A Search for Best Error Metrics to Predict Discrimination of Original and Spectrally Altered Musical Instrument Sounds*

ANDREW B. HORNER, AES Member
(horner@cs.ust.hk)

Department of Computer Science, Hong Kong University of Science and Technology, Kowloon, Hong Kong

JAMES W. BEAUCHAMP, AES Fellow
(jwbeauch@uiuc.edu)

School and Music and Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

AND

RICHARD H. Y. SO
(rhyso@ust.hk)

Department of Industrial Engineering and Engineering Management, Hong Kong University of Science and Technology, Kowloon, Hong Kong

The correspondence of various spectral difference error metrics to human discrimination data was investigated. Time-varying harmonic amplitude data were obtained from the spectral analysis of eight musical instrument sounds (bassoon, clarinet, flute, horn, oboe, saxophone, trumpet, and violin). Sounds were resynthesized with various levels of random spectral alteration, ranging from 1 to 50%. Listeners were asked to discriminate the randomly altered sounds from reference sounds resynthesized from the original data. Then several formulas designed to predict discrimination performance were evaluated by calculating the correspondence between the discrimination data and the associated spectral difference measurements. Averaged over the eight instruments, the best correspondence was achieved using a spectral error metric based on linear harmonic amplitude differences normalized by rms amplitude and raised to a power $a$. While an optimum correspondence of 91% was achieved for $a = 0.64$, good correspondence occurred over a wide range of $a$. For linear harmonic amplitudes without rms normalization, good correspondence occurred within a narrower range, with a maximum correspondence of 88%. Correspondence was approximately 80% for decibel-amplitude differences over an even narrower range. Other error metrics such as those based on critical-band grouping of components worked well but did not give any improvement over the method based on harmonic amplitudes, and in some cases yielded worse results. Spectral differences using a small number of representative frames emphasizing attack and decay transients yielded results slightly better than using all frames.

0 INTRODUCTION

How to best predict the perceptual difference between two musical instrument sounds has been a long-standing problem in computer music. Listening tests are ideal for measuring such differences, but they are not always possible or practical. Therefore a numerical error metric that accurately predicts average listener discrimination between similar sounds is highly desirable. The metrics evaluated in this paper are based on the time-varying amplitudes of the harmonics of sustained musical instru-
ment sounds. In all of the sounds tested, time-varying partial frequencies are replaced by fixed harmonic frequencies in order to focus listener attention on the perception of harmonic amplitude (or spectral envelope) differences.

Error formulas typically measure spectral differences using harmonic amplitudes on the basis of either harmonics or critical bands. The metric can normalize linear harmonic amplitudes by rms amplitude, or use decibel amplitudes. Usually either a time-averaged or a peak error is used, which can include all time frames or only a subset of the frames.

Spectral differences are especially important in applications such as spectral modeling and data reduction. For example, the optimization of parameters for frequency modulation resynthesis requires an error metric to measure how closely the synthetic result matches the original [1]. An error metric is also useful for measuring the synthesis error resulting from wavetable additive resynthesis [2] or the use of piecewise linear approximations to amplitudes-versus-time functions [3].

Plomp [4] considered the correspondence between an error metric and discrimination data in his early work on timbre differences. His metric was a decibel spectral error using one-third-octave bands. Investigating static spectra of musical instruments and vocals, Plomp found that spectral differences correlated quite well (80–85%) with listeners’ judgments of timbral dissimilarity, and concluded that differences in timbre can be predicted well from spectral differences.

The authors of this paper recently measured the discrimination of eight resynthesized sustained musical instruments from corresponding sounds whose spectra were altered randomly by various amounts [5]. The current paper extends this work by determining how well various error metrics match human discrimination. It examines how discrimination correspondence varies from instrument to instrument to see whether one error metric stands out as best, or whether there are several good metrics. Results for absolute spectral differences, squared (Euclidean) differences, and absolute differences raised to other powers are considered. The performance of decibel amplitudes versus linear amplitudes, and of harmonic amplitudes versus amplitudes averaged over critical bands are compared.

In other related work digital audio researchers have developed objective measures for perceived audio quality [6]–[7], especially for perceptually coded audio signals. The current work, which focuses on the spectral alteration of simple harmonic instrument tones, is a specialized refinement within this general framework. It is also an important special case in music synthesis. Random spectrum alteration is a form of spectral coloration, where a small change to a simple harmonic instrument tone might be imperceptible, whereas a larger change might be perceptible but not annoying. Even a large change would be only slightly annoying compared to the more audible linear distortions of a digital audio compression algorithm pushing the limit with respect to bit-rate reduction, and nonlinear distortions would be yet more annoying. It has received less attention than linear and nonlinear distortion, since it is a relatively benign artifact in perceptually coded audio. One of the goals of the current work is to determine which aspects of the harmonic instrument tone carry more information, and may therefore be more important than others when assessing alterations. This is useful for quality assessment of music, since an appropriate spectrotemporal weighting has not yet been found. We will consider weighting of low and high harmonics as well as how to average time frames.

In recent years error metrics have also been used for sound source identification and timbre classification using time- or frequency-domain-based feature vectors. In a series of three papers Fujinaga and his colleagues used a k-nearest-neighbor (k-NN) classifier, a genetic algorithm (GA), and several features extracted from 1338 musical instrument spectra for timbre recognition. In the first study [8], based on steady-state spectra, the GA was used to pair down 352 features to seven, including spectral centroid, standard deviation, skewness, and amplitudes of the first two harmonics. While the average recognition rate for the 39 instruments was 50.3%, for three instruments it was 80.5%. In the second study [9], when changes of features were incorporated, recognition accuracy was improved to 63.6% for 39 instruments and 98.3% for three instruments. In this case, while the spectrum-related features again included centroid, standard deviation, and skewness, they were augmented by measures of the rate of change and the amount of change of these features. In the third study [10], for real-time timbre recognition, spectral irregularity [11] and tristimulus [12] features were introduced. Weights for the different features were computed using a genetic algorithm, and the 39-instrument recognition rate was increased to 68%. In 1999 Brown [13] devised separate training and test sets based on feature vectors consisting of constant-Q cepstral coefficients, which were derived from oboe and saxophone spectra with one-third-octave resolution. Classification was done using a combination of k-means clustering and Gaussian probability density functions for the training set and a Bayes decision rule for identification. Saxophone and oboe sounds were identified correctly 90 and 96% of the time, respectively. In 2001 Brown et al. [14] expanded the 1999 study to include four instruments (flute, oboe, clarinet, and saxophone) and a much larger number of features. In this case cepstral coefficients gave an average of 77% correct whereas “spectral irregularity” (defined as bin-to-bin differences of the constant-Q spectrum) yielded 84% accuracy. In 2002 Agostini et al. [15] measured 18 features (including centroid, bandwidth, inharmonicity, and harmonic skewness) from 1007 tones taken from 27 instruments. Four classification methods, quadratic discriminant analysis (QDA), canonical discriminant analysis (CDA), k-nearest-neighbor (k-NN), and support vector machines (SVM), were compared. For the 27 instrument correct classification ranged from 60.3% (CDA) to 69.7% (SVM), whereas instrument family correct classification ranged from 72.9% (CDA) to 80.8% (QDA). Herrera-Boyer et al. [16] reviewed a large number of timbre classification schemes but arrived at no conclusion as to the best method.
These methods of classification are probably not applicable to the problem defined in this paper. Classification generally assumes that spectra representing different instruments are significantly different from one another, whereas we are assuming that the differences are rather small. Furthermore, automatic classification is quite a different problem from the problem of correlating a metric with psychoacoustic data. In the former case there is no need to take perception into account. Nevertheless some of the ideas of spectral classification might be useful in future studies on metrics for estimating spectral similarity. For the most part they beyond the scope of the current study.

