relative-amplitude error, ε', as defined by Eq. (4), is slightly different from ε. For our purposes, we randomly generated the $\{r_k\}$ and then calculated ε'. If ε' was in the range $[\varepsilon-0.01, \varepsilon+0.01]$, we accepted the $\{r_k\}$; otherwise, we continued to generate and test the $\{r_k\}$ until ε' fell in the desired range.

For example, Fig. 2 shows an original spectral envelope and two randomly altered spectral envelopes, each with 24% error. Note that the first altered spectral envelope is smoother than the original, while the second one is more irregular than the original.

Spectral centroid has been shown to be strongly correlated with one of the most prominent dimensions of timbre as derived by multidimensional scaling (MDS) experiments (Gray and Gordon, 1978; Wessel, 1979; Krumhansl, 1989; Iverson and Krumhansl, 1993; Krimphoff *et al.*, 1994; Kendall and Carterette, 1996; Lakatos, 2000). Normalized cen-

troid versus time functions for original and altered sounds are defined as

$$NSC(t_n)=\frac{\Sigma_{k=1}^{K} kA_k(t_n)}{\Sigma_{k=1}^{K} A_k(t_n)} \quad \text{and}$$

$$NSC'(t_n)=\frac{\Sigma_{k=1}^{K} kA'_k(t_n)}{\Sigma_{k=1}^{K} A'_k(t_n)}. \tag{5}$$

Preserving the spectral centroid after random spectral alteration has been applied eliminates this parameter from being a correlate of discrimination. Therefore, the $\{r_k\}$ were only accepted if the peak (over time) centroid of the original and altered spectra were matched within $\pm 2.5\%$.

The final spectrum alteration algorithm then consists of the following steps:

(1) Pick a vector $\{r_k\}$ such that, for each harmonic k, $1-2\varepsilon<r_k<1+2\varepsilon$.
(2) Generate modified harmonic amplitude envelopes, $A'_k(t)=r_kA_k(t)$.
(3) Calculate the relative-amplitude spectral error, ε' [see Eq. (4)].
(4) If ε' is outside the range $[\varepsilon-0.01,\varepsilon+0.01]$, go to step (1) to repick $\{r_k\}$.
(5) Calculate the peak spectral centroids of the original and randomly altered sounds [see Eq. (5)].
(6) If the peak centroid NSC' of the altered spectrum is outside the range $[0.975\cdot NSC_{\text{peak}},1.025\cdot NSC_{\text{peak}}]$, go to step (1) to repick $\{r_k\}$.
(7) End.

G. Resynthesis method

Resynthesis was accomplished by additive (or Fourier) synthesis of the altered harmonic sine waves:

$$\hat{s}(t)=\sum_{k=1}^{K} A''_k(t)\cos(2\pi kf_at+\theta_k(0)), \tag{6}$$

where $A''_k(t)$ is the linear interpolation of the $A'_k(t_n)$ amplitude envelope data between frames, so that

$$A''_k(t)=\frac{t_{n+1}-t}{T_o}A'(t_n)+\frac{t-t_n}{T_o}A'_k(t_{n+1}),$$
$$t_n\leqslant t\leqslant t_{n+1}. \tag{7}$$

While linear interpolation introduces a smoothing operation to the harmonic amplitude functions, it has a very minor effect on the resulting sound. Also, both the randomly altered sounds and the original reference sounds are resynthesized by additive synthesis in the same fashion, with strictly fixed harmonic frequencies, before being compared in the listening experiment.

III. EXPERIMENTAL METHOD

A. Subjects

Twenty listeners participated in the experiment. They were undergraduate students at the Hong Kong University of Science and Technology, ranging in age from 20 to 23 years, who reported no hearing problems. They included 17 males and 3 females, and their experience playing a musical instrument was recorded. For the 13 listeners with musical experience, the range in musical experience varied from 1 to 5 years with a mean of 2.5 years. The listeners were paid to compensate for their time spent in the experiment.

B. Stimuli

The eight musical instruments used belong to the air column (air reed, single reed, lip reed, double reed) and bowed string families: bassoon, clarinet, flute, horn, oboe, saxophone, trumpet, and violin. Each sound was analyzed and resynthesized using the reference analysis data with no frequency variations and no inharmonicity. The fixed harmonic frequencies encouraged listeners to focus their attention exclusively on the amplitude data. They were prevented from detecting cues stemming from the amplitudes of originally weak sinusoids with noisy harmonic frequency variations possibly made more audible by random spectral alteration. Also, the original sustained sounds had very nearly strictly harmonic frequencies with relatively small frequency deviations.

