Detection of time-varying harmonic amplitude alterations due to spectral interpolations between musical instrument tones

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

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

Richard H. Y. So[c)]
Department of Industrial Engineering and Logistics Management, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong

(Received 18 April 2008; revised 16 October 2008; accepted 22 October 2008)

Gradated spectral interpolations between musical instrument tone pairs were used to investigate discrimination as a function of time-averaged spectral difference. All possible nonidentical pairs taken from a collection of eight musical instrument sounds consisting of bassoon, clarinet, flute, horn, oboe, saxophone, trumpet, and violin were tested. For each pair, several tones were generated with different balances between the primary and secondary instruments, where the balance was fixed across the duration of each tone. Among primary instruments it was found that changes to horn and clarinet timbres were most easily discriminable, while changes to saxophone and trumpet timbres were least discriminable. Among secondary instruments, the clarinet had the strongest effect on discrimination, whereas the bassoon had the least effect. For primary instruments, strong negative correlations were found between discrimination and their spectral incoherences, suggesting that the presence of dynamic spectral variations tends to increase the difficulty of detecting time-varying alterations such as spectral interpolation.
© *2009 Acoustical Society of America.* [DOI: 10.1121/1.3025916]

PACS number(s): 43.75.Cd, 43.66.Jh [DD] Pages: 492–502

I. INTRODUCTION

Musical instruments are recognizable even when the sound has been substantially altered by a spectrum equalizer or room acoustics. Previous work has shown that the detection of time-invariant spectral alterations in musical instrument tones is more difficult when the original tone has significant time-varying spectral fluctuations, while spectral "jaggedness" has relatively little effect (Horner *et al.*, 2004). Do the same results hold for time-variant spectral alterations such as spectral interpolation? Answering this question is the primary objective of this paper.

In small amounts, time-invariant spectral alterations can change a clarinet sound into other clarinetlike sounds. In larger amounts, they can change a clarinet sound into unknown synthetic sounds. But, due to their time-invariance, they cannot produce all possible clarinet sounds, or sounds from other instruments, such as the violin. An alternative alteration method that includes time-varying changes is spectral interpolation. Spectral interpolation combines two or more sounds to create a new sound with an intermediate spectrum. For example, a spectrally interpolated sound could be formed by resynthesizing a spectral mix consisting of 60% of a viola spectrum (primary) and 40% of a clarinet spectrum (secondary).

Some related previous studies have considered the effect of simple spectral alterations of static spectra, speech, and audio signals. Plomp (1970) considered the correspondence between an error metric and discrimination databased on static musical instrument and vowel spectra. He concluded that differences in timbre can be predicted well from spectral differences. Toole and Olive (1988) explored the effect of adding a single resonance to the spectra of noise, music recordings, and speech. They found that resonances are more easily heard in noise input signals than in speech and music input signals. Watkins (1991) and Watkins and Makin (1996) investigated the effect of spectral-envelope distortion on vowel sounds in speech, and found that the perception of a sound can be influenced by the sound that precedes it. This suggests that spectral variations might also have a significant influence on discrimination.

A related area of research is *spectral profile analysis* (Green, 1988), which 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. Most importantly, spectral profile analysis usu-

[a)]Electronic mail: horner@cs.ust.hk
[b)]Electronic mail: jwbeauch@uiuc.edu
[c)]Electronic mail: rhyso@ust.hk

Downloaded 12 Mar 2011 to 143.89.90.104. Redistribution subject to ASA license or copyright; see http://asadl.org/journals/doc/ASALIB-home/info/terms.jsp

ally only considers static spectra while the current study is concerned with time-varying alterations in dynamic spectra. In addition, spectral profile analysis has typically used log-spaced rather than harmonic-spaced components, which Versfeld and Houtsma (1991) found to produce very different results. Another difference is that profile analysis spectra are usually flat unlike acoustic instrument spectra, though some researchers have also used "perturbed" or "jagged" spectra (Kidd *et al.*, 1991) and noted that spectral jaggedness increases the threshold of detection (Lentz and Richards, 1998). Yet another basic difference is that profile analysis studies have usually only attempted to determine the threshold of discriminating a change in a single spectral component, or at most a few components. Bernstein *et al.* (1987) observed that thresholds for single-component changes cannot be used to accurately predict thresholds for multiple-component changes.

It is difficult to generalize profile analysis results to dynamic harmonic spectra, although the above results suggest that multiple-component changes to acoustic spectra are more difficult to detect than are single-component changes. Also, comodulation-masking-release results (Mendoza *et al.*, 1996) suggest that alterations to coherent spectra are easier to detect than are alterations to incoherent spectra. This suggests that time-varying alterations may be still more difficult to detect.

The most relevant previous work studied time-invariant (static) alteration of musical instrument spectra, where each harmonic amplitude was multiplied by a time-invariant random scalar (Horner *et al.*, 2004). This work found that listeners had more difficulty discriminating alterations to instrument sounds containing more pronounced spectral variations. Spectral incoherence (SI) and normalized centroid deviation (NCD) were both found to have strong negative correlations with discrimination scores. This suggested that dynamic spectral variations increase the difficulty of detecting spectral alterations. However, the same study found relatively low correlation for spectral irregularity (SIR), a measure of the jaggedness of a spectrum.

In a recent related study, Gunawan and Sen (2008) tested discrimination thresholds for perturbing musical instrument spectral envelopes by attenuating bands within them, as a function of center frequency and bandwidth. Perturbing center frequencies varied from 689 to 19 294 Hz, while bandwidths ranged from 1378 to 11 025 Hz, depending on the center frequency. Using tones of three musical instruments [each performed at E_4^b (~311.1 Hz)], they found that spectral sensitivity was governed by only the first few harmonics and sensitivity did not improve when extending the bandwidth any higher. On the other hand, sensitivity was found to decrease if changes were made only to the higher frequencies and continued to decrease as the bandwidth was widened. Note that this study, like the study of Horner *et al.* (2004), was based on static processing of the time-varying spectral envelope.

While the studies mentioned above have considered the effect of simple spectral alterations, relatively little work has addressed the effect of time-variant spectral alterations such as those created by spectral interpolation of dynamic musical instrument spectra. Grey (1975) studied the effect of instrument time-variant interpolation in his Ph.D. thesis, cross-fading pairs of instrument tones to create new hybrid tones. For each instrument pair 11 tones were presented to the listener with interpolation levels increasing in steps of 10% going from the primary instrument (0%) to the secondary instrument (100%). The interpolation level was fixed across the duration of each tone. The listener was told to identify the point at which they identified the initial appearance of the secondary instrument in the sound. Results indicated that listeners tended to delay the perception of the secondary instrument well beyond the 50% interpolation point, thus demonstrating a hysteresis effect in timbre perception.

Related to spectral interpolation is the issue of *blend* (Kendall and Carterette, 1993; Sandell, 1995), which is defined as whether the instruments fuse into a single composite timbre, segregate into distinct timbral entities, or fall somewhere in between the two extremes. Sandell (1995) investigated spectral centroid and other factors in determining blended instrument pairings and found that lower average values of both centroid and onset duration for a pair of tones led to increased blends, as did closeness in value for the two factors.

This study will investigate the listeners' ability to discriminate changes to the time-varying spectral amplitudes of musical sounds caused by different degrees of spectral interpolation. Are measures of spectral variation (SI and NCD) strongly correlated with discrimination in spectrally interpolated tones? Does spectral jaggedness (irregularity) have more of an effect in interpolated tones? We will address these questions. We will also determine which instruments are least and most affected by spectral interpolation. Conversely, we will determine which instruments have the least and most effect as secondary instruments in the interpolation pair.

Section II presents the techniques used for analysis and synthesis of the stimuli, followed by a discussion of the discrimination experiment in Sec. III. Section IV reviews various spectral correlate measures, such as spectral incoherence and irregularity. Finally, in Sec. V, we present the discrimination results in terms of the effect of instrument, interpolation level, and spectral variations.

II. STIMULUS PREPARATION

Eight sustained musical instrument sounds were selected as prototype signals for stimulus preparation. Sounds of a bassoon, clarinet, flute, horn, oboe, saxophone, trumpet, and violin, performed at E_4^b (~311.1 Hz), were used to represent the several Western wind and the bowed string instrument families. Five of the sounds were taken from the McGill University Master Samples recordings (CD version), two from the Prosonus Sound Library (bassoon and oboe), and one (trumpet) was recorded at the University of Illinois Urbana-Champaign School of Music. All sounds were recorded at 16 bits, 44.1 kHz. Except for the bassoon and horn, these sounds were also used by McAdams *et al.* (1999), who gave more details about their characteristics. All eight sounds were also used by Horner *et al.* (2004). After parameter equalization (see Sec. II A), each sound was used as a stimu-

Downloaded 12 Mar 2011 to 143.89.90.104. Redistribution subject to ASA license or copyright; see http://asadl.org/journals/doc/ASALIB-home/info/terms.jsp

lus, including attack, sustain, and decay segments of the sounds. Sounds were chosen to be representative of each instrument's timbre, in that typical expressive elements were included.

The sounds were first subjected to time-variant spectrum analysis using a computer-based phase-vocoder method (Beauchamp, 2007). This phase vocoder is different from most in that it allows a fixed analysis frequency (311.1 Hz) to be tuned to the estimated fundamental frequency of the input signal. Beauchamp (2007) and Horner *et al.* (2004) gave further details on the phase-vocoder analysis method. Briefly, the method uses a fast Fourier transform (FFT) with Hamming window whose duration (~6.4 ms) is set to approximately twice the period of the input signal in order to minimize leakage between harmonics. The FFT bin frequencies are thus integer multiples of approximately 155.5 Hz. However, only the even multiples, corresponding to the input signal's harmonics, are retained. Prior to FFT analysis, the signal is up-sampled in order to create a power-of-2 number of samples for the window.

A. Temporal and loudness equalization and frequency flattening

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. Briefly, this was done by first identifying the attack and decay portions of each sound and then cross-fading the beginning of interior of the sound (using a cubic spline cross-fade function) with a later portion of the sound so as to reduce the duration to 2 s. This process was performed on the analysis data prior to resynthesis. Details are given by McAdams *et al.* (1999).

Attack and decay duration are also potential factors in discrimination. For example, if a tone with a short attack is interpolated with a tone with a long attack, the listener may be able to detect a difference compared to the original tone simply based on the longer attack time. In order that attack and decay times would not be factors in the study, the sounds were standardized to 0.05 s attack times and 0.15 s decay times through interpolation of the analysis data during these segments. Most attack and decay times were very close to the standardization values. The authors noted no significant perceptual differences between the standardized tones and the original 2 s tones. Where minor audible differences did occur, the standardized tone was judged as realistic as the original.

Next, amplitude multipliers were determined by a loudness program (Moore *et al.*, 1997) in order that each sound had a loudness of 87.4 phons. An iterative procedure adjusted the amplitude multiplier starting from a value of 1.0 until the resulting phons were within 0.1 phons of 87.4 [which corresponded to the trumpet sound played through headphones at 78 dB sound pressure level (SPL)].

In addition, frequency variations and inharmonicity were eliminated from the sounds in order that they would not be factors in this study. This was done by setting each harmonic's frequency equal to the product of its harmonic number and the fixed analysis frequency (311.1 Hz) in the analysis data prior to resynthesis, resulting in flat, equally spaced frequency versus time envelopes.

Sounds produced by the equalization and frequency flattening methods described above are referred to as *reference sounds* from here on in this paper.

B. Spectral interpolation and resynthesis

Spectral interpolation was performed on the analysis data of each pair of instruments, after which the sounds were regenerated by additive synthesis. Interpolation was accomplished in the frequency domain by calculating the weighted sum of the instrument pairs' harmonic amplitudes as follows:

$$A'_k(t_n) = (1 - l)A_{p,k}(t_n) + lA_{s,k}(t_n), \quad (1)$$

where $A'_k(t_n)$, $A_{p,k}(t_n)$, and $A_{s,k}(t_n)$ are the kth harmonic (linear) amplitudes of the interpolated, primary, and secondary instrument tones at time t_n, t_n is the time corresponding to the nth frame (i.e., $t_n = n\Delta t$ and $\Delta t = 6.4$ ms), and $0 < \ell \leq 0.5$ is the interpolation level. For example, if we wanted an interpolation of 60% viola with 40% clarinet, the viola would be the primary instrument, the clarinet would be the secondary instrument, and the interpolation level ℓ would be 0.4. The interpolation level remains fixed across the duration of the tone. It should be noted that since the instantaneous phases and frequencies corresponding to the primary and secondary instruments are equal, this method is equivalent to *mixing* frequency-flattened versions of the equalized tones in the time domain.

Interpolated sounds were synthesized for each pair of the eight instruments for interpolation levels of 5%, 10%, 15%, 20%, 25%, 30%, 40%, and 50%, yielding a total of 56 interpolated sounds for each primary instrument. These increments were chosen to provide adequate resolution of the interpolation process. However, going beyond 50% is not necessary because doing so reverses the role of primary and secondary instruments.

Both the reference sounds and the interpolated sounds were resynthesized by additive synthesis 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 (HKUST), ranging in age from 19 to 23 years, who reported no hearing problems. The listeners were paid to compensate for their time spent in the experiment.

B. Design of experiment

The experiment uses an unbalanced factorial design with three independent factors: primary instrument (eight levels: bassoon, clarinet, flute, horn, oboe, saxophone, trumpet, and

Downloaded 12 Mar 2011 to 143.89.90.104. Redistribution subject to ASA license or copyright; see http://asadl.org/journals/doc/ASALIB-home/info/terms.jsp

violin), secondary instrument (eight levels: bassoon, clarinet, flute, horn, oboe, saxophone, trumpet, and violin), and interpolation level (eight levels: 5%, 10%, 15%, 20%, 25%, 30%, 40%, and 50%). The participant is the random factor (20 levels), and the data collected from the 20 participants are used as repeated measures. In a full factorial design, the primary and secondary instruments can be the same. However, this may cause confusion because the resulting interpolated sound will then be the same as the uninterpolated reference sound, regardless of the interpolation level. To avoid this confusion, the authors decided not to present conditions where the primary and secondary instruments are the same. In summary, each participant will take part in 448 ($8\times7\times8=448$) conditions (or 8 blocks of 56 conditions). An unbalanced general linear model is used in the analysis of variance (ANOVA) analyses and type III sums of squares are used to calculate the F-values (see Sec. V).

C. Test procedure

Within each of the 448 conditions, a two-alternative forced-choice 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* represents one of the 56 interpolated sounds. (Note that the reference sound is always the primary sound.) 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 interpolated 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 using a keyboard. The computer would not accept a response until at least the first pair had been played. A custom program written at HKUST ran on an Intel PC to control the experiment.

For each instrument, a block of 224 trials was presented to each of the listeners (four trial structures $\times$ 56 interpolated sounds). The order of presentation of these 224 trials was randomized. For each interpolation, 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 224 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 45 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 50 min (including breaks). The total duration of the experiment was about 6.5 h.

Listeners were seated in a "quiet room" with less than 40 dB SPL background noise level (mostly due to computers and air conditioning). Sound signals were converted to analog by a SoundBlaster Audigy soundcard and then presented through Sony MDR-7506 headphones at 87.4 phons. The Audigy DAC utilized 24 bits with a maximum sampling rate of 96 kHz and a 100 dB signal-to-noise ratio. The sounds were actually played at 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

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 (Grey and Gordon, 1978; Wessel, 1979; Krumhansl, 1989; Iverson and Krumhansl, 1993; Krimphoff *et al.*, 1994; Kendall and Carterette, 1996; Lakatos, 2000). The time-varying normalized spectral centroid based on a sound's harmonic amplitudes, $\{A_k(t_n)\}$, is defined as

$$\mathrm{NSC}(t_n)=\frac{\sum_{k=1}^{K}kA_k(t_n)}{\sum_{k=1}^{K}A_k(t_n)}, \tag{2}$$

where K is the number of harmonics and $\mathrm{NSC}(t_n)$ can be thought of as the amplitude-averaged harmonic number at time t_n. It is "normalized" because spectral centroid is frequently given in frequency units, and, in this case, frequency has been normalized out of the definition by division by the fundamental frequency (311.1 Hz). Discrimination of spectral alterations may be affected by some aspect of the spectral variations of the reference sounds. Several different measures of these variations are possible. Grey (1977) referred to "spectral fluctuation" as an interpretation of a dimension revealed by 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 "flux" as the root-mean-squared deviation of the normalized spectral centroid over time. We call it NCD, which is given by

$$\mathrm{NCD}=\sqrt{\frac{1}{N}\sum_{n=0}^{N-1}(\mathrm{NSC}(t_n)-\mathrm{NSC}_{xx})^2}, \tag{3}$$

where NSC_{xx} could be the time-averaged, rms, or maximum value of NSC. Time-averaged measurements of NCD were used in the current study.

Beauchamp and Lakatos (2002) measured spectral fluctuation in terms of SI, 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.

We define the coherent version of a time-varying spectrum to be one that has the same average spectrum and the same instantaneous rms amplitude as the reference sound, but unlike the reference, all harmonic amplitudes vary in time proportional to the rms amplitude and, therefore, in

Downloaded 12 Mar 2011 to 143.89.90.104. Redistribution subject to ASA license or copyright; see http://asadl.org/journals/doc/ASALIB-home/info/terms.jsp

TABLE I. ANOVA table illustrating the main effects and two-way interactions of primary instrument (eight primary instruments), secondary instrument (eight secondary instruments), and interpolation level (eight levels: 5%, 10%, 15%, 20%, 25%, 30%, 40%, and 50%) 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 interpolation level, there were 28 trials (four repetitions with seven different secondary instrument interpolations in the test sounds). Because the primary and secondary instruments are never the same, an unbalanced ANOVA was conducted. The *p*-values generated from ANOVAs for all the main effects and two-way interactions are the same when using type I sum of squares and type III sum of squares. The ANOVAs using type III sum of squares are shown (Johnson and Wichern, 1992).

Source	DF	Sum of squares	Mean square	*F*-value	Pr>*F*
Primary (P) instrument	7	19.75	2.82	73.51	<0.0001 [a]
Secondary (S) instrument	7	13.83	1.97	51.49	<0.0001 [a]
Interpolation level	7	102.75	14.68	382.49	<0.0001 [a]
P-instrument and S-instrument	41	12.52	0.31	7.96	<0.0001 [a]
P-instrument and interpolation level	49	5.31	0.11	2.83	<0.0001 [a]
S-instrument and interpolation level	49	6.04	0.12	3.21	<0.0001 [a]
Measurement error	8799	337.66	0.04		
Corrected total	8959	398.40			

[a]*p*-values are consistent with the results of nonparametric statistical tests.

fixed ratios to each other. Put another way, the coherent spectrum's harmonic amplitudes normalized by the rms amplitude are fixed.

The coherent version of the kth harmonic amplitude is defined by

$$\hat{A}_k(t_n) = \frac{\bar{A}_k A_{\text{rms}}(t_n)}{\sqrt{\sum_{k=1}^{K} \bar{A}_k^2}}, \tag{4}$$

where $\bar{A}_k$ is the time-averaged amplitude of the kth harmonic and

$$A_{\text{rms}}(t_n) = \sqrt{\sum_{k=1}^{K} A_k^2(t_n)} \tag{5}$$

is the instantaneous rms amplitude.

Then, the SI of the reference spectrum is defined as

$$\text{SI} = \left(\frac{\sum_{n=0}^{N-1} \sum_{k=1}^{K} (A_k(t_n) - \hat{A}_k(t_n))^2}{\sum_{n=0}^{N-1} (A_{\text{rms}}(t_n))^2} \right)^{1/2}, \tag{6}$$

where N is the sound's total number of frames.

With the definitions of Eqs. (4) and (6), SI varies between 0 and 1 with higher values indicating more incoherence (more dynamic). SI is a measure of spectral fluctuation, while NCD is a measure of the normalized centroid change over time. Since it is possible for NCD to be significant while SI is relatively small and for SI to be large while NCD is small, SI and NCD are approximately independent measures, although they may be correlated for a particular set of musical sounds.

Krimphoff (1993) also introduced the concept of SIR, which was defined by Beauchamp and Lakatos (2002) as

$$\text{SIR} = \frac{1}{N} \sum_{n=0}^{N-1} \times \frac{\sum_{k=2}^{K-1} A_k(t_n) |A_k(t_n) - (A_{k-1}(t_n) + A_k(t_n) + A_{k-1}(t_n))/3|}{A_{\text{rms}}(t_n) \sum_{k=2}^{K-1} A_k(t_n)}. \tag{7}$$

This formula defines the average difference between a spectrum and a spectrally smoothed version of it. First, for each frame, it is amplitude-averaged over harmonics, then it is normalized by rms amplitude, and finally it is time-averaged over all of the frames.

V. RESULTS

This section analyzes the discrimination data with respect to interpolation level and *relative-amplitude spectral error* (Horner *et al.*, 2004, 2006). The latter error measure (also known as *relative spectral error*) is necessary since interpolations between spectrally similar instruments such as the violin and viola are much more difficult to detect than interpolations between very different instruments such as the violin and bassoon. Therefore, interpolation level is not a good independent predictor of discrimination, and an alternative predictor such as relative-amplitude spectral error is needed to explain the results. This error is given by the formula

$$\varepsilon = \frac{1}{N} \sum_{n=0}^{N-1} |A_k(t_n) - A'_k(t_n)|, \tag{8}$$

where $A_k(t_n)$ and $A'_k(t_n)$ are the kth harmonic amplitudes of the reference and interpolated tones, respectively.

A. Data analysis methods

A test of normality indicated that the discrimination data were not normally distributed even after various transforma-

Downloaded 12 Mar 2011 to 143.89.90.104. Redistribution subject to ASA license or copyright; see http://asadl.org/journals/doc/ASALIB-home/info/terms.jsp

TABLE II. The effects of primary instrument on the ability of listeners to discriminate interpolated tones with nearest-neighbor error levels (5%, 10%, 15%, 20%, 25%, 30%, 40%, and 50%). Within each column, primary instruments with the lowest discrimination scores are listed at the top, and those with the highest discrimination scores are listed at the bottom. The ranking is based on the results of Student–Newman–Keuls tests, and is consistent with the results of nonparametric tests. When two or more primary instruments produced similar discrimination scores (i.e., not significantly different at the $p=0.05$ level), they are grouped and labeled with the same capital letter (A–D). Only the group with the smallest (labeled as A) and the largest discrimination scores (labeled as either B or C or D) are shown.

	Secondary instruments															
	Bs		Cl		Hn		Sx		Fl		Ob		Tp		Vn	
Primary instruments	Sx	A	Sx	A	Sx	A	Vn	A	Tp	A	Tp	A	Sx	A	Sx	A
	Tp		Ob	A	Bs	A	Fl	A	Sx	A	Vn	A	Vn	A	Cl	A
	Fl		Fl		Tp	A	Tp	A	Bs	A	Sx	A	Ob		Fl	A
	Vn		Vn		Vn		Cl	B	Ob		Fl		Cl		Tp	
	Ob		Tp		Ob		Bs	B	Vn		Cl		Fl		Ob	
	Cl	D	Bs	D	Fl		Ob	B	Hn		Bs		Bs		Bs	D
	Hn	D	Hn	D	Cl	C	Hn	B	Cl	D	Hn	D	Hn	D	Hn	D

tions (e.g., sqrt[arcsin(data)]) (Stevens, 2002). Therefore, the data were analyzed using both parametric and nonparametric statistical tests (parametric: ANOVA, Student–Newman–Keuls tests; nonparametric: Friedman ANOVA by rank). The results of parametric and nonparametric tests were found to be consistent. In the rest of the paper, the main effects and the two-way interactions were tested with both ANOVA and nonparametric tests.

B. Effects and interactions of interpolation level and instrument

Discrimination scores, given in terms of percent correct, were computed for each interpolation 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. Table I shows the results of the ANOVA studying the main effects of interpolation level, primary instrument, secondary instrument, and their two-way interactions effects. Inspection of Table I indicates that all main effects and interactions are significant ($p<0.001$). These significant results are also consistent with the results of nonparametric analyses (Kruskal–Wallis H tests, Mann–Whitney U, and Friedman two-way ANOVA were tested).

The interactions among the effects of primary instrument and secondary instruments have been studied using Friedman ANOVAs, and the results are shown in Tables II and III.

Table II shows that the saxophone (Sx) as a primary instrument resulted in the significantly lowest discrimination scores when mixed with all other instruments. This means that when other instruments are mixed with the saxophone, listeners have trouble detecting any difference. On the other hand, the horn (Hn) as a primary instrument resulted in the significantly highest discrimination scores when mixed with the other instruments (the only exception was with the flute). This suggests that when other tones are mixed with the horn, it is relatively easy to detect the difference.

Table III shows that the bassoon (Bs) as a secondary instrument resulted in the significantly lowest discrimination scores when mixed with the other instruments (the only exception was with the clarinet). On the other hand, the clarinet (Cl) as a secondary instrument resulted in the significantly highest discrimination scores when mixed with the other instruments (the only exception was with the oboe). This suggests that the bassoon easily blends with other instrument tones, while the clarinet tends to stand out.

TABLE III. The effects of secondary instrument on the ability of listeners to discriminate interpolated tones with nearest-neighbor error levels (5%, 10%, 15%, 20%, 25%, 30%, 40%, and 50%). Within each column, secondary instruments with the lowest discrimination scores are listed at the top, and those with the highest discrimination scores are listed at the bottom. The ranking is based on the results of Student–Newman–Keuls tests and is consistent with the results of nonparametric tests. When two or more secondary instruments produced similar discrimination scores (i.e., not significantly different at the $p=0.05$ level), they are grouped and labeled with the same capital letter (A–E). Only the group with the smallest (labeled as A) and the largest discrimination scores (labeled at either C or D or E) are shown.

	Primary instruments															
	Bs		Cl		Hn		Sx		Fl		Ob		Tp		Vn	
Secondary instruments	Fl	A	Vn	A	Bs	A	Bs	A	Bs	A	Bs	A	Bs	A	Tp	A
	Hn	A	Ob	A	Fl	A	Tp		Ob		Tp		Ob	A	Ob	A
	Ob	A	Tp		Ob	A	Hn		Vn		Vn		Fl		Bs	A
	Vn		Bs		Vn		Vn		Sx		Cl		Hn		Sx	
	Tp	C	Sx		Sx		Ob		Tp	D	Fl		Vn		Hn	C
	Sx	C	Fl	E	Tp		Fl		Cl	D	Hn		Sx		Fl	C
	Cl	C	Hn	E	Cl	D	Cl	C	Hn	D	Sx	C	Cl	D	Cl	C

Downloaded 12 Mar 2011 to 143.89.90.104. Redistribution subject to ASA license or copyright; see http://asadl.org/journals/doc/ASALIB-home/info/terms.jsp

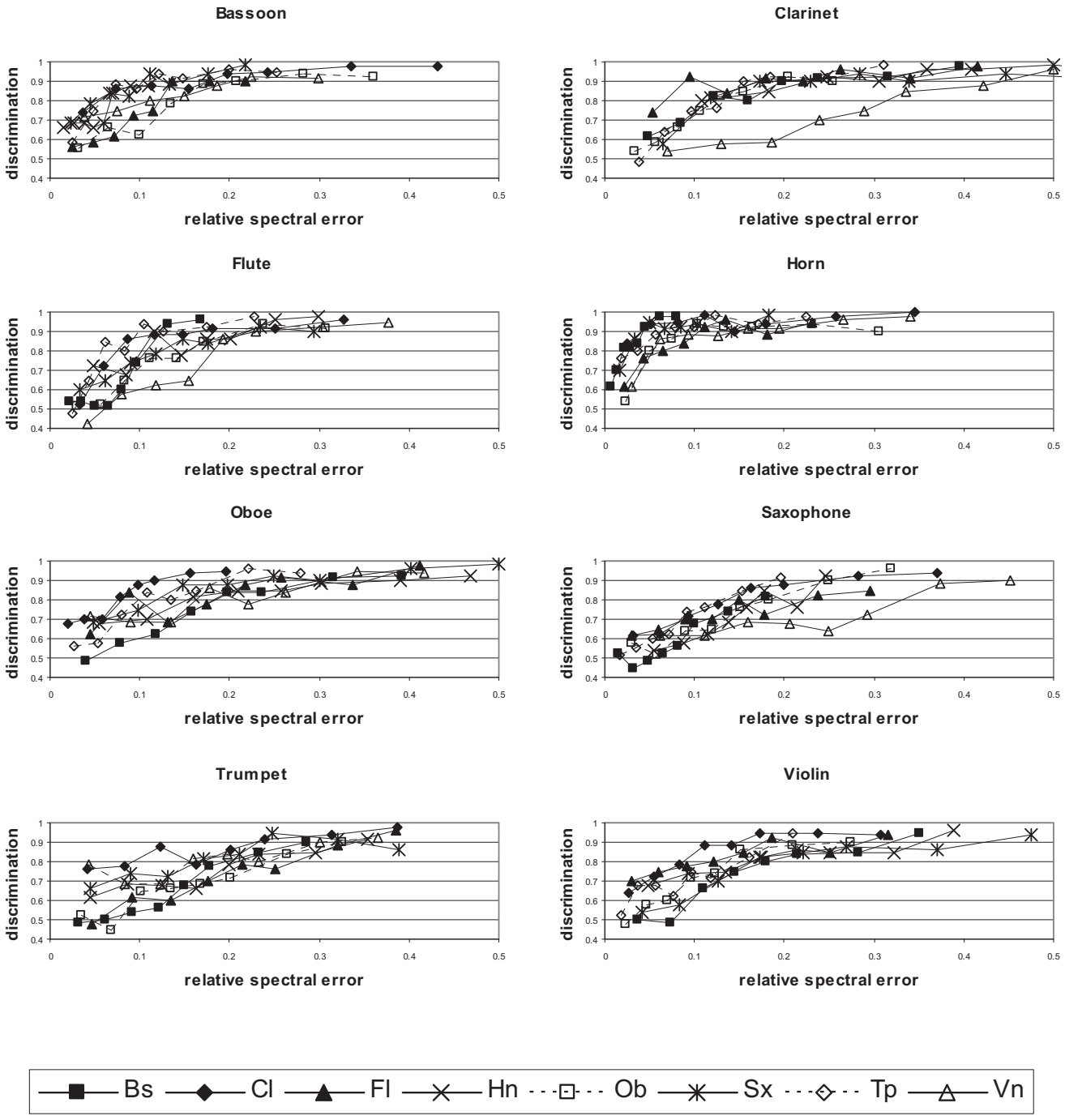

FIG. 1. Mean discrimination scores for the eight primary instruments. Each line of points corresponds to interpolations with one of the other seven secondary instruments (some outlying lines are labeled).

C. Use of relative-amplitude spectral error to explain the results

Although the interpolation levels have significant effects on discrimination, these effects have significant interactions both with the effects of primary instrument and secondary instrument (see Tables II and III). Intuitively, this is easy to understand since interpolations between spectrally similar instruments such as the violin and viola are much more difficult to detect than interpolations between very different instruments such as the violin and saxophone. These differences are better reflected in spectral error metrics such as relative-amplitude spectral error [see Eq. (8)], which is a more "universal" measure than spectral interpolation because it works for time-variant cases other than spectral interpolation. Recent previous work has shown that relative-amplitude spectral error captures about 90% of the variance in discrimination data for instruments similar to those presented in this paper (Horner *et al.*, 2006).

Therefore, relative spectral errors have been calculated for each combination of primary instrument, secondary instrument, and interpolation level. Tests of correlation indicate that the calculated relative spectral errors are significantly

Downloaded 12 Mar 2011 to 143.89.90.104. Redistribution subject to ASA license or copyright; see http://asadl.org/journals/doc/ASALIB-home/info/terms.jsp

correlated with interpolation level ($p<0.001$, Spearman correlation test) as well as discrimination ($p<0.001$, Spearman). Results of regression analyses indicate that $\log_{10}$ (relative error) can explain 48% of the data variance in the mean discrimination scores of the 20 listeners ($R^2=0.489$). Because relative spectral error accounts for the relative difference between two interpolated tones, it is a better predictor variable than interpolation level. In the subsequent analyses, discrimination scores will be examined as a function of relative spectral error.

For all instruments, larger interpolation levels always result in larger relative spectral error levels. It can also be trivially observed that as the relative spectral level increases, the mean discrimination increases asymptotically to between 90% and 100%. Figure 1 shows discrimination scores averaged over the 20 listeners on the eight primary instruments, where each line of points corresponds to interpolations with one of the other seven secondary instruments.

For example, the graph labeled "clarinet" shows its interpolation with the other seven instruments, and the "Vn" line corresponds to the clarinet-violin interpolations. In particular, the leftmost point on the Vn line (at 7% relative spectral error) is an interpolation with 95% clarinet and 5% violin. Moving rightward, the next point along the line (at 13% error) is an interpolation with 90% clarinet and 10% violin. The points farther to the right on this line increase the violin to 15%, 20% 25%, 30%, 40%, and 50%, and decrease the clarinet correspondingly downward.

Interestingly, the horn shows a consistently faster convergence to a ceiling value than the other primary instruments in Fig. 1 no matter which secondary instrument it is interpolated with. The slowest convergence occurs when the clarinet is the primary instrument and the violin is the secondary instrument—a pronounced outlier among the curves. The other curves in the clarinet graph are very similar (with the exception of the two leftmost flute points), with very little deviation. Perhaps this is due to the prominence of the odd harmonics in the clarinet, which results in fairly uniform discrimination changes.

Figure 2 shows a graph of average discrimination as a function of relative spectral error for each primary instrument, where points have been grouped into error bins of 5% (e.g., 5%, 10%, 15%, 20%, etc.) using nearest-neighbor grouping. Bins with only one element have been dropped from the figure. The curves vary considerably from instrument to instrument, with a range of 25% for error levels up to 15%, and a range of 15% for error levels between 20% and 30%.

Inspection of Fig. 2 indicates that the horn has significantly higher average discrimination scores than the other primary instruments ($p<0.05$, Student–Newman–Keuls tests), meaning that listeners found it relatively easy to detect interpolations to the horn, in agreement with the results in Table II. The bassoon also had significantly higher discrimination scores ($p<0.05$, Student–Newman–Keuls tests). These two instruments tend to blend or fuse with others in a smooth way, allowing other instruments to transparently shine through them quite readily. The saxophone and trumpet have significantly lower average discrimination scores than the other primary instruments ($p<0.05$, Student–Newman–Keuls tests), which basically agrees with the results in Table II. These instruments tend to be more opaque and dominate other instruments, making it more difficult to detect interpolations to them.

These observations are also in agreement with the advice given by Rimsky-Korsakov (1964) in his classic orchestration book, where relative strengths of the instruments are given for the various instruments in order to balance them when playing at the same dynamic level. While the other woodwinds are given a weight of 1, Rimsky-Korsakov (1964) assigned the saxophone a weight of 3 and the trumpet a weight of 4 reflecting their relative strength and ability to project.

Figure 3 displays the same data as Fig. 1, but it is arranged by secondary instrument. For example, in the bassoon graph, the "Hn" line represents interpolations with the bassoon as a secondary instrument and the horn as a primary instrument. The other lines in the bassoon graph represent interpolations with the other primary instruments (relative spectral error is given with respect to each of the primary instruments).

Figure 4 shows a graph of average discrimination for each secondary instrument. There is not as much variation between the instruments as in Fig. 1. The clarinet has significantly higher average discrimination scores than the other primary instruments ($p<0.05$, Student–Newman–Keuls tests), meaning that listeners found it relatively easy to detect the prominent odd harmonics of the clarinet when they were interpolated with the primary instrument. This agrees with the results in Table III. The trumpet is also significantly higher ($p<0.05$, Student–Newman–Keuls tests). On the other hand, the violin has significantly lower average discrimination scores than the other primary instruments ($p<0.05$, Student–Newman–Keuls tests), though there is considerable variation among the violin scores. This differs from the results in Table III, which found the bassoon to be the lowest with respect to interpolation level. The unusually

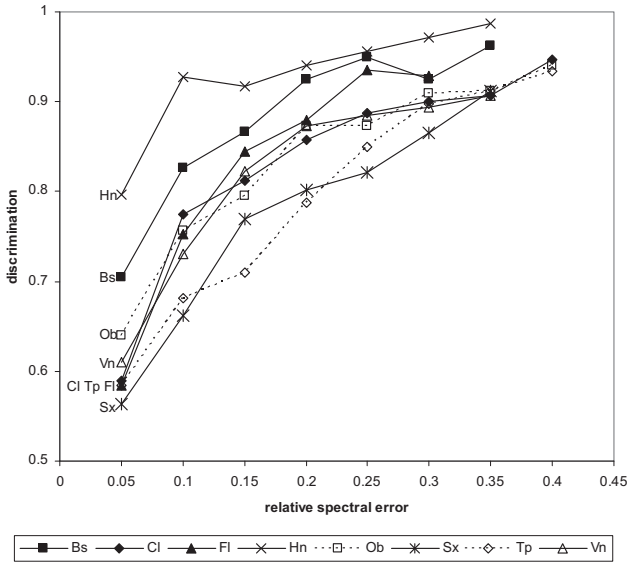

FIG. 2. Average discrimination scores for the eight primary instruments.

Downloaded 12 Mar 2011 to 143.89.90.104. Redistribution subject to ASA license or copyright; see http://asadl.org/journals/doc/ASALIB-home/info/terms.jsp

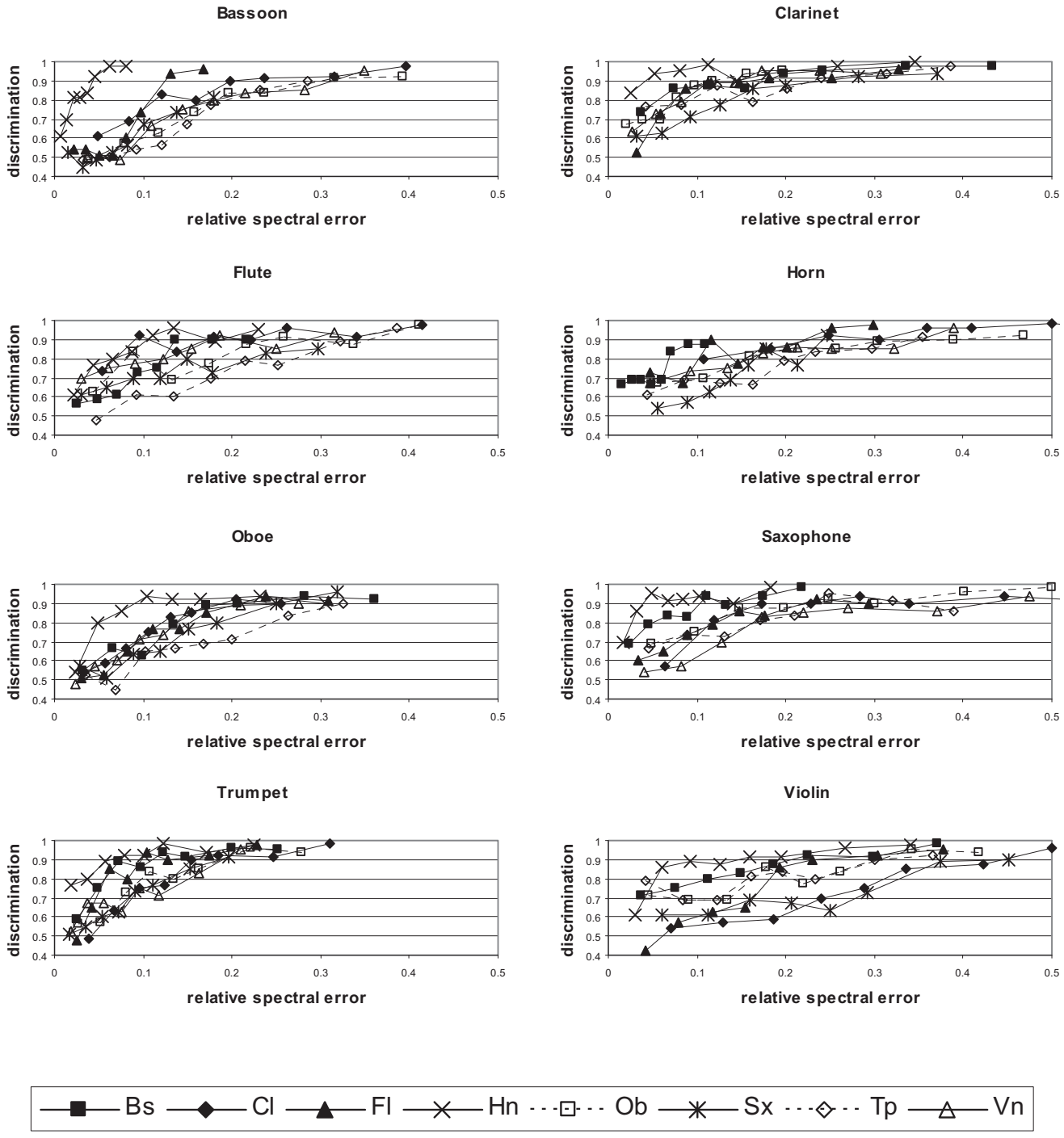

FIG. 3. Mean discrimination scores for the eight secondary instruments. Each line of points corresponds to interpolations with one of the other seven primary instruments (some outlying lines are labeled).

strong variations probably account for the difference. The saxophone also exhibits some unusual behavior, having among the highest discrimination scores for low-error levels and the lowest scores for high-error levels.

D. Correlation of spectral incoherence, normalized centroid deviation, and spectral irregularity with discrimination

Discrimination scores were correlated with physical measures SI, NCD, and SIR [see Eqs. (3), (6), and (7)] of the reference sounds to determine whether these measures of spectral variation were significantly related to discrimination. Table IV gives SI, spectral centroid deviation, SIR, and average discrimination scores for error levels of 10%, 15%, and 20% for each of the eight primary instruments. Table V gives the same values for each of the eight secondary instruments.

For primary instruments, strong negative correlations were found between the discrimination scores and SI with 10%, 15%, and 20% error levels ($r(6)=-0.76$, $p<0.05$ at 10%; $r(6)=-0.74$, $p<0.05$ at 15%; and $r(6)=-0.74$, p

Downloaded 12 Mar 2011 to 143.89.90.104. Redistribution subject to ASA license or copyright; see http://asadl.org/journals/doc/ASALIB-home/info/terms.jsp

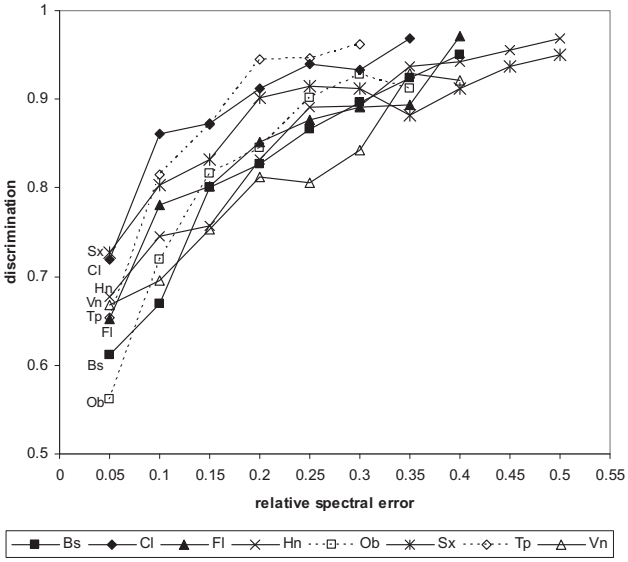

FIG. 4. Average discrimination scores for the eight secondary instruments.

<0.05 at 20%, Spearman correlation). No statistically reliable correlation relationship was found for NCD or SIR. For secondary instruments, no statistically reliable correlation relationship was found for SI, NCD, or SIR.

VI. DISCUSSION

The results presented in Sec. V show that, for the instruments tested, SI is strongly correlated with difficulty of detecting spectral interpolations. It seemed reasonable to expect that increased incoherence would make detection of a time-varying spectral change such as spectral interpolation more difficult, especially since our previous work had found that SI was strongly correlated with difficulty of detecting static random spectral alterations.

On the other hand, based on our previous work, we also expected NCD to affect discrimination, but correlations between this parameter and discrimination were found to be not significant. Though SI and centroid deviation are theoretically quite independent, in our previous work they appeared to be tightly correlated measures of time-variant spectral variation. Apparently, centroid deviations in the primary instrument have little affect on the detection of spectral interpolations, perhaps because such deviations would be somewhat averaged out by interpolation with the secondary instrument. There is no such averaging in random spectral alteration, perhaps accounting for the difference.

Both our previous and current studies found SIR to have no significant correlation to discrimination. Time-varying spectral variations, especially those of SI, seem to be more important than jaggedness of a spectrum with respect to difficulty of detecting spectral alterations and interpolations.

The results of this study can potentially lead to methods of estimating the perception of timbral difference based on spectral difference. For the spectral interpolation case, relative spectral error is by no means a perfect predictor of discrimination. R^2 correspondence was about 48% for this case as opposed to 90% found for static spectral variations (Horner *et al.*, 2006). In the future, it would be interesting to compare different measures of spectral distance in terms of their correlations with discrimination results as well as to results of listener judgments of timbral distance. Higher correspondences may be obtained from timbral distance judgments rather than discrimination, and if so, in the opinion of these authors, such judgments would yield just as meaningful, if not more meaningful results as those based on discrimination. Another way to get at the problem of spectral distance for the time-varying case would be to measure correspondences for the case of data-reduced time-varying synthesis of musical sounds. Three possibilities, where the fidelity of synthesis can be easily varied, are (1) frequency-modulation synthesis (Horner *et al.*, 1993a), (2) principal-component synthesis (Horner *et al.*, 1993b), and (3) critical-band wavetable synthesis (Beauchamp and Horner, 1995). In each of these cases, unlike the static spectral variation method, ratios between original and synthetic harmonic amplitudes vary with time.

Our investigation provides an excellent framework for timbre hybridization. In future work, it will be interesting to see which instrument spaces yield the most promising sounds. For example, it is easy to imagine that the Australian didgeridoo would produce a fascinating blend of timbres with almost any other instrument in small amounts. Mongolian throat singing could also produce some amazingly rich timbres as primary instrument, and a fascinating way of "muting" other instruments when used as the secondary instrument.

These results may also be relevant for orchestration theory, as well as for the identification of salient timbre dimensions, an important topic in music information retrieval.

TABLE IV. SI, NCD, SIR, and average 10%, 15%, and 20% error level discrimination scores for the eight primary instruments.

Instrument	Spectral incoherence	Normalized centroid deviation	Spectral irregularity	10% discrim.	15% discrim.	20% discrim.
Bs	0.075	0.400	0.093	0.826	0.866	0.925
Cl	0.085	0.700	0.174	0.775	0.812	0.857
Fl	0.118	0.600	0.129	0.752	0.845	0.88
Hn	0.057	0.200	0.073	0.927	0.917	0.94
Ob	0.069	0.800	0.137	0.756	0.795	0.873
Sx	0.101	0.600	0.195	0.662	0.769	0.801
Tp	0.184	1.600	0.039	0.681	0.710	0.787
Vn	0.193	1.400	0.131	0.730	0.822	0.873

Downloaded 12 Mar 2011 to 143.89.90.104. Redistribution subject to ASA license or copyright; see http://asadl.org/journals/doc/ASALIB-home/info/terms.jsp

TABLE V. SI, NCD, SIR, and average 10%, 15%, and 20% error level discrimination scores for the eight secondary instruments.

Instrument	Spectral incoherence	Normalized centroid deviation	Spectral irregularity	10% discrim.	15% discrim.	20% discrim.
Bs	0.075	0.400	0.093	0.668	0.800	0.827
Cl	0.085	0.700	0.174	0.860	0.872	0.912
Fl	0.118	0.600	0.129	0.780	0.801	0.851
Hn	0.057	0.200	0.073	0.745	0.756	0.832
Ob	0.069	0.800	0.137	0.719	0.815	0.845
Sx	0.101	0.600	0.195	0.802	0.831	0.901
Tp	0.184	1.600	0.039	0.814	0.871	0.944
Vn	0.193	1.400	0.131	0.695	0.753	0.812

ACKNOWLEDGMENTS

We wish to express our appreciation to Xavier Rodet for his suggestion to undertake this spectral interpolation detection study as a follow-up to our earlier work on spectral alteration detection. We would like to thank Cammy Zhuang, for running subjects in the listening experiment, and Simon Cheuk-Wai Wun, for writing the program for Intel personal computers used for the listening tests. This work was supported in part by the Hong Kong Research Grants Council's CERG Project Nos. 613806 and 613508.

Beauchamp, J. W., and Horner, A. (**1995**). "Wavetable interpolation synthesis based on time-variant spectral analysis of musical sounds," presented at the 98th AES Convention, Audio Engineering Society Preprint No. 3960, pp. 1–17.

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*, Proceedings of the Seventh International Conference on Music Perception and Cognition (University of New South Wales, Sydney, Australia), pp. 592–595.

Beauchamp, J. W. (**2007**). in *Analysis, Synthesis, and Perception of Musical Sounds*, edited by J. W. Beauchamp (Springer, New York), pp. 1–89.

Bernstein, L. R., Richards, V. M., and Green, D. M. (**1987**). in *Auditory Processing of Complex Sounds*, edited by W. A. Yost and C. S. Watson (Lawrence Erlbaum Associates, Hillsdale, NJ), pp. 6–15.

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

Grey, J. M. (**1975**). "An exploration of musical timbre," Ph.D. thesis, Stanford University, Stanford, CA; also available as Stanford Department of Music Report No. STAN-M-2, Stanford University, Stanford, CA.

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.

Gunawan, D., and Sen, D. (**2008**). "Spectral envelope sensitivity of musical instrument sounds," J. Acoust. Soc. Am. **123**, 500–506.

Horner, A., Beauchamp, J. W., and Haken, L. (**1993a**). "Machine tongues XVI: Genetic algorithms and their application to FM matching synthesis," Comput. Music J. **17**, 17–29.

Horner, A., Beauchamp, J. W., and Haken, L. (**1993b**). "Methods for multiple wavetable synthesis of musical instrument tones," J. Audio Eng. Soc. **41**, 336–356.

Horner, A., Beauchamp, J. W., and So, R. (**2004**). "Detection of random alterations to time-varying musical instrument spectra," J. Acoust. Soc. Am. **116**, 1800–1810.

Horner, A., Beauchamp, J. W., and So, R. (**2006**). "A search for best error metrics to predict discrimination of original and spectrally altered musical instrument sounds," J. Audio Eng. Soc. **54**, 140–156.

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

Johnson, R. A., and Wichern, D. W. (**1992**). *Applied Multivariate Statistical Analysis*, International ed. (Prentice-Hall, Englewood Cliffs, NJ).

Kendall, R. A., and Carterette, E. C. (**1993**). "Identification and blend of timbres as a basis for orchestration," Contemporary Music Review **9**, 51–67.

Kendall, R. A., and Carterette, E. C. (**1996**). *Difference Thresholds for Timbre Related to Spectral Centroid*, Proceedings of the Fourth International Conference on Music, Perception and Cognition (Faculty of Music, McGill University, Montreal, Canada), pp. 91–95.

Kidd, G., Jr., 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 (Acoustic analysis and timbre perception)," 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 (Timbre characteristics for complex sounds. II. Acoustical analysis and psychophysical measurements)," J. Phys. **4**, 625–628.

Krumhansl, C. L. (**1989**). in *Structure and Perception of Electroacoustic Sounds and Music*, edited by S. Nielzén 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**). in *Frequency Analysis and Periodicity Detection in Hearing*, edited by R. Plomp and G. F. Smoorenburg (Sijthoff, Leiden, The Netherlands), pp. 405–408.

Rimsky-Korsakov, N. (**1964**). *Principles of Orchestration* (Dover, New York), pp. 22–23, 33.

Sandell, G. J. (**1995**). ""Roles for spectral centroid and other factors in determining "blended" instrument pairings in orchestration," Music Percept. **13**, 209–246.

Stevens, J. (**2002**). *Applied Multivariate Statistics for the Social Sciences*, 4th ed. (Lawrence Erlbaum Association, Hillsdale, NJ), pp. 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. **43A**, 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**, 45–52.

Downloaded 12 Mar 2011 to 143.89.90.104. Redistribution subject to ASA license or copyright; see http://asadl.org/journals/doc/ASALIB-home/info/terms.jsp