Evaluation of Mel-Band and MFCC-Based Error Metrics for Correspondence to Discrimination of Spectrally Altered Musical Instrument Sounds*

Andrew B. Horner, *AES Member*
(horner@cse.ust.hk)

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

James W. Beauchamp, *AES Life Fellow*
(jwbeauch@illinois.edu)

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

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

Several mel-band-based metrics and a single MFCC-based error metric were evaluated for best correspondence with human discrimination of single tones resynthesized from similar musical instrument time-varying spectra. Results show high levels of correspondence that are very close and often nearly identical to those found previously for harmonic and critical-band error metrics. The number of spectrum-related terms in the metrics required to achieve 85% R^2 correspondence is about five for harmonics, ten for mel bands, and ten for MFCCs, leading to the conjecture that subjects discriminate more on the basis of the first few harmonics than on the broad spectral envelope.

0 INTRODUCTION

Recent work [1] investigated the R^2 correspondence of various harmonic-based spectral error metrics to human discrimination data for a set of sustained musical tones which differ from a reference set by gradated amounts of random spectral error. A maximum correspondence of 91% was found for a relative-amplitude spectral error metric, with several other metrics achieving correspondences nearly as high. Metrics using critical bands, based on Zwicker and Terhardt's formula [2], were the most robust, although their peak correspondences were not better than metrics using equally weighted harmonic amplitudes, and in some cases they yielded slightly worse results. A question arose as to whether a metric based on mel bands or mel-frequency cepstral coefficients (MFCCs) could yield an improvement.

The mel-band and mel-frequency-cepstral-coefficient spectrum analysis methods are designed to provide a useful data reduction of sonic spectra. MFCCs in particular have been used extensively and successfully for speech recognition applications and, more recently, have been employed in a wide variety of music applications. In 1997 De Poli and Prandoni [3] used

*Manuscript received 2010 April 29; revised 2011 February 10.

them for categorizing musical instrument spectral envelope data. In 2000 Logan [4] compared the use of MFCCs for modeling speech and music signals. In 2001 Aucouturier and Sandler [5] investigated their use for segmenting complex musical textures. In 2003 D'haes and Rodet [6] also used cepstral coefficients with mel-frequency warping in their work on feature detection in the identification of music instruments. In 2004 Weng et al. [7] used MFCCs for musical instrument identification. Terasawa et al. [8] used MFCCs for measuring perceptual distance in timbre space in 2005. Recently (2009) Brent [9] employed them for percussive timbre identification.

Since MFCCs have been so effective for speech recognition and so promising in music information retrieval problems, we decided it would be interesting to see how well they function as musical error metrics for the case of human discrimination of spectrally altered individual musical instrument sounds. Of course, MFCCs are most often used in music information retrieval applications that involve much more complex signals than single tones, but it is also interesting to see how well they work as error metrics for the single-tone case.

In the following, previous work on error metrics [1] is reviewed in detail. First we give background details on stimulus preparation, listening tests, and discrimination data interpretation. Then several mel-band metrics and a single MFCC-based error metric are presented. Results are given and discussed for these metrics as compared to corresponding harmonic and critical-band metrics presented in the previous study. Finally conclusions are drawn about the relative effectiveness and robustness of mel-band-based and MFCC-based error metrics.

1 PREVIOUS WORK ON ERROR METRICS

How to best measure the timbral difference between two musical sounds is a longstanding problem in music perception. Listening tests are ideal for measuring such differences, but they are not always possible or practical. Therefore a numerical error metric that correlates well with average listener discrimination between individual sounds is highly desirable. The metrics evaluated in this paper are based on the time-varying amplitudes of the harmonics, or groups of harmonics, in sustained musical instrument sounds. In all of the sounds tested, time-varying partial frequencies are replaced by fixed harmonic frequencies in order to focus listener attention on timbre perception based on the time-varying amplitude spectra. However, we note that the frequency variations of the sounds used in this study were barely audible to begin with.

Error formulas typically measure spectral differences using harmonic amplitudes on the basis of either harmonics, critical bands, or some other spectral grouping or feature. 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 difference measurements are important for applications such as spectral modeling and data reduction [10]–[12].