The sounds were stored in 16-bit integer format on a hard disk. All "reference sounds" (resynthesized using the analysis data with strictly fixed harmonic frequencies) were equalized for duration and loudness. The randomly altered sounds for each instrument were resynthesized with the method described in Sec. II G on an Intel PC.

Ten different randomly altered sounds were synthesized for each of five spectral error levels (8%, 16%, 24%, 32%, and 48%), yielding a total of 50 modified sounds for each instrument. The initial random multipliers, r_k, used were the same for each of the instruments. However, the random-alteration algorithm guaranteed that the relative-amplitude spectral errors, as measured by Eq. (4), were within $\pm 1\%$ of the prescribed error levels while matching the reference sounds in duration, fundamental frequency, and peak centroid. Like the reference sounds, the randomly altered sounds were also synthesized using strictly fixed harmonic frequencies. Also, using the Moore–Glasberg loudness program (Moore *et al.*, 1997), it was determined that loudnesses of the altered sounds matched those of the reference sounds within 2 phons.

C. Test procedure

A two-alternative forced-choice (2AFC) discrimination paradigm was used. The listener heard two pairs of sounds and chose which pair was "different." Each trial structure was one of AA-AB, AB-AA, AA-BA, or BA-AA, where A represents the reference sound and B one of the 50 randomly altered sounds. This paradigm had the advantage of not being as susceptible to variations in listeners' criteria across experimental trials compared to the simpler A-B method. All four combinations were presented for each randomly altered sound. 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. For each trial, the user was prompted with "which

pair is different, 1 or 2?'' and the response was given by the keyboard. The computer would not accept a response until all four sounds in a trial had been played.

For each instrument, a block of 200 trials was presented to each of the listeners (four trial structures×5 error levels×10 different random alterations). The order of presentation of these 200 trials was randomized. For each random alteration, discrimination performance was averaged using the results of the four trials for each listener. Because these four trials were presented in random order within the 200 trials, the effects of possible learning were averaged out. The same trials were presented to each listener, although in a different random order. The duration of each block was about 40 min, and listeners took 5–10-min breaks after each block. Eight blocks were presented to each listener, corresponding to the eight instruments. The order of presentation of the instruments was also randomized for each listener. Sessions of four blocks were scheduled on two different days. The average time to complete a block was 45 min (including breaks). The total duration of the experiment was about 6 h. A custom program written at HKUST ran on an Intel PC to control the experiment.

Listeners were seated in a ''quiet room'' with less than 40 dB SPL background noise level (mostly due to computers and air conditioning). The headphones also provided some reduction of the noise level. Sound signals were converted to analog by a SoundBlaster Audigy soundcard and then presented through Sony MDR-7506 headphones at a level of approximately 75 dB SPL as measured with a sound-level meter. The Audigy DAC utilized 24 bits with a maximum sampling rate of 96 kHz and a 100 dB S/N ratio. The sounds were actually played at 22.05 or 44.1 KHz.

At the beginning of the experiment, the listener read the instructions and asked any necessary questions of the experimenter. Five test trials (chosen at random from the instruments) were presented before the data trials for each instrument.

IV. SPECTRAL CORRELATE MEASURES

Discrimination of spectral alterations may be affected by some aspect of the spectral variations of the original sounds. Several different measures of these variations are possible. Gray (1977) referred to ''spectral fluctuation'' as an interpretation of a dimension revealed by multidimensional scaling (MDS) of musical instrument dissimilarity judgments. ''Spectral flux'' was qualitatively described by Krumhansl (1989) as ''the temporal evolution of the spectral components'' and by McAdams *et al.* (1999) as ''the change in shape of a spectral envelope over time.'' Krimphoff (1993) defined *spectral variation* as the average correlation between adjacent time points in the spectrum:

$$VS = \frac{1}{N}\sum_{n=1}^{N}\frac{\Sigma_{k=1}^{K}A_k(t_{n-1})A(t_n)}{A_{rms}(t_{n-1})A_{rms}(t_n)}, \tag{8}$$

where

$$A_{rms}(t_n) = \sqrt{\sum_{k=1}^{K}A_k^2(t_n)} \tag{9}$$

is the instantaneous rms amplitude and N is the total number of frames. With the definition of Eq. (8), VS is theoretically between 0 and 1, but because spectral amplitudes vary slowly, VS is expected to be close to 1, with closeness to 1 being an indicator of how static the signal is. Instead of this measure, we opted to use a spectral incoherence measure as defined by Eqs. (11) and (12) below.