We did try one spectral feature correspondence. The tristimulus [12] is based on three features: fundamental intensity, intensity of harmonics 2 to 4, and intensity of harmonics 5 and above, each normalized by the total intensity and varying with time. Using an error metric based on these features, we found an average correspondence of 0.36, substantially below that achieved by using the metrics considered in this study.

Hereafter the technique used for analyzing and synthesizing the stimuli is presented, followed by a comparison of error metrics for calculating spectral differences. The listening test procedure is outlined, and discrimination results are given along with their correspondences with the various error metrics, which are discussed in detail. Finally conclusions are made about which error metrics are best for applications such as music resynthesis.

1 STIMULI PREPARATION

Eight sustained (nonpercussive) musical instrument sounds were selected as prototype signals for stimulus preparation. (Except for the horn and bassoon, these reference sounds were also used by McAdams et al. [11], and all of them were used by Horner et al. [5].) In order that noise was not a major factor and that we could focus on the effects of spectral alteration, original sounds with minimal breath noise were selected. The sounds were first subjected to time-varying spectrum analysis using a computer-based phase vocoder method [17]. This phase vocoder is different from most in that it allows the tuning of a fixed analysis frequency \( f_a \) to coincide with the estimated fundamental frequency of the input signal. Thus the analysis bin frequencies are integer multiples of \( f_a \). The analysis method yields frequency deviations between harmonics of \( f_a \) and the corresponding frequencies of the input signal, which are assumed to be approximately harmonic relative to the fundamental.

1.1 Signal Representation

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

\[
s(t) = \sum_{k=1}^{K} A_k(t) \cos \left[ 2\pi \int_0^t \left[ k f_a + \Delta f_k(\tau) \right] d\tau + \theta_k(0) \right]
\]

where

- \( s(t) \) = sound signal
- \( t \) = time in seconds (interpreted as either continuous time or discrete samples)
- \( \tau \) = integrand dummy variable representing time
- \( k \) = harmonic number
- \( K \) = number of harmonics
- \( A_k(t) \) = amplitude of \( k \)th harmonic at time \( t \)
- \( f_a \) = analysis frequency
- \( \Delta f_k(t) \) = \( k \)th harmonic’s frequency deviation;
  - \( \Delta f_k(t) = k f_a + f_k(t) \) is the estimated frequency of the \( k \)th harmonic
- \( \theta_k(0) \) = initial phase of \( k \)th harmonic

The parameters used for resynthesis in this study are \( A_k(t) \). The frequency deviations \( \Delta f_k(t) \) were intentionally set to zero to restrict listener attention to the harmonic amplitude data. Although \( A_k(t) \) and \( f_k(t) \) were only stored as samples occurring every \( T_0 = 1/2f_a \), the synthetic signals were generated at a much higher data rate (sample frequency 22 050 or 44 100 Hz) by using linear interpolation between these values.

1.2 Prototype Sustain Sounds

Bassoon, clarinet, flute, horn, oboe, saxophone, trumpet, and violin sounds performed at approximately 311.1 Hz (E'0) were used to represent the wind and the bowed string instrument families. Five of the sounds were taken from the McGill University Master Samples recordings, two were taken from the Prosonus Sound Library (bassoon and oboe), and one (trumpet) was recorded at the UIUC School of Music. McAdams et al. [11] gives more details about the characteristics of these sounds, except for the horn and bassoon.

1.3 Analysis Method

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

1) Segmentation of the input signal into contiguous frames whose lengths are equal to twice the analysis period \( 2T_0 \) and overlap by \( T_0 = T_f/2 \), the time between frames. The input signal is interpolated using a bandlimited method to produce a power-of-2 number of samples per analysis period \( T_a = 1/f_a \).

2) Multiplication of each frame by a Hamming window function.

3) Fast Fourier transform (FFT) of each frame; components that are not positive integer multiples of \( f_a \) are discarded.

4) Computation of the magnitude (amplitude) and phase of each harmonic of \( f_a \).

5) Harmonic frequency deviations are computed from phase differences, but they are not used in this study

6) Harmonic amplitude and frequency deviation data are stored as \( f_a/\tau \) floats per frame in an “analysis file,” which becomes the basis for further data processing. Since the number of frames \( N \) in a sound with duration \( T_{\text{dur}} \) is \( T_{\text{dur}}/T_0 = 2T_{\text{dur}}/f_a \), the size of an analysis file is approximately \( 8T_{\text{dur}}/f_a \) bytes, or four times the size of the corre-
sponding monaural sound file with 16-bit (2-byte) samples.

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

The analysis system may also be viewed as a set of contiguous bandpass filters which have identical bandwidths \( f_a \) and are centered at the harmonics of the analysis frequency \( f_a \). The basic assumption is that the signal consists of harmonic sine waves, which line up with the filters such that each filter outputs one of the sine waves. The analysis gives the amplitude and frequency of each sine wave. When the sine waves are summed, the signal is almost perfectly reconstructed. In fact, the sine-wave sum can be viewed as a signal created by processing the input by a superposition of bandpass-filter characteristics. For Hamming window filters it can be shown that this sum varies a small amount over the range \( [f_a/2, f_a/2] \), with maxima of \( 0 \) dB occurring at the harmonic frequencies and minima of \(-1.4 \) dB at the halfway points. Fig. 1 shows a block diagram of the basic analysis/synthesis system.

1.4 Duration and Loudness Equalization

In order that sound duration would not be a factor in the study, most of the sounds were shortened to a 2-s duration by special interpolation of the analysis data [11]. Next amplitude multipliers were determined by a loudness program [18] in order that each sound would have an estimated loudness of 87.4 phons. An iterative procedure adjusted the amplitude multiplier starting from a value of 1.0 until the resulting loudnesses were within 0.1 phon of 87.4 (which corresponded to the trumpet sound played through headphones at 78 dB SPL).

1.5 Harmonic Frequency Flattening

In order that frequency variations and inharmonicity would not be factors in this study, these were eliminated from the sounds by setting each harmonic’s frequency equal to the exact product of its harmonic number \( k \) and the fixed analysis frequency \( f_a \), resulting in flat equally spaced frequency envelopes, that is,

\[
    f_k = kf_a. \tag{2}
\]

Frequency flattening was previously shown to have an effect on discrimination by Grey and Moorer [19] and Charbonneau [20]. However, McAdams et al. [11] found it to have a minimal effect compared to other alterations. The effect is strongly influenced by the amount of frequency deviation in the original sound. Frequency flattening was used in this study to avoid having frequency fluctuations influence listeners’ judgments, especially considering that the effect of these fluctuations may be amplified when harmonic amplitudes are altered. Since the original tones were all nearly strictly harmonic with no appreciable vibrato, the effect of flattening the frequency deviations was relatively minor.