Plomp [13] considered the correspondence between an error metric and discrimination data in his early work on speech vowel and musical timbre differences. His metric treated the decibel outputs of one-third-octave bands as vectors and measured the Euclidean difference between the vectors. Investigating static spectra of musical instruments and vowels, Plomp found that this metric correlated quite well (80–85%) with listener judgments of timbral dissimilarity and concluded that differences in timbre can be predicted well from such spectral differences.

In a previous study [14], published in 2004, the authors of this paper measured the discrimination of eight resynthesized sustained musical instruments from corresponding sounds whose spectra were altered randomly by various amounts. Harmonic amplitudes were perturbed randomly while preserving spectral centroid. The original and altered sounds were also duration and loudness equalized, and frequency flattened to restrict listener attention to the harmonic amplitude data.

A follow-up paper [1], published in 2006, extended this work to determine how well various error metrics matched human discrimination. Fig.1 shows an overview of the error metric evaluation procedure. First spectrally altered tones are generated based on the original musical instrument tones. Next a listening test measures the ability of human listeners to discriminate the altered tones from the originals. Finally R^2 correspondences between the discrimination scores and spectral distances given by particular error metrics provide a measure of how well each error metric accounts for variations in the discrimination data [15].

Various harmonic and critical-band error metrics were compared in the 2006 study. Results for sums of squared (Euclidean) differences and absolute differences raised to other powers were considered. We found a best correspondence of 91% using an amplitude-normalized (relative) spectral error metric based on linear harmonic amplitude differences normalized by rms amplitude and raised to a power a, with good correspondence over a wide range of a. For linear harmonic amplitudes without amplitude normalization, good correspondence occurred within a narrower range of a, with a maximum correspondence of 88%. Correspondence was approximately 80% for decibel-amplitude differences over an even narrower range. Error metrics based on critical-band grouping of components worked well and improved the robustness of the metrics by widening the range of good correspondences with respect to a. However, they did not give any peak improvement over the method based on harmonic amplitudes, and in some cases they yielded slightly worse results.

2 STIMULI, LISTENING TESTS, AND DATA INTERPRETATION

The musical sound stimuli, listening test results, and data interpretation methods are identical to those of our 2006 metric study [1] and are briefly reviewed here. More details can be found in the 2006 study.

2.1 Stimulus Preparation

The reference stimuli consist of quasiperiodic signals taken from sounds performed by the following eight sustained musical instruments performed at approximately $f_0 = 311.1$ Hz (E^b_4): bassoon, clarinet, flute, horn, oboe, saxophone, trumpet, and violin. Note that the stimuli are limited to the same eight sustained tones used in the 2006 study. Time-variant harmonic analysis was performed by an F_0-synchronous short-time Fourier transform program [16], [17]. These signals were normalized as follows. 1) Durations were shortened to 2 seconds without altering attacks or decays. 2) Loudnesses were normalized to 87.4 phons using the LOUDEAS program of Moore et al. [18]. 3) Partial frequencies were set to fixed harmonic values so that $f_k = kf_0$ Hz, where the harmonic number $k = 1, \ldots, K$, with the number of harmonics K ranging between 30 and 70, depending on the instrument. Note that harmonic amplitudes time-varied as in the original recorded tones except for duration compression. The resynthesized reference signals conformed to the following sinusoidal model:

$$s(t) = \sum_{k=1}^{K} A_k(t)\cos(2\pi k f_0 t + \theta_k) \quad (1)$$

where $A_k(t)$ is the amplitude of the kth harmonic and θ_k is its starting phase.

Test spectra were produced by randomly varying the harmonic amplitudes by time-invariant multipliers so that $A'_k(t) = r_k A_k(t)$. The multipliers were confined to certain limits, that is, $r_k = 1 \pm 2e$. For each value of error level e, ranging from 0.01 to 0.50, r_k spectra were selected and modified so as to match the spectral centroids of the corresponding reference signals, and amplitudes were scaled in order that loudnesses were matched to the standard 87.4 phons.

2.2 Listening Tests

Twenty subjects aged 18 to 23 participated in the listening tests. A two-alternative forced-choice discrimination paradigm was used, where the listener's task was to identify which of two tone pairs presented in succession was the "different pair." Four different trial structures were used: AA–AB, AA–BA, AB–AA, and BA–AA, where A represents the reference sound, and B was one of the randomly altered sounds. Since there were 50 error levels and four trial structures, each subject processed 200 trials for each instrument.