Krimphoff also defined ''flux'' as the root-mean-squared deviation of the normalized spectral centroid over time given by

$$FL = \sqrt{\frac{1}{N}\sum_{n=0}^{N-1}(NSC(t_n) - NSC_{xx})^2}, \tag{10}$$

where NSC_{xx} could be the average, rms, or maximum value of NSC. A time-average value was used in the current study.

Beauchamp and Lakatos (2002) measured spectral fluctuation in terms of *spectral incoherence*, a measure of how much a spectrum differs from a coherent version of itself. Larger incoherence values indicate a more dynamic spectrum, and smaller values indicate a more static spectrum. A perfectly static spectrum has an incoherence of zero.

The coherent spectrum is defined to have the same average spectrum as the original and the same instantaneous rms amplitude, but, unlike the original, all harmonic amplitudes vary in time proportional to the rms amplitude and, therefore, in fixed ratios to each other. Put another way, harmonic amplitudes normalized by the rms amplitude are fixed. The coherent version of the kth harmonic amplitude is defined by

$$\hat{A}_k(t_n) = \frac{\overline{A}_k A_{rms}(t_n)}{\sqrt{\Sigma_{k=1}^{K}\overline{A}_k^2}}, \tag{11}$$

where $\overline{A}_k$ is the time-averaged amplitude of the kth harmonic. Then, spectral incoherence of the original spectrum is defined as

$$SI = \left(\frac{\Sigma_{n=0}^{N-1}\Sigma_{k=1}^{K}(A_k(t_n) - \hat{A}_k(t_n))^2}{\Sigma_{n=0}^{N-1}(A_{rms}(t_n))^2}\right)^{1/2}. \tag{12}$$

With the definitions of Eqs. (11) and (12), spectral incoherence (SI) varies between 0 and 1 with higher values indicating more incoherence (more dynamic). Both VS and SI are measures of spectral fluctuation, but FL is a measure of the normalized centroid change over time. Since it is possible for FL to be significant while SI is relatively small (or VS is close to 1) and for SI to be large while FL is small, SI (or VS) and FL are approximately independent measures, although they may be correlated for a particular set of musical sounds.

Krimphoff (1993) also introduced the concept of *spectral irregularity*, which was redefined by Beauchamp and Lakatos (2002) as

$$SIR = \frac{1}{N}\sum_{n=0}^{N-1} \frac{\Sigma_{k=2}^{K-1} A_k(t_n)\left|A_k(t_n) - (A_{k-1}(t_n) + A_k(t_n) + A_{k-1}(t_n))/3\right|}{A_{rms}(t_n)\Sigma_{k=2}^{K-1} A_k(t_n)}. \quad (13)$$

This formula defines the difference between a spectrum and a spectrally smoothed version of it, averaged over both harmonics and time and normalized by rms amplitude. It may be hypothesized that spectra that are irregular or jagged to begin with are more likely to be confused by listeners with randomly altered versions of the spectra.

In general, the expectation is that the more complex an original signal is, either in terms of spectral fluctuation over time or in terms of spectral jaggedness, the less perceptible the effect of random alteration on a spectrum would be.

V. RESULTS

A. Data analysis methods

A test of normality indicated that the discrimination data were not normally distributed even after various transformations (e.g., sqrt[arcsin(data)]) (Stevens, 2002). Therefore, the data were analyzed using both parametric and nonparametric statistical tests (parametric: ANOVAs, Student Newman–Keuls tests; nonparametric: Friedman ANOVAs by ranks, Wiloxcon signed-ranks tests, Kruskal–Wallis tests, and Mann–Whitney U tests). The results of parametric and nonparametric tests were found to be similar for the overall main effects and two-way interaction effects, although they differed in some detailed analyses (e.g., statistical groupings of data using parametric Student Newman–Keuls tests were sometimes different from the results of Wilcoxon signed-ranks tests). In the rest of the paper, the main effects and the two-way interactions were tested with both ANOVAs and nonparametric tests. For detailed analyses of interactions, results of nonparametric tests are presented because the data were not normally distributed.