The duration- and loudness-equalized frequency-flattened sounds then served as the reference sounds for this study, and their corresponding spectral data sets are henceforth referred to as the analysis data.

1.6 Random Spectrum Alteration

Random spectral alteration was performed on the analysis data, after which the sounds were generated by the additive resynthesis method. Random alteration was done by multiplying each harmonic amplitude by a time-invariant random scalar \( r_k \).

\[
    A_k(t) = r_k A(t). \tag{3}
\]

Eq. (3) describes a linear stationary process. The goal of this random spectral alteration is to perturb each harmonic amplitude, without changing the spectral centroid or loudness. By uniformly picking \( r_k \) in the range \( [1 - 2\varepsilon, 1 + 2\varepsilon] \), the “expected” error is approximately \( \varepsilon \), though the “actual” error, or the average harmonic-amplitude fractional error \( \varepsilon_k \) (see Section 2.3), deviates slightly from \( \varepsilon \). For this study, we generated 50 sounds for each instrument, where the expected error \( \varepsilon \) (henceforth referred to as the error level) ranged from 1 to 50% in increments of 1%. For example, for 50% error \( r_k \) was picked from the range \([0, 2]\).

In order to avoid introducing extreme variations in the harmonic amplitudes, which could make individuals stand out, they were grouped in critical bands within which they shared the same random scalar. We used a formula given by Zwicker and Terhardt [21], which gives the critical bandwidth for any band-center frequency \( f_c \).

\[
    \Delta f_{cb} = 25 + 75 \left[ 1 + 1.4 (0.001 f_c + 2)^{0.69} \right]. \tag{4}
\]

Fig. 2 shows an original spectral envelope and five randomly altered spectral envelopes with errors of 10, 20, 30, 40, and 50%. The envelopes become more irregular with increasing error, departing more and more from the shape of the original. The spectral errors introduced in these
randomly altered spectral envelopes are similar to those encountered in FM and wavetable parameter optimization of music instrument tones [1], [2], where errors from one band of harmonics can vary fairly independently from its neighbors.

The 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 [22]–[28]. Normalized centroid-versus-time functions for original and altered sounds can be defined as

\[
\text{NSC}(t) = \frac{\sum_{k=1}^{K} kA_k(t)}{\sum_{k=1}^{K} A_k(t)} \quad \text{and} \quad \text{NSC}'(t) = \frac{\sum_{k=1}^{K} kA'_k(t)}{\sum_{k=1}^{K} A'_k(t)}.
\] (5)

To preserve the spectral centroid after random spectrum alteration has been applied, the altered spectra are tilted in order to match the original centroid. Tilting is accomplished by scaling the altered harmonic amplitudes by raising the frequency to a to be determined power \( p \),

\[
A'_k(t) = C^{p}A_k(t) .
\] (6)

Starting with \( p = 0 \), \( p \) is iterated using Newton’s method until the normalized centroid of the altered harmonic amplitudes is within ±0.1 of the target value based on the original amplitudes. While it would be possible to match the original centroid-versus-time function on a frame-by-frame basis, we found that it was adequate and probably more desirable (in terms of minimum artifacts) to use a fixed value of \( p \) for each sound based on the sound's average harmonic spectrum.

Finally a fixed amplitude multiplier \( C \) is determined such that each altered sound has the same loudness as the resynthesized original sound (87.4 phons), as measured by the LOUDEAS program [18], so that the final harmonic amplitudes are given by

\[
A''_k(t) = CA'_k(t) = Ck^pA_k(t) = Ck^pr_kA_k(t).
\] (7)

The random spectrum alteration algorithm therefore consists of the following steps:

1) Pick initial values for \( r_k \) such that \( 1 - 2e < r_k < 1 + 2e \).
2) Apply random alteration: \( A'_k(t) = r_kA_k(t) \).
3) Centroid equalization:
   a. Calculate the average spectra of original and altered sounds.
   b. Calculate the spectral centroids of the average spectra from step 3a.
   c. Iteratively tilt the altered time-varying spectrum using Newton’s method until the average centroids match.
4) Iterative loudness equalization using the LOUDEAS program of Moore et al. [18].
5) End.

1.7 Resynthesis Method

For synthesis at the signal sample rate, linear interpolation of the \( A''_k(t_n) \) amplitude envelope data is used between frames, so that

\[
A''_k(t) = \frac{t_{n+1} - t}{T_0} A''_k(t_n) + \frac{t - t_n}{T_0} A''_k(t_{n+1})
\] (8)

where \( n \) is the frame number, \( t_n = nT_0 \) is the beginning time of frame \( n \) with \( T_0 \) the time between frames, and \( A''_k(t_n) \) is the amplitude of the \( k \)th harmonic at time \( t_n \) after the random multiplier, centroid, and loudness equalizations are applied. While linear interpolation introduces a smoothing operation to the harmonic amplitude functions, in most cases its effect is hardly noticeable. Moreover,
only sounds that have undergone the same basic processing operations, including the original reference sounds, are compared in the listening experiment.

Additive resynthesis is accomplished by additive (or Fourier) synthesis of the altered harmonic sine waves,

$$\delta(t) = \sum_{k=1}^{K} A_{tk}^o(t) \cos(2\pi k f t + \theta_k(0)).$$

(9)

Note that the frequency deviations of Eq. 1(1) have been removed.

2 ERROR METRICS

There are several different approaches for computing spectral differences, including harmonic amplitude differences, relative-amplitude and decibel-amplitude spectral errors, amplitude envelope errors, frequency-weighted amplitude differences, and use of a subset of the frames versus all frames in the calculations. We will outline these various error metrics in this section.

2.1 Harmonic-Amplitude Error Metrics

Probably the simplest error metric is the average simple distance measure based on linear harmonic amplitudes treated as vectors, which we call linear-amplitude spectral error,

$$e_{dase} = \frac{1}{N} \sum_{n=1}^{N} a \sqrt{\sum_{k=1}^{K} |A_k(t_n) - A_k^o(t_n)|^a}$$

(10)

where

- $n =$ analysis frame number
- $t_n =$ time in seconds of analysis frame $n$
- $N =$ number of analysis frames
- $k =$ harmonic number
- $K =$ number of harmonics
- $A_k(t_n) =$ amplitude of resynthesized original signal’s $k$th harmonic at time $t_n$
- $A_k^o(t_n) =$ amplitude of spectrally altered signal’s $k$th harmonic at time $t_n$
- $a =$ arbitrary exponent applied to each amplitude difference. While $a$ is most commonly set to 1 or 2, it may have a different optimum value.

Normally for the metric calculations we take $N = 20$, where 10 points equally spaced in time are taken from the “attack” portion of the sound (defined as the time from the onset up to the point of maximum rms amplitude of the original signal) and the rest are equally spaced in time over the remaining portion of the sound. Two advantages of using a subset are that error computation is cheaper and a subset can provide more emphasis on perceptually important time regions such as the sound’s attack and decay. Several previous studies (see [1], [2], [29]) have utilized spectral error metrics based on 10 equally spaced spectral frames taken from the attack (defined as the time period before the peak rms amplitude point) and 10 equally spaced frames taken from the rest of the sound. It has never been shown empirically whether using all available spectral frames or only a few carefully chosen representative spectral frames provides a better correspondence to perceptual difference. Results obtained related to the aforementioned problem are discussed in Section 4.3.1.

Note that the highest amplitude harmonics of the highest amplitude time frames make the strongest contributions to the linear spectrum error. This emphasizes the sustained part of most sounds, since they are usually the loudest. Statistical procedures, such as principal-components analysis (PCA), have been used to reduce additive synthesis data to wavetable synthesis data by minimizing the error in Eq. (10) [2], [30]–[32]. Also a least-squares approach has been used to determine optimum amplitude-versus-time envelopes for frequency modulation and wavetable synthesis by minimizing Eq. (10) [2]. Both applications minimize the Euclidean spectral distance where $a = 2$. Also, double frequency modulation matching work has used the linear spectral error with $a = 1$ [33], [34].

Alternatively one might argue that decibel differences are a better measure of how humans hear. The decibel-amplitude spectral error can be formulated as

$$e_{rasewdn} = \frac{1}{N} \sum_{n=1}^{N} a \sqrt{\sum_{k=1}^{K} \frac{|L_k(t_n) - L_k^o(t_n)|^a}{\sum_{k=1}^{K} L_k(t_n)}}.$$  

(12)

We refer to this error measure, whose values lie between 0 and 1, as the relative-amplitude spectral error with simple normalization. Authors of papers on wavetable and frequency modulation resynthesis models have used Eq. (12) with $a = 2$ [1], [2], [35] and sometimes with $a = 1$. It is also possible to normalize by both the original and the altered harmonic amplitudes. We call the resulting error measure relative-amplitude spectral error with dual normalization,

$$e_{rasewd} = \frac{1}{N} \sum_{n=1}^{N} a \sqrt{\sum_{k=1}^{K} \frac{|L_k(t_n) - L_k^o(t_n)|^a}{\sum_{k=1}^{K} L_k^o(t_n)}},$$  

(13)

McAdams et al. [16] used a variation on Eq. (13) with an alternative normalization method, which we will refer to as relative-amplitude spectral error with maximum normalization,

$$e_{rasewdm} = \frac{1}{N} \sum_{n=1}^{N} a \sqrt{\sum_{k=1}^{K} \frac{|L_k(t_n) - L_k^o(t_n)|^a}{\max[L_k(t_n), L_k^o(t_n)]}},$$  

(14)

Rather than averaging the harmonic differences, one could consider only the worst difference at each time frame to emphasize the occurrence of a single large spectral differ-
ence. The following is the maximum relative-amplitude spectral error,

\[ e_{\text{max}} = \frac{1}{N} \sum_{n=1}^{N} \sqrt{\frac{\max_{k=1}^{K} |A_k(t_n) - A_k^*(t_n)|^2}{\sum_{k=1}^{K} A_k^2(t_n)}}. \]  

(15)

It is also possible to take the root after the summation in the relative-amplitude spectral error [Eq. (12)]. This emphasizes larger amplitude differences. The following is the rms relative-amplitude spectral error,

\[ e_{\text{rms}} = \frac{1}{N} \sqrt{\frac{\sum_{n=1}^{N} \sum_{k=1}^{K} |A_k(t_n) - A_k^*(t_n)|^2}{\sum_{k=1}^{K} A_k^2(t_n)}}. \]  

(16)

### 2.2 Critical-Band-Amplitude Error Metrics

Eqs. (10)–(16) define error metrics based on the amplitudes of individual harmonics. They can be rewritten to depend on the combined amplitudes of critical bands in order to better represent the human auditory system. To accomplish this, an algorithm was devised to sort the harmonics into critical bands whose bandwidths are defined by Eq. (4). The amplitude of a critical band is defined as the rms value of the sinusoid amplitudes within the band,

\[ A_k(t_n) = \sqrt{\sum_{k=1}^{K} A_k^2(t_n)} \]  

(17)

Then the linear-amplitude critical-band error is defined as

\[ e_{\text{labe}} = \frac{1}{N} \sum_{n=1}^{N} \sum_{k=1}^{C} |A_k(t_n) - A_k^*(t_n)|^a \]  

(18)

where \( c \) is the critical band number, and \( C \) is the number of critical bands. Similarly the decibel-amplitude critical-band error is defined as

\[ e_{\text{dabe}} = \frac{1}{N} \sum_{n=1}^{N} \sum_{k=1}^{C} |L_k(t_n) - L_k^*(t_n)|^a \]  

(19)

where \( L_k(t_n) = 20 \log[A_k(t_n)] \) and \( L_k^*(t_n) = 20 \log[A_k^*(t_n)] \). Eq. (19) is similar to that used by Plomp [4] in his study of the correspondence of error metrics and discrimination data, except that he used one-third-octave bands instead of critical bands.

Similar to Eqs. (12)–(14) the relative-amplitude critical-band error is defined as

\[ e_{\text{rabe}} = \frac{1}{N} \sum_{n=1}^{N} \sqrt{\frac{\sum_{c=1}^{C} |A_k(t_n) - A_k^*(t_n)|^2}{\sum_{c=1}^{C} A_k^2(t_n)}} \]  

(20)

where \( A_k^a \) represents simple, dual, or maximum normalization. The maximum relative-amplitude critical-band error is defined as

\[ e_{\text{max}} = \frac{1}{N} \sum_{n=1}^{N} \sqrt{\frac{\max_{c=1}^{C} |A_k(t_n) - A_k^*(t_n)|^2}{\sum_{c=1}^{C} A_k^2(t_n)}} \]  

(21)

and the rms relative-amplitude critical-band error as

\[ e_{\text{rabe}} = \frac{1}{N} \sqrt{\frac{\sum_{n=1}^{N} \sum_{c=1}^{C} |A_k(t_n) - A_k^*(t_n)|^2}{\sum_{c=1}^{C} A_k^2(t_n)}}. \]  

(22)

### 2.3 Harmonic-Amplitude Envelope Error Metric

The error metrics defined in Sections 2.1 and 2.2 weight each spectral frame equally. Alternatively we can weight each harmonic equally with a relative-harmonic-amplitude envelope error given by

\[ e_{\text{rhaee}} = \frac{1}{K} \sum_{k=1}^{K} \sqrt{\frac{\sum_{n=1}^{N} |A_k(t_n) - A_k^*(t_n)|^2}{\sum_{n=1}^{N} A_k^2(t_n)}}. \]  

(23)

Taking into account the random multipliers and the loudness and centroid equalizations [see Eq. (7)], we can substitute \( A_k^w = c k^p r_k A_k \) so that Eq. (23) can be simplified to

\[ e_{\text{rhaee}} = \frac{1}{K} \sum_{k=1}^{K} [1 - C k^p r_k] \]  

(24)

which is independent of \( a \). (It would depend on \( a \) only if different \( r_k \) were picked for each time frame.) Note that the envelope error of Eq. (23) corresponds to the average harmonic-amplitude fractional error \( e' \) mentioned in Section 1.6.