2.3 Data Interpretation

Discrimination scores were averaged over the 20 subjects and four trial structures for each instrument/error-level combination, resulting in 50 scores for each instrument, or 400 scores altogether. For the error-metric application, the data for the eight instruments were combined, and for each error metric a fourth-order regression polynomial function was calculated. The R^2 correspondence measures the degree to which data conform to the regression polynomial. We determined that for a relative-amplitude spectral error metric, which in our previous study gave the best (91%) peak correspondence, R^2 varied very weakly with the regression order. This metric is given by

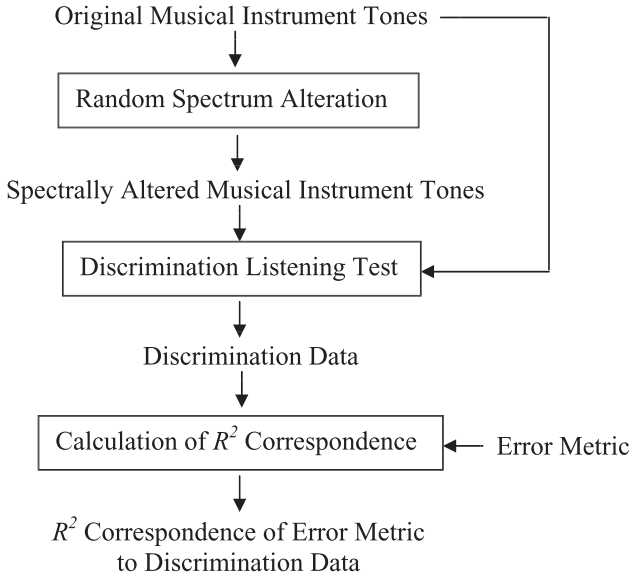


Fig. 1. Overview of error metric evaluation procedure.

$$\varepsilon_{\text{rase}} = \frac{1}{N}\sum_{n=1}^{N}\sqrt[a]{\frac{\sum_{k=1}^{K}|A_k(t_n) - A'_k(t_n)|^a}{\sum_{k=1}^{K}A_k^a(t_n)}} \quad (2)$$

where t_n is the time in seconds of analysis frame n and N is the number of frames.

Use of this metric to predict discrimination requires the actual regression polynomial which, in turn, depends on the data. However, since the regression curve is increasing monotonically, minimizing the error metric also minimizes discrimination. Thus the regression function is not needed for applications that require minimization rather than prediction.

3 MEL-BAND ERROR METRICS

Previous work on error metrics was based on amplitudes of individual harmonics or the combined amplitudes of critical bands. For the current study they have been rewritten to depend on combined amplitudes of mel bands, or cepstral coefficients of the mel bands.

3.1 Metrics Based on Mel-Band Amplitudes

Mel bands are based on the frequency-to-mel relationship, which was originally measured by Stevens and Volkmann [19]. This relationship, originally published as a data plot, has been approximated by various functions [20]. One of the most popular of these is given by O'Shaughnessy [21],

$$\text{mel}(f) = 2595\log_{10}(1 + f/700) \quad (3)$$

where f is the frequency in Hz, and mel(f) is the mel frequency in mels.

For our application the frequency range $0 \le f \le 30 \times 311.1 = 9333.0$ Hz is translated into the mel frequency range $0 \le \text{mel}(f) \le 3000.7$, and this range is divided into 27 contiguous overlapping triangular-shaped bands with center mel frequencies separated by $\Delta f_m = 3000.7/27 = 111.1$ mel, each having a width of 222.2 mel. Thus the mth band extends from $(m-1)\Delta f_m$ to $(m+1)\Delta f_m$ mel for $1 \le m \le 27$, and the center mel frequency of each band is given by $f_m = m\Delta f_m$. Harmonics within the mth band are those whose frequencies are such that $f_{m-1} < \text{mel}(f_k) < f_{m+1}$, where $f_k = k311.1$ Hz.