B. Effects and interactions of error level, instrument, and musical experience

Discrimination scores were computed for each random alteration across the four trial structures for each listener. Because the presentation order of each of the four trials was randomized, any potential effects of learning were averaged out. Figure 3 shows the scores averaged over the 20 listeners for the 50 random alterations (ten versions of each of five error levels) on the eight instruments. It can be observed that as the error level increases from 8% to 48%, the mean discrimination increases asymptotically towards 100%. For 8% error levels, most scores are in the range 50% to 60%. The ranges are wider and more variable for intermediate errors such as 16% error, where the scores are between 60% and 85% (except for the trumpet, which is about 10% lower). For 24%, 32%, and 48% error, most scores are between 80% and 95%, 88% and 100% and 92% and 100%, respectively.

Figure 4 shows a graph of average discrimination versus error level for each instrument. The scores are in close agreement, especially for the 8%, 32%, and 48% error cases. Even for 16% and 24% error levels, the average discriminations for the various instruments are within 12% of one another.

ANOVA analyses of the results used the ten different random alterations for each instrument as repeated measures to test the main effects of instrument (eight instruments), error level (five levels: 8%, 16%, 24%, 32%, 48%), musical experience (two groups: no musical experience, at least 1 year of musical experience) and their two-way interactions (see Table I).

Inspection of Table I indicates that both experience and instrument, and instrument and error level, have strong significant two-way interactions and experience and error level have weak significant interactions. Further analyses of the weak interactions between the effects of musical experience and the effects of error level indicate that for all 16 combinations of instrument and experience, the effects of error level are significant and the trends are the same ($p < 0.0001$, Friedman ANOVAs by ranks). This suggests that the effects of error level are strong and consistent for all conditions.

Mean discrimination scores for listeners with no experience in playing a musical instrument and listeners with at least one year of experience as functions of the five error levels are compared in Fig. 5. Mann-Whitney U tests were conducted to compare the discrimination scores between listeners with and without musical experience at each error level. Consistent with Fig. 5, no significant effects of experience were found when error levels were either 8% or 48% ($p > 0.05$). However, when error levels were 16%, 24%, and 32%, the scores of musically experienced listeners were found to be significantly higher than those of listeners with no musical experience ($p < 0.001$, Mann-Whitney U tests).

Since there were significant two-way interacting effects with experience and instrument, and instrument and error level, Friedman ANOVAs by rank and Wilcoxon signed-ranks tests were conducted to examine the interaction patterns between instrument and error level for data collected from listeners with and without musical experience. The results of these analyses are presented in Table II. Table II shows that when the error levels were 8% or 48%, instrument did not have a significant effect on discrimination. However, when the error levels were 16%, 24%, or 32%, instrument did have a significant effect. Also, for all listeners regardless of their musical experience, the violin and trumpet were consistently associated with lower discrimination scores, and the bassoon and horn were consistently associated with higher discrimination scores.

C. Correlation of spectral incoherence, spectral centroid change, and spectral irregularity with discrimination

Since discrimination varies significantly with instrument for intermediate error levels, we correlated the discrimination

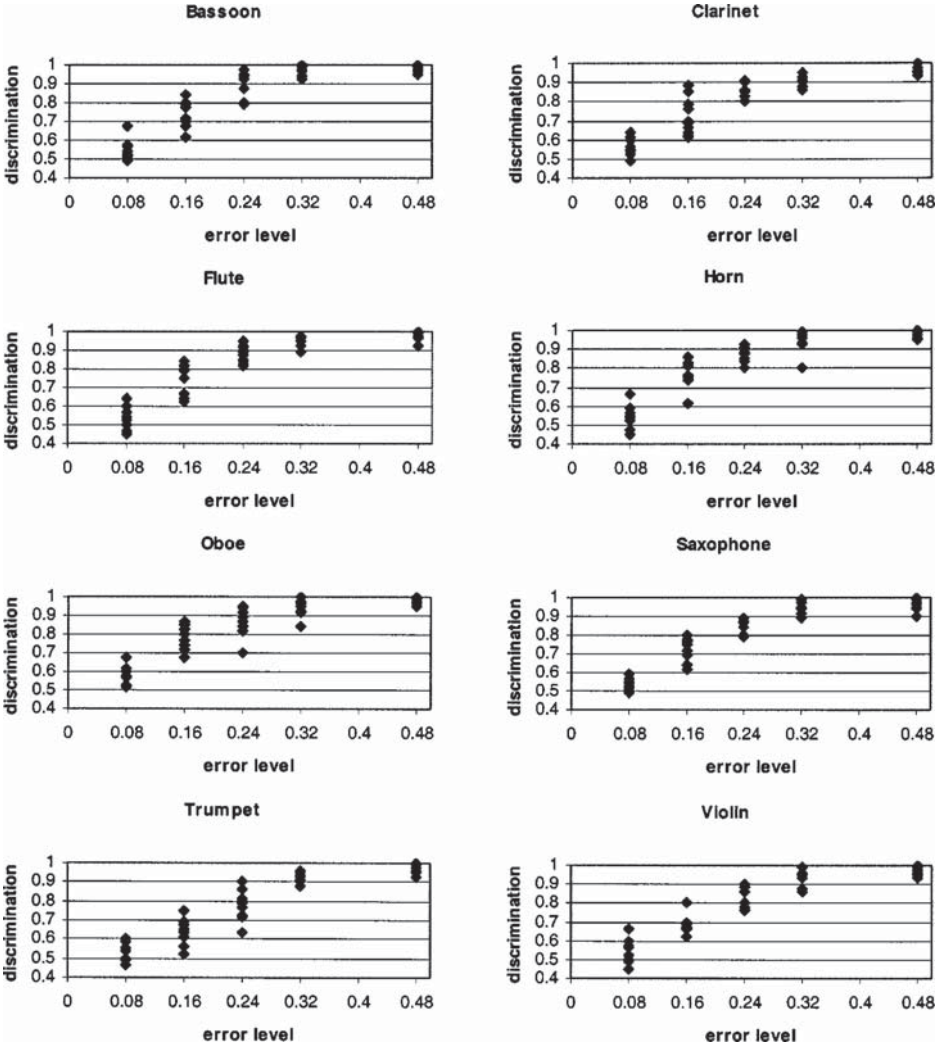

FIG. 3. Mean discrimination scores for ten randomly altered sounds versus error level for the eight instruments.

scores with spectral incoherence, normalized centroid deviation, and spectral irregularity to determine whether these different measures of spectral variation were significantly related to discrimination. Table III gives spectral incoherence, spectral centroid deviation, spectral irregularity, and average discrimination values for error levels of 16% and 24% for each of the eight instruments.

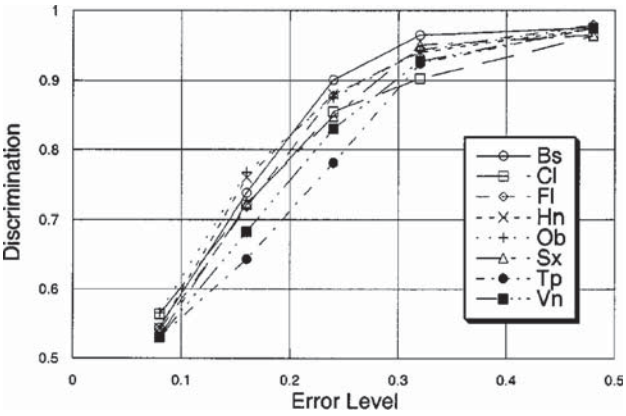

FIG. 4. Average discrimination scores versus error level for the eight instruments.

Table III shows that the trumpet and violin have the largest incoherence values, as well as the lowest discrimination scores. Conversely, the bassoon and horn have among the lowest spectral incoherence values and the highest discrimination scores. This suggests that spectral incoherence is related to difficulty in detecting random spectral alterations. When discrimination scores were averaged over the ten alterations and the 20 subjects (as shown in Fig. 4), strong negative correlations were found for the 16% and 24% error levels [$r(6) = -0.91$, $p < 0.005$ at 16%; $r(6) = -0.64$, $p < 0.1$ at 24%, Spearman correlation].

Correlations between discrimination scores and normalized centroid deviations were also calculated. Again, strong negative correlations were found [$r(6) = -0.72$, $p < 0.05$ at 16%; $r(6) = -0.85$, $p < 0.01$ at 24%, Spearman correlation]. On the other hand, the correlations between discrimination scores and spectral irregularity were relatively weak and not statistically significant [$r(6) = 0.12$, $p = 0.78$ at 16%; $r(6) = -0.17$, $p = 0.69$ at 24%, Spearman correlation].

VI. DISCUSSION

The results presented in the previous section show that, for the instruments tested, both spectral incoherence and

TABLE I. ANOVA table illustrating the main effects and two-way interactions of instrument (eight instruments), error level (8%, 16%, 24%, 32%, and 48% error), and musical experience (no experience versus 1 year or more) on the data collected from 20 listeners participating in the discrimination experiment. Data are the percentage of correct discrimination scores (100%, 75%, 50%, 25%, and 0%) over each group of four trials containing the same reference and altered sounds. For each error level, there were 40 trials with 10 different random spectral alterations in the test sounds.

Source	DF	Sum of squares	Mean square	F value	Pr>F
Instrument	7	40.56	5.59	10.69	0.0001[a]
Error level	4	3186.14	796.53	1469.79	0.0001[b]
Musical experience	1	15.61	15.61	28.80	0.0001[c]
Instrument and error level	28	42.03	1.50	2.77	0.0001[d]
Experience and instrument	7	16.14	2.31	4.26	0.0001[d]
Experience and error level	4	5.36	1.34	2.47	0.0425[e]
Measurement error	7948	4307.31	0.54		
Corrected total	7999	7613.14			

[a]The significant main effects were confirmed with nonparametric Friedman ANOVA by ranks: $p<0.0001$, chi-square=3337.4.
[b]The significant main effects were confirmed with nonparametric Friedman ANOVA by ranks: $p<0.0001$, chi-square=77.87.
[c]The significant main effects were confirmed with nonparametric Kruskal–Wallis test: $p<0.001$, chi-square =50.3.
[d]Nonparametric statistical tests also indicated a significant two-way interacting effect.
[e]Nonparametric statistical tests did not indicate a significant two-way interacting effect.

spectral centroid deviation are strongly correlated with difficulty of detecting random spectral alterations. While it seemed reasonable to expect that increased incoherence would make detection of random spectral alterations more difficult, it was not obvious that increased centroid deviation would have the same effect. Even though spectral incoherence and centroid deviation are theoretically quite independent, in our case they appear to be tightly correlated measures of time-variant spectral variation. On the other hand, spectral irregularity was expected to affect discrimination, but correlations between this parameter and discrimination were found to be small. For example, the trumpet and violin have similar spectral incoherences and centroid deviations and similar 16% and 24% error level discrimination scores, but the trumpet has a much lower spectral irregularity than the violin has. Time-varying spectral variations seem to be more important than "jaggedness" of a spectrum with respect to difficulty of detecting spectral alterations.

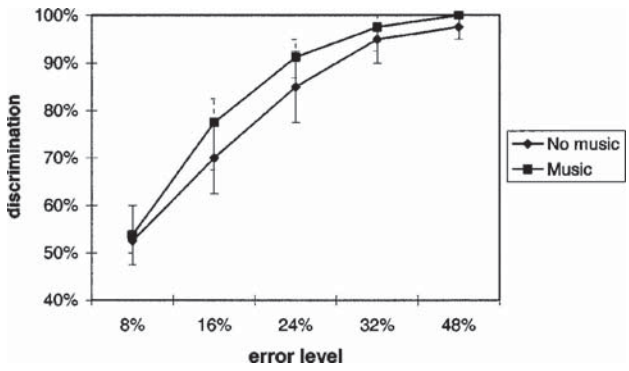

FIG. 5. Median of discrimination scores collected from listeners with and without musical experience as functions of error level (8%,16%,24%,32%,48%). Scores from the ten repeated runs per each condition have been averaged first before the median was calculated. The interquartile ranges are also shown.

From Fig. 4 it can be seen that the 16% error level roughly corresponds to a 75% discrimination threshold, where subtle differences between original and randomly altered sounds are sometimes detected. However, at the 24% error level, we have found (in informal listening tests) that random spectral alteration produces interesting collections of "similar, yet different" sounds compared to the originals. Random spectrum alteration could then provide an efficient time-invariant method for generating diverse sets of musical sounds with the same temporal behavior and spectral centroid. This method may possibly provide an improvement for existing synthesizers, which have often been criticized as sounding too tepid, presumably because repeated notes, played at the same amplitude, typically sound exactly the same. Knowing how to change sounds rapidly by random spectrum alteration so that the sounds sound subtly different from one note to the next may result in a process for production of more dynamic and interesting synthetic musical sound sequences.

The distribution of error among the harmonics may make a large difference in the perceptual effect of spectral modification. A previous spectral matching study using a method that often lumped most of the error into one or two of the most prominent harmonics (Horner, 2001) found 75% discrimination thresholds at about 8% error. This is approximately the profile analysis result found by Kidd *et al.* (1991), who observed that jagged spectra with a single altered partial have 75% discrimination thresholds at error levels roughly corresponding to 8%. The current study found that evenly distributing the error among all the partials appears to double the 75% discrimination threshold approximately to 16% error. This indicates that errors distributed over all the harmonics are more difficult to detect than those lumped into one or two of the most prominent harmonics.

TABLE II. This table presents the effects of instrument on the ability of listeners with and without musical experience to discriminate randomly altered tones with five levels of error. Within each column, the instruments (BS: bassoon; CL: clarinet; FL: flute; HN: horn; OB: oboe; SX: saxophone; TP: trumpet; VN: violin) are listed in descending order of discrimination scores collected from 20 listeners for the five error levels (8%, 16%, 24%, 32%, and 48%). The table shows the results of Wilcoxon signed-ranks tests and Friedman ANOVA tests, where instruments on the ranked lists associated with similar (i.e., not significantly different at $p=0.05$ level) data were grouped into the same group as represented by the same capital letter ("A" to "D").

Listeners with no musical experience

8%		16%				24%					32%				48%	
SX	A	OB	A			BS	A				SX	A			FL	A
BS	A	HN	A	B		HN	A	B			BS	A	B		OB	A
VN	A	SX	A	B		SX	A	B			FL	A	B		HN	A
OB	A	BS	A	B		OB	A	B	C		HN	A	B		SX	A
CL	A	FL		B		FL	A	B	C		OB	A	B	C	TP	A
FL	A	VN		B		CL		B	C		TP		B	C	VN	A
TP	A	CL		B	C	VN			C	D	VN			C	BS	A
HN	A	TP			C	TP				D	CL			C	CL	A

Listeners with musical experience

8%		16%					24%					32%				48%	
OB	A	HN	A				BS	A				BS	A			BS	A
CL	A	OB	A	B			HN	A				VN	A	B		VN	A
HN	A	CL	A	B			FL	A	B			OB	A	B	C	HN	A
FL	A	BS	A	B	C		CL	A	B			HN		B	C	FL	A
TP	A	FL	A	B	C		OB	A	B	C		FL		B	C	TP	A
BS	A	SX		B	C	D	VN		B	C	D	SX		B	C	CL	A
VN	A	VN			C	D	SX			C	D	TP		B	C	OB	A
SX	A	TP				D	TP				D	CL			C	SX	A

VII. CONCLUSIONS

The results of the current study show a significant monotonic relationship between discrimination and the amount of random spectrum alteration. The trend of increasing discrimination scores with increasing error level is consistent for all combinations of musical experience and instrument, although the discrimination scores are significantly higher for listeners with at least one year of musical experience when the errors are 16%, 24%, and 32%. Significant interacting effects between error level and instrument were found and investigated. When error levels were 16%, 24%, and 32%, musically experienced listeners discriminated significantly better than those with no experience and instrument had a significant effect.

In general, discrimination was poor for sounds with less than 16% error, the 75% discrimination threshold. Sounds with 16% and 24% random alteration yielded the widest range of discrimination, while sounds with 32% and 48% error levels were clearly and consistently distinguishable by most listeners. Random spectrum alteration leads to lower discrimination scores than other spectral modification methods, such as profile analysis, which do not evenly distribute the error among all harmonics.

Listeners had more difficulty discriminating alterations to instrument sounds containing more pronounced spectral variations. Average spectral incoherences and spectral centroid deviations were found to have a strong negative correlation with average discrimination scores. This suggests that dynamic spectral variations result in increased difficulty of detecting spectral alterations. Such a high correlation was not found for spectral irregularity, a measure of the "jaggedness" of a spectrum.

TABLE III. Spectral incoherence, normalized centroid deviation, spectral irregularity, and average 16% and 24% error level discrimination scores for the eight instruments.

Instrument	Spectral incoherence	Normalized centroid deviation	Spectral irregularity	16% error average discrimination	24% error average discrimination
HN	0.057	0.2	0.073	0.765	0.889
OB	0.069	0.8	0.137	0.761	0.871
BS	0.075	0.4	0.093	0.738	0.901
CL	0.085	0.7	0.174	0.721	0.855
SX	0.101	0.6	0.195	0.723	0.848
FL	0.118	0.6	0.129	0.718	0.878
TP	0.184	1.6	0.039	0.643	0.781
VN	0.193	1.4	0.131	0.683	0.830

ACKNOWLEDGMENTS

This work was supported in part by the Hong Kong Research Grants Council's CERG Project Nos. HKUST6194/02E and HKUST6167/03E. We would like to thank Francis Hing-Cheung Wong for running subjects in the control experiment, and Simon Cheuk-Wai Wun for writing the program for Intel personal computers used for the listening tests. Thanks to the anonymous reviewers for their careful and helpful comments.