### 2.4 Frequency-Weighted Relative-Harmonic-Amplitude Error Metric

If a frequency-dependent term \( k^b \) is introduced as an amplitude weight in the relative-amplitude spectral error metric [Eq. (12) with \( a = 1 \)], low or high frequencies can be emphasized. The frequency-weighted relative-harmonic-amplitude error then becomes

\[ e_{\text{fwrae}} = \frac{1}{N} \sum_{n=1}^{N} \left[ \sum_{k=1}^{K} k^b |A_k(t_n) - A_k^*(t_n)|^2 \right] \]  

(25)

Eq. (25) emphasizes high frequencies when \( b > 0 \), low frequencies when \( b < 0 \), and reduces to relative-amplitude spectral error [Eq. (12)] when \( b = 0 \).

### 3 EXPERIMENTAL METHOD

#### 3.1 Listening Test Organization

The 20 subjects who participated in the listening tests were undergraduate students at the Hong Kong University of Science and Technology, ranging in age from 18 to 23 years, who reported no hearing problems. They included 10 “musicians” (six males, four females) and 10 nonmusicians (four males, six females). Musicians were defined as those having at least five years of practice on an instrument, and nonmusicians were defined as never having played a musical instrument. The subjects were paid for their participation.

The authors acknowledge that the chances of failing to detect an existing significant effect (probability of type II error) could be further reduced by using more subjects [36]. However, 20 subjects were considered adequate because a previous study on a similar topic was able to detect a significant effect at an alpha level of 0.05 with 20 subjects [5]. In other words, it has been demonstrated that the effect of the spectral error on discrimination was strong...
enough to be detected within a 5% type I error in the presence of data variance of 20 randomly selected subjects.

3.2 Listening 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, AA–BA, AB–AA, or BA–AA, where A represents the reference sound and B one of 50 randomly altered sounds. This paradigm has the advantage of not being as susceptible to variations in subjects’ criteria across experimental trials compared to the simpler A–B method, that is, there is no need to subtract for guessing.

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. On each trial the user was prompted with “which pair is different, 1 or 2?” and gave a response using the computer keyboard. The computer would not accept a response until at least the first pair was played. The Intel PC command-line program which controlled the experiment was custom written in a computer science laboratory at HKUST.

For each instrument a block of 200 trials was presented to each subject (four trial structures x 50 error levels). The duration of each block was about 40 minutes. Eight blocks were presented corresponding to the eight instruments. Therefore the total duration of the experiment was about 6 hours. Listeners took 5–10-min breaks between blocks and finished the test in two separate 3-hour sessions. Only a few listeners’ scores degraded significantly over the duration of the test. We eliminated these as not reliable enough to take their results into account. From an initial group of 28 subjects we used the scores of the 20 most reliable listeners.

Subjects were seated in a “quiet room” with a 40-dB SPL background noise level (mostly due to computers and air conditioning). The stimuli, which were stored on a hard disk in 16-bit integer format, were converted to analog by SoundBlaster Audigy soundcards and then presented through Sony MDR-7506 headphones at 87.4 phons. The Audigy DAC uses 24 bit 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, depending on the instrument.

At the beginning of the experiment each subject read a set of instructions and was given the opportunity to ask any necessary questions. Five test trials (chosen at random from the altered sounds) were presented prior to the data trials for each instrument. The order of presentation of the 200 trials was random within each block, and the order of presentation of the instruments was randomized for each subject.

4 RESULTS

4.1 Discrimination Scores

Fig. 3 shows discrimination-versus-error-level scores averaged over the 20 subjects for each of the 50 random alteration error levels for each instrument. A discrimination score of 100% represents perfect distinguishability, whereas 50% represents random guessing. Most points are close to the superimposed fourth-order polynomial regression S curves, which provide the best least-squares fits to the data. Mean discrimination scores are close to 100% in the presence of 50% spectral errors. This suggests that the 20 subjects were not guessing, and they were able to detect audible differences introduced by the spectral errors. If the subjects were guessing, the mean discrimination scores would be close to 50%, which was the case when no spectral errors were added.

We hypothesized that a cluster plot of discrimination results versus percentage error would form an S shape consisting of two points of inflection. The use of an S-shaped fourth-order polynomial, which affords the possibility of two inflection points, to fit the data agrees well with this hypothesis. In fact we found that when the percentage error was close to zero (say 0–10%), listeners appeared to be making random choices, with discrimination results close to 50%. As the percentage error increased from 10% to a certain level (in our case, about 15%) listeners found it easier to discriminate between the two stimuli, and discrimination improved as the percentage error increased. When the percentage error increased beyond about 30%, the rate of increase reduced toward an asymptotic level of 100%.

Most points are close to the regression curves, especially in the clarinet and oboe. The most scattered graph is that of the horn, but even it follows the regression. Since the same random harmonic-amplitude multipliers were used for all eight instruments, it is somewhat puzzling that a few outlier points occur for some but not all instruments. The most obvious example is that at the 27% error level the bassoon, flute, saxophone, trumpet, and violin data have distinct outliers. This phenomenon also occurs at the 36% error level in the bassoon, flute, horn, and violin graphs. However, the multipliers at these error levels do not always result in outliers; for example, in the clarinet graph both the 27% and the 36% error-level data points are near the regression curve. This phenomenon may have something to do with the instrument-specific resonance structures of the original spectra.

Fig. 4 combines the instrument discrimination data of Fig. 3 into a single graph. The semitransparent rectangles indicate the 25th and 75th percentiles of the data collected when errors are at the ranges of 0–0.1; 0.1–0.2; 0.2–0.3; 0.3–0.4; and 0.4–0.5, respectively. For error levels up to 10% most scores are in the range of 40–60%. These scores are close to an indistinguishability level of 50%, corresponding to random guessing. The range is wider and more variable for intermediate error levels between 15 and 25%, where the scores cover nearly the full range from 50 to 100%. Intermediate error levels correspond to “somewhat distinguishable” cases. For error levels more than 30%, most scores are above 90% and are “very distinguishable.” An analysis of variances (ANOVA) with Student–Newman–Kuel (SNK) analyses was conducted to investigate the main effects of percentage error on the discrimination data. The analyses identified three statistically separate groups of mean discrimination data (p < 0.05): 1) mean data up to 58%; 2) mean data from 64 to 87%; and 3) mean data from 91 to 98%. The first group corresponds exactly to data collected from experimental
conditions with percentage errors of 0 to 12%, whereas the other two groups approximately corresponded to data from 13 to 31% and from 32 to 50%. In other words, despite the intersubject variability (Fig. 4), statistical analyses indicate that discrimination scores associated with percentage errors between 32 and 50% have over 95% chance to be greater than those scores associated with percentage errors between 13 and 31%, which in turn have over 95% chance to be greater than those scores associated with percentage errors of 12% or below.

4.2 Correspondence of Error Metrics and Discrimination Scores

Each of the error metrics of Section 2 was calculated to determine its correspondence with the discrimination data. For example, Fig. 5 shows discrimination scores versus relative-amplitude spectral error [Eq. (12)] with the error weighting exponent $\alpha = 1$. The fourth-order polynomial regression fit to the data increases from less than 50% discrimination and gradually converges to near 100%. Again, the semitransparent rectangles indicate the 25th and 75th percentiles of the data collected when spectral errors are the ranges of 0–0.1; 0.1–0.2; 0.2–0.3; 0.3–0.4; and 0.4–0.5, respectively.

Regression analysis provides a measure of how much variance each error metric accounts for in the discrimination data. The coefficient of determination or squared multiple correlation coefficient $R^2$ [37] measures how well the data values fit a regression curve, and thus the correspondence between the discrimination scores and a particular error metric. $R^2$ is calculated on the 400 data points corresponding to the 50 error levels for each of the eight instruments and is defined as

$$R^2 = \frac{\sum_{i=1}^{400} (d_i - \bar{d})^2}{\sum_{i=1}^{400} d_i^2}$$

(26)

Fig. 3. Mean subject discrimination scores for randomly altered sounds versus error level for eight instruments. — fourth-order polynomial regression trends; ♦—average discrimination score over 20 listeners at a particular error level.
where \( d_i \) is the \( i \)th discrimination score, \( d_i' \) is the regression curve approximation of the \( i \)th discrimination score, and \( \overline{d} \) is the mean discrimination score. For example, if \( R^2 = 0 \), the error metric explains none (that is, 0\%) of the variation in the discrimination data. On the other hand, \( R^2 = 1 \) means that all the data points lie on the regression curve, and all (that is, 100\%) of the variation in the discrimination scores is explained by the error metric. With \( R^2 = 0.9 \) the error metric accounts for 90\% of the variance in the discrimination data.

### 4.2.1 Linear-Amplitude Error Metric Results

Fig. 6 shows \( R^2 \) plotted against exponent \( a \) for the linear-amplitude spectral error metric of Eq. (10). This metric accounts for about 83–88\% of the variance when \( 0.5 \leq a \leq 2 \). The absolute spectral difference (with \( a = 1 \)) is better than the Euclidean spectral distance (with \( a = 2 \)). The best \( R^2 \) correspondence is 88\% at \( a = 0.7 \).

Decibel-amplitude spectral error \( R^2 \) results are shown in Fig. 7. The maximum correspondence converges to about 80\% as \( a \) increases, worse than that of the linear spectral error, which converges down to about the same level. Low \( a \) values give poor correspondence. For \( a = 1 \) only about 66\% of the variance is accounted for.

Fig. 8 shows \( R^2 \) plotted versus \( a \) for the relative-amplitude spectral error metric of Eq. (12). The maximum correspondence is 91\% (at \( a = 0.64 \)). Moreover, the curve is quite flat, with correspondences of 87\% at \( a = 2 \) and 85\% at \( a = 0.3 \) and \( a = 3 \). Thus it does not matter much whether \( a = 0.5 \), \( a = 1 \), or \( a = 2 \): the results are good in all cases. This robustness demonstrates an advantage of the relative-amplitude spectral error over the linear and dB spectral error metrics. The correspondence curve for the relative-amplitude spectral error with dual normalization [Eq. (13)] is almost identical to Fig. 8 and has a maximum correspondence of 91\% at \( a = 0.68 \). Likewise the correspondence for the relative-amplitude spectral error with maximum amplitude normalization [Eq. (14)] is nearly identical to Fig. 8, with a maximum correspondence about 0.5\% better than those of Eqs. (12) and (13) (91\% at \( a = 0.68 \)).
It is a bit surprising that the result of Eq. (14) is slightly superior to that of Eq. (12); nevertheless, Eq. (12) seems mathematically more elegant.

Fig. 9 shows $R^2$ versus $a$ for the maximum relative-amplitude spectral error metric of Eq. (15). Although the metric is reasonably good for $a > 1$, low $a$ values yield poor correspondence. The maximum correspondence is 83% at $a = 1.90$. The correspondence curve for the rms relative-amplitude spectral error metric [Eq. (16)] is nearly identical to that of Fig. 8 [Eq. (12)], with a maximum correspondence of 91% at $a = 0.64$. In other words, whether the root is taken before or after the summation has practically no effect.

### 4.2.2 Critical-Band Error Metrics

It seemed logical to us that the critical-band error metrics discussed in Section 2.2 would correlate better with the discrimination data than the harmonic amplitude difference metrics. However, we are not able to demonstrate such an advantage in terms of the maximum correspondence achieved. Fig. 10 shows $R^2$ versus $a$ for the linear-amplitude critical-band error metric of Eq. (18). The maximum correspondence is 87% at $a = 0.48$. It is similar to the linear-amplitude harmonic error in Fig. 6, but the peak is slightly less. It yields good correspondence for $a$ as low as 0.16.

On the other hand, as shown in Fig. 11, the correspondence for the decibel-amplitude critical-band error metric of Eq. (19) is significantly better than that of the decibel-amplitude harmonic error metric of Eq. (11) shown in Fig. 7 ($p < 0.001$, Wilcoxon signed ranked test) and much more robust with respect to the variation of $a$. The maximum correspondence is 87% at $a = 1.3$. Still, it is not as good as the relative-amplitude metrics, such as Eq. (12), shown in Fig. 8, in terms of peak correspondence and robustness.

---

Fig. 6. $R^2$ versus $a$ for linear-amplitude spectral error [Eq. (10)]. Maximum correspondence 88% at $a = 0.7$.

Fig. 7. $R^2$ versus $a$ for decibel-amplitude spectral error [Eq. (11)]. Maximum correspondence converges to about 80% as $a$ increases.

Fig. 8. $R^2$ versus $a$ for relative-amplitude spectral error [Eq. (12)]. Maximum correspondence 91% at $a = 0.64$. Of the metrics tested, this was the best result.

Fig. 9. $R^2$ versus $a$ for maximum relative-amplitude spectral error [Eq. (15)]. Maximum correspondence 83% at $a = 1.90$.

Fig. 10. $R^2$ versus $a$ for linear-amplitude critical-band error [Eq. (18)]. Maximum correspondence 87% at $a = 0.48$.

Fig. 11. $R^2$ versus $a$ for decibel-amplitude critical-band error [Eq. (19)]. Maximum correspondence 87% at $a = 1.3$. 
The correspondence for the relative-amplitude critical-band error metric of Eq. (20), plotted in Fig. 12, is excellent for all $a$ values in the range given and shows significant improvement over the harmonic error for $a < 0.5$. While it is somewhat surprising that the critical-band measure is slightly worse than relative-amplitude harmonic error metric (see Fig. 8) for most $a$ values, they are not significantly different ($p = 0.75$, Wilcoxon signed ranked test). Relative-amplitude critical-band error metrics with dual and maximum normalization give results almost identical to Fig. 12, as does the rms relative-amplitude critical-band error metric. Fig. 13 shows that the maximum relative-amplitude critical-band error metric of Eq. (21) rises faster than the comparable harmonic error (see Fig. 9) for $a < 1$, but the two errors are nearly identical for $a \geq 1$.

4.2.3 Relative-Amplitude Envelope Error Metric

The correspondence of the relative-amplitude envelope error metric [defined by Eqs. (23) and (24)] is invariant with $a$, so a flat discrimination versus $a$ curve results in 79%. Here the relative-amplitude envelope error is equivalent to the actual error $e'$ mentioned in Section 1.6. The correspondence for the error level $e$ is 81%. Therefore the error level and the actual error are in close agreement, although other metrics have better values of correspondence.

4.2.4 Frequency-Weighted Relative-Harmonic-Amplitude Error

It might be expected that frequency weighting of harmonic amplitudes, which introduces an additional degree of freedom and takes advantage of possible increased listener sensitivity to certain frequency regions, would offer an improvement over the frequency-insensitive relative-amplitude spectral error metric. However, we are only able to demonstrate a slight advantage.

Fig. 14 shows $R^2$ versus $b$ for the frequency-dependent relative spectral error of Eq. (25), where $a$ is set to 1.0. Both positive and negative $b$ values are included. The maximum correspondence is 91% for $b = 0.54$, which corresponds to weighting higher harmonics slightly more than lower ones. Note, however, that the correspondence is close to 90% at $b = 0$, corresponding to the $a = 1.0$ result for the relative-amplitude spectral error. Also optimizing both $a$ and $b$ yields little improvement. Therefore, overall, frequency weighting improves the correspondence by less than 1%.

4.3 Sensitivity Analyses

4.3.1 Effect of Number of Frames on Correspondence

For all of the metrics discussed in previous sections we assumed $N = 20$, with 10 equally spaced frames taken from the attack and 10 equally spaced frames taken from the rest of the sound. As an alternative we calculated the correspondence of the relative-amplitude spectral error given by Eq. (12) with $N = N_{\text{max}}$, where $N_{\text{max}}$ is the total number of frames (approximately 1250 for these sounds) available in the analysis. For this case correspondence versus $a$ was nearly identical to that of Fig. 8 (where the number of frames was $N = 20$), although using all frames actually gave a slightly inferior result. So using 20 representative spectra, with 10 in the attack and 10 over the rest of the sound, appears to yield marginally better correspondence than using all frames (and is much faster to compute).

We also found excellent correspondence using fewer than 20 frames. Fig. 15 shows $R^2$ plotted versus the number of frames used in the relative-amplitude spectral error calculation (with $a = 1$). Again, half the points were taken from the attack and half from the remainder of the tone. (For an odd number of points, the middle point coincided...
with the maximum rms amplitude. We note that even a single frame gives a correspondence of 89% and speculate that this strong result is due to random spectrum alteration being time-invariant. Thus listeners tended to focus on overall spectral envelope differences, which could be captured in a single frame measurement. Of course, this result would be different for time-varying alterations, which we did not explore. For example, if the subset of frames used by the error metric had very little alteration compared to those not used by the error metric, then the discrimination correspondence would be relatively poor.

4.3.2 Effects of Instruments

Fig. 16 superimposes \( R^2 \)-versus-\( a \) curves for the relative-amplitude spectral error metric on the eight instruments (bassoon, clarinet, flute, horn, oboe, saxophone, trumpet, and violin). Inspection of Fig. 16 indicates that all instruments follow a similar trend. Wilcoxon signed ranked tests conducted to compare the \( R^2 \) values, for all values of \( a \), from different instruments indicated the following results: 1) \( R^2 \) was significantly higher for flute and oboe \((p < 0.001)\); 2) this was followed by \( R^2 \) for saxophone, clarinet, and trumpet; 3) \( R^2 \) values for trumpet, bassoon, and violin were not significantly different from each other \((p > 0.2)\); 4) \( R^2 \) values associated with horn were the lowest \((p < 0.001)\). While the statistical tests can indicate a consistent and reliable ranking of the \( R^2 \) values, the impact of the ranking on the absolute values of \( R^2 \) is not large except for violin and horn, the two extreme cases. These analyses were repeated for the rms relative-amplitude spectral error metric and similar results were obtained.

4.3.3 Effects of Order of Regression Fit

Based on our hypothesis that discrimination-versus-error data would follow a curve with two inflection points, we decided to use a fourth-order regression fit, as discussed in Section 4.1. However, we also tried replacing the fourth-order regression function with third- and fifth-order polynomials. Inspection of Fig. 17 indicates that there is only a very slight observable difference among the three regressions for the relative-amplitude spectral error metric [Eq. (12)]. Further examination of the \( R^2 \) data indicates that the maximum differences occur at the peak \( R^2 \) values, but the differences are less than 1%. The foregoing analyses were repeated for the rms relative-amplitude spectral error metric [Eq. (16)], and similar results were obtained.

5 DISCUSSION

Several of the metrics produced excellent peak correspondences with the discrimination data. However, the range over which the parameter \( a \) gave near-peak correspondence varied considerably. Table 1 gives the maximum \( R^2 \) and the range of the parameter \( a \) over which \( R^2 \) is within 5% of the maximum \( R^2 \) for each error metric. The relative-amplitude spectral error [Eq. (12)] explains over 90% of the variation in the discrimination data. This metric is very robust, with good results for absolute differences, Euclidean differences, and differences raised to other powers. In the past (see [1]–[3]) we found the relative-amplitude spectral error to be an excellent tool for optimizing parameters for applications such as frequency modulation and wavetable synthesis.

All forms of normalization [Eqs. (12)–(14)] performed well, as did the rms relative-amplitude spectral error [Eq. (16)]. The best results for linear-amplitude spectral error [Eq. (10)] were about as good as the relative-amplitude spectral error, but the metric was considerably less robust in terms of sensitivity to the power of the spectral difference. This perhaps accounts for why some researchers have noticed artifacts resulting from principal-component-analysis-based methods for data reduction of additive syn-
thesis data [2], [30]–[32], since principal-component analysis only optimizes the linear spectral error with $a = 2$.

We were surprised that the critical-band-based error metrics [Eqs. (18)–(22)] yielded no better and sometimes slightly inferior results compared to those based on harmonics for most $a$ values. On the other hand the critical-band errors were among the least sensitive to changes in the power $a$ of all the metrics studied, with excellent performance for all powers between 0.3 and 3. The decibel-amplitude spectral error [Eq. (11)], the maximum relative-amplitude spectral error [Eq. (15)], and the relative-amplitude envelope error [Eq. (23)] gave the worst results of the metrics tested, with only about 80% correspondence. However, decibel-amplitude differences with critical bands [Eq. (19)] gave results comparable to the relative-amplitude spectral error.

Another surprise was that using a frequency dependence in the metric offered only a slight improvement in the correspondence. We would have expected that emphasizing the lower harmonics might help, since they are usually more prominent, but this was not the case. Amplitude-based weighting seems to work fine without any frequency weighting.

Computing spectral differences using the relative-amplitude spectral error metric with 20 representative frames resulted in a slightly better correspondence than using all frames. This allows more emphasis on the perceptually important attack and decay and is also about two orders of magnitude more efficient. However, another surprise was that less than 20 frames also yielded excellent correspondence (see Fig. 15), and even one frame yielded a correspondence of 89%, which is only 2% less than the best correspondence of any of the metrics with $N = 20$.

While this study concentrated on measuring the correspondence with the fourth-order regression curve for the discrimination data, the relative-amplitude spectral error metric appears to be robust with third, fourth-, and fifth-order regression fits across all eight instrument sounds tested, independent of the number of frames used in the calculation. Therefore the order of regression fit, at least within the 3-to-5 range, does not seem to be important.

6 CONCLUSIONS

All error metrics tested have been found to have at least a reasonable correspondence (70% or more) with the discrimination of random spectrum modifications, and several had excellent correspondences just over 90%. Relative-amplitude spectral error and rms relative-amplitude spectral error had excellent peak correspondence and good robustness with respect to the power $a$. Relative-amplitude critical-band errors achieved similar peak correspondence, and in general were somewhat more robust with respect to variations in $a$. Frequency weighting did not provide significant improvement. However, correspondence using a small number of well-selected frames in the relative-amplitude spectral difference metric yielded a slight improvement over taking all frames, verifying the effectiveness of this useful shortcut. Even one frame gave excellent results.

Absolute spectral differences ($a = 1$) outperformed Euclidean spectral differences ($a = 2$) on nearly every metric, with $a = 0.6$ about optimal. The strong correspondence of metrics such as relative-amplitude spectral error allows music syntheists to use them with confidence, since they account for over 90% of the variance in the discrimination data.

The relative-amplitude spectral error metric provides a means for estimating how much additive synthesis data reduction of musical sounds is possible without incurring significant perceptual alteration. For example, if a 20% data reduction results in a relative spectral error of 8%, Fig. 5 shows that this change corresponds to an almost indistinguishable discrimination level of 50–70%. Moreover, the strong correspondences of the relative-amplitude and rms relative-amplitude spectral error metrics to the empirically collected discrimination data have been shown to be robust against the use of third-, fourth-, and fifth-order regression fits. The strong correspondences were

<table>
<thead>
<tr>
<th>Error Metric</th>
<th>Maximum $R^2$</th>
<th>$a$ Value of Lower Bound</th>
<th>$a$ Value of Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear-amplitude spectral error</td>
<td>0.877</td>
<td>0.44</td>
<td>0.70</td>
</tr>
<tr>
<td>Decibel-amplitude spectral error</td>
<td>0.785</td>
<td>1.66</td>
<td>3.00+</td>
</tr>
<tr>
<td>Relative-amplitude spectral error with simple normalization</td>
<td>0.908</td>
<td>0.30</td>
<td>0.64</td>
</tr>
<tr>
<td>Relative-amplitude spectral error with dual normalization</td>
<td>0.906</td>
<td>0.30</td>
<td>0.68</td>
</tr>
<tr>
<td>Relative-amplitude spectral error with maximum normalization</td>
<td>0.914</td>
<td>0.32</td>
<td>0.70</td>
</tr>
<tr>
<td>Maximum relative-amplitude spectral error</td>
<td>0.834</td>
<td>0.94</td>
<td>1.90</td>
</tr>
<tr>
<td>Rms relative-amplitude spectral error</td>
<td>0.908</td>
<td>0.30</td>
<td>0.64</td>
</tr>
<tr>
<td>Linear-amplitude critical-band error</td>
<td>0.873</td>
<td>0.20</td>
<td>0.48</td>
</tr>
<tr>
<td>Decibel-amplitude critical-band error</td>
<td>0.874</td>
<td>0.46</td>
<td>1.30</td>
</tr>
<tr>
<td>Relative-amplitude critical-band error with simple normalization</td>
<td>0.903</td>
<td>0.06</td>
<td>0.54</td>
</tr>
<tr>
<td>Relative-amplitude critical-band error with dual normalization</td>
<td>0.903</td>
<td>0.06</td>
<td>0.60</td>
</tr>
<tr>
<td>Relative-amplitude critical-band error with maximum normalization</td>
<td>0.908</td>
<td>0.10</td>
<td>0.60</td>
</tr>
<tr>
<td>Maximum relative-amplitude critical-band error</td>
<td>0.836</td>
<td>0.86</td>
<td>1.84</td>
</tr>
<tr>
<td>Rms relative-amplitude critical-band error</td>
<td>0.903</td>
<td>0.06</td>
<td>0.52</td>
</tr>
<tr>
<td>Relative-amplitude envelope error</td>
<td>0.791</td>
<td>0.02–</td>
<td>(all)</td>
</tr>
<tr>
<td>Frequency-dependent relative-amplitude spectral error</td>
<td>0.907</td>
<td>0.70</td>
<td>0.54</td>
</tr>
<tr>
<td>Relative-amplitude spectral error using all analysis frames</td>
<td>0.907</td>
<td>0.30</td>
<td>0.68</td>
</tr>
</tbody>
</table>
also maintained when regression analyses were conducted for each of the eight instrument sounds.

7 FUTURE WORK

Some issues for further investigation are the following.

1) With current metrics, determine the minimum number of frequency and time values and their spacings required for good correspondence.

2) Explore new ways to combine or select frequency components that improve discrimination correspondence (since critical-band amplitudes did not result in significant improvement).

3) In hopes of discovering an error metric with a discrimination correspondence greater than 91%, test combinations of error metrics such as spectrotemporal features like spectral irregularity and spectral flux [11] (used in instrumental timbre classification [15], [16]).

4) Restrict the listening data to “golden ear” subjects, and compare the results with those of ordinary subjects.

5) Use random spectrum alteration to produce new musical sounds that have the same pitch, duration, loudness, spectral centroid, and error level (with respect to a reference sound) but are clustered in perceptually distinct groups, which can be further analyzed to determine salient timbral parameters.

6) Compare metrics for their ability to predict the discrimination of time-varying spectral alterations such as occur with FM or wavetable resynthesis.

8 ACKNOWLEDGMENT

This work was supported in part by the Hong Kong Research Grant Council’s Projects HKUST6167/03E and HKUST6135/05E. We would like to thank Simon Cheuk-Wai Wun for his excellent listening test, and Wu Yuk Ying and Francis Hing-Cheung Wong for managing the listening test.

9 REFERENCES


1995 he was a visiting researcher at the Institut de Recherche et Coordination in Acoustique/Musique (IRCAM) in Paris, France.

During the 1970s, 1980s, and 1990s Dr. Beauchamp produced over 30 concerts of electronic and computer music at the UIUC. He has taught courses in musical acoustics and computer music for the UIUC School of Music as well as courses in electronic music synthesis and audio for the Dept. of Electrical and Computer Engineering at UIUC. In 1997 he retired from the UIUC, but continues his affiliation with it as professor emeritus. His current research interests are in sound analysis algorithms, sound synthesis models, perceptual effects of musical sound parameter modifications, automatic pitch detection, timbre classification, and instrument separation.

Dr. Beauchamp is a coeditor and a coauthor of *Music by Computers* (1969) and has published numerous articles in professional journals on computer music synthesis and analysis, musical acoustics, and audio. He is the primary author/developer of the Music 4C sound synthesis software package and the SNDAN sound analysis/synthesis package for both Unix workstations. He is a fellow of the Audio Engineering Society and the Acoustical Society of America (ASA) and is currently chair of the Technical Committee for Musical Acoustics of the ASA. He was the president (1981–1983) and a member of the Board of Directors (1980–1983, 1986–1993) of the International Computer Music Association (ICMA) and cohosted the International Computer Music Conferences (ICMCs) at the UIUC School of Music in 1975 and 1987.

Richard H. Y. So is associate professor of human factors and head of the computational ergonomics research team, Department of Industrial Engineering and Logistics Management, Hong Kong University of Science and Technology. His research interests include biological-inspired computational models of spatial vision, and spatial hearing.

Dr. So is a registered member of the Ergonomics Society (UK), a founding council member of the Hong Kong Ergonomics Society, and a senior member of the American Institute of Aeronautics and Astronautics.