The mth triangular band characteristic is given by

$$W_m[\text{mel}(f)] = \begin{cases} \dfrac{1 - |f_m - \text{mel}(f)|}{\Delta f_m}, & |f_m - \text{mel}(f)| < \Delta f_m \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

Overlaid triangular bandpass characteristics and their intersections with the harmonics of 311.1 Hz translated into mels are shown in Fig. 2. While the harmonic amplitudes depicted here all have unit amplitude, the amplitudes for musical sound stimuli naturally vary with the harmonic number.

The effective linear amplitude of the mth mel band, defined as the square root of the sum of the squared amplitudes of the harmonics whose frequencies lie within the band, where each harmonic is weighted by the band characteristic,[1] is given by

$$\alpha_m(t_n) = \sqrt{\sum_{k=k_{m_1}}^{k_{m_2}} W_m[\text{mel}(f_k)]A_k^2(t_n)} \quad (5)$$

where

m	mel-band number
k	harmonic number
k_{m_1}	lowest harmonic in mth mel band
k_{m_2}	highest harmonic in mth mel band
$W_m()$	mth triangular band characteristic defined above

[1] In some implementations the weight is also squared. We have determined that this has negligible effect on correspondence results.

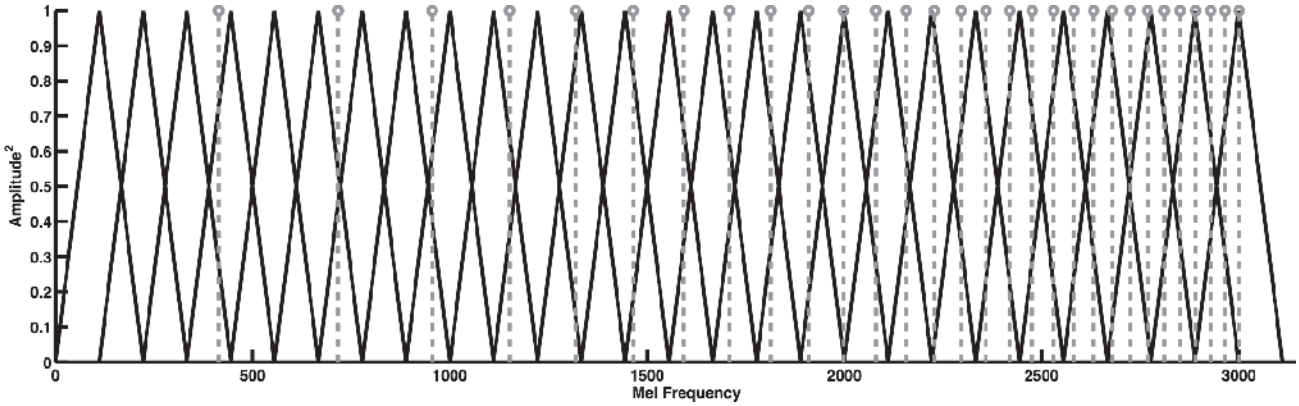


Fig. 2. Mel-band filter bank with overlaid (dashed vertical lines) harmonics of 311.1 Hz translated to mel frequency scale.

f_k constant frequency of kth harmonic, $= k311.1$ Hz

$A_k(t_n)$ amplitude of resynthesized original signal's kth harmonic at time t_n.

For the corresponding spectrally modified harmonic amplitudes $A'_k(t_n)$ the mth mel-band linear amplitude becomes $\alpha'_m(t_n)$, using the same formula as Eq. (5) with primes appropriately inserted. For this study the number of harmonics used in metric calculations was limited to 30 for all instruments, which for our fundamental frequency (311.1 Hz) closely matches a maximum mel value of 3000, an approximate standard [20].

A simple mel-band error metric is an average distance measure based on effective mel-band amplitudes treated as vectors, which we call linear-amplitude mel-band error,

$$\varepsilon_{\text{lambe}} = \frac{1}{N}\sum_{n=1}^{N}\sum_{m=1}^{M}|\alpha_m(t_n) - \alpha'_m(t_n)|^a \tag{6}$$

where

m mel-band number

M number of mel bands used in computation (normally 27)

N number of analysis frames, $= 20$

$\alpha_m(t_n)$ amplitude of resynthesized original signal's mth mel band at time t_n

$\alpha'_m(t_n)$ amplitude of spectrally altered signal's mth mel band 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.)