Allen, J. B. (**1977**). "Short term spectral analysis, synthesis, and modification by discrete Fourier transform," IEEE Trans. Acoust., Speech, Signal Process. **25**, 235–238.

Beauchamp, J. W. (**1993**). "Unix workstation software for analysis, graphics, modifications, and synthesis of musical sounds," 94th Conv. Audio Eng. Soc. (Berlin), Audio Eng. Soc. Preprint 3479 (L-I-7).

Beauchamp, J. W., and Lakatos, S. (**2002**). "New Spectro-Temporal Measures of Musical Instrument Sounds Used for a Study of Timbral Similarity of Rise-Time- and Centroid-Normalized Musical Sounds," in Proc. 7th Int. Conf. on Music Perception and Cognition, Univ. of New South Wales, Sydney, Australia, pp. 592–595.

Bernstein, L. R., Richards, V. M., and Green, D. M. (**1987**). "The detection of spectral shape changes," in *Auditory Processing of Complex Sounds*, edited by W. Yost and C. Watson (Erlbaum, Hillsdale, NJ), pp. 6–15.

Buus, S. (**1985**). "Release from masking caused by envelope fluctuations," J. Acoust. Soc. Am. **78**, 1958–1965.

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

Green, D. M. (**1988**). *Profile Analysis: Auditory Intensity Discrimination* (Oxford U.P., New York).

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 modification 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.

Hall, J. W., Haggard, M. P., and Fernandes, M. A. (**1984**). "Detection in noise by spectrotemporal pattern analysis," J. Acoust. Soc. Am. **76**, 50–56.

Horner, A. (**2001**). "A Simplified Wavetable Matching Method Using Combinatorial Basis Spectra Selection," J. Audio Eng. Soc. **49**, 1060–1066.

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 Proc. 4th Int. Conf. Music, Perception and Cognition (Montreal), Faculty of Music, McGill University, pp. 91–95.

Kidd, Jr., G., Mason, C. R., Uchanski, R. M., Brantley, M. A., and Shah, P. (**1991**). "Evaluation of simple models of auditory profile analysis using random reference spectra," J. Acoust. Soc. Am. **90**, 1340–1354.

Krimphoff, J. (**1993**). "Analyse acoustique et perception du timbre," unpublished DEA thesis, Université du Maine, Le Mans, France.

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

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

Lakatos, S. (**2000**). "A common perceptual space for harmonic and percussive timbres," Percept. Psychophys. **62**, 1426–1439.

Lentz, J. J., and Richards, V. M. (**1998**). "The effects of amplitude perturbation and increasing numbers of components in profile analysis," J. Acoust. Soc. Am. **103**, 535–541.

McAdams, S., Beauchamp, J. W., and Meneguzzi, S. (**1999**). "Discrimination of musical instrument sounds resynthesized with simplified spectrotemporal parameters," J. Acoust. Soc. Am. **105**, 882–897.

Mendoza, L., Schultz, M. L., and Schulz, R. A. (**1996**). "Comodulation masking release as a function of masking noise-band temporal envelope similarity in normal hearing and cochlear impaired listeners," J. Acoust. Soc. Am. **99**, 2565.

Moore, B. C. J., Glasberg, B. R., and Baer, T. (**1997**). "A model for the prediction of thresholds, loudness and partial loudness," J. Audio Eng. Soc. **45**, 224–240.

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, Eliden, The Netherlands), pp. 405–408.

Stevens, J. (**2002**). *Applied Multivariate Statistics for the Social Sciences*, 4th ed. (Erlbaum, Hillsdale, NJ), p. 264.

Toole, F. E., and Olive, S. E. (**1988**). "The modification of timbre by resonances: perception and measurement," J. Audio Eng. Soc. **36**, 122–142.

Versfeld, N. J., and Houtsma, A. J. M. (**1991**). "Perception of spectral changes in multi-tone complexes," Q. J. Exp. Psychol. A **43**, 459–479.

Watkins, A. J. (**1991**). "Central auditory mechanisms of perceptual compensation for spectral-envelope distortion," J. Acoust. Soc. Am. **90**, 2942–2955.

Watkins, A. J., and Makin, S. J. (**1996**). "Effects of spectral contrast on perceptual compensation for spectral-envelope distortion," J. Acoust. Soc. Am. **99**, 3749–3757.

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