For metric calculations we use $N = 20$, where 10 points equally spaced in time are taken from the "attack" portion of the sound and the rest are equally spaced in time over the remainder of the sound. Two advantages of using a subset are that 1) error computation is cheaper, and 2) a subset can provide more emphasis on perceptually important time regions such as the sound's attack and decay. In the previous error metric paper [1] the authors showed that for this stimulus set using a few carefully chosen representative spectral frames actually provides a better correspondence to perceptual differences than using all frames. Note that the highest amplitudes of the highest amplitude time frames make the strongest contributions to the linear error. This emphasizes the sustained part of most sounds, which is usually the loudest.

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

$$\varepsilon_{\text{dambe}} = \frac{1}{N}\sum_{n=1}^{N}\sum_{m=1}^{M}|L_m(t_n) - L'_m(t_n)|^a \tag{7}$$

where $L_m(t_n) = 20\ \log_{10}[\alpha_m(t_n)]$ and $L'_m(t_n) = 20\ \log_{10}[\alpha'_m(t_n)]$. Eq. (7) is similar to that used by Plomp [13] in his study of the correspondence of error metrics and discrimination data, except that he used time-invariant spectra and one-third-octave bands instead of mel bands. Both Eqs. (6) and (7) emphasize spectral frames with higher amplitudes, and thus emphasize the perceptually important attack and decay.

Alternatively linear amplitude differences can be normalized using a relative-amplitude mel-band error,

$$\varepsilon_{\text{rambewsn}} = \frac{1}{N}\sum_{n=1}^{N}\sqrt[a]{\frac{\sum_{m=1}^{M}|\alpha_m(t_n) - \alpha'_m(t_n)|^a}{\sum_{m=1}^{M}[\alpha_m(t_n)]^a}}. \tag{8}$$

We refer to this error measure, whose values lie between 0 and 1, as the relative-amplitude mel-band error with simple normalization.

It is also possible to normalize by both the original and the altered harmonic amplitudes. We call the resulting error measure relative-amplitude mel-band error with dual normalization,

$$\varepsilon_{\text{rambewdn}} = \frac{1}{N}\sum_{n=1}^{N}\sqrt[a]{\frac{\sum_{m=1}^{M}|\alpha_m(t_n) - \alpha'_m(t_n)|^a}{\sum_{m=1}^{M}|\alpha_m(t_n)\alpha'_m(t_n)|^{a/2}}}. \tag{9}$$

An alternative normalization method was used by McAdams et al. [22]. In its mel-band version we refer to it as relative-amplitude mel-band error with maximum normalization,

$$\varepsilon_{\text{rambewmn}} = \frac{1}{N}\sum_{n=1}^{N}\sqrt[a]{\frac{\sum_{m=1}^{M}|\alpha_m(t_n) - \alpha'_m(t_n)|^a}{\sum_{m=1}^{M}\left\{\max[\alpha_m(t_n), \alpha'_m(t_n)]\right\}^a}}. \tag{10}$$

Still another possibility is to consider only the largest harmonic difference at each time frame, resulting in a maximum relative-amplitude mel-band error,

$$\varepsilon_{\text{mambe}} = \frac{1}{N}\sum_{n=1}^{N}\sqrt[a]{\frac{\max_{1\le m\le M}[|\alpha_m(t_n) - \alpha'_m(t_n)|^a]}{\sum_{m=1}^{M}[\alpha_m(t_n)]^a}}. \tag{11}$$

Finally the root can be taken after the summation in the relative-amplitude mel-band error [Eq. (8)], thus emphasizing larger amplitude differences. The following is the rms relative-amplitude mel-band error,

$$\varepsilon_{\text{rrambe}} = \sqrt[a]{\frac{1}{N}\sum_{n=1}^{N}\frac{\sum_{m=1}^{M}|\alpha_m(t_n) - \alpha'_m(t_n)|^a}{\sum_{m=1}^{M}[\alpha_m(t_n)]^a}}. \tag{12}$$

3.2 Metric Based on Mel-Frequency Cepstral Coefficients

Mel-frequency cepstral coefficients (MFCCs) are computed by applying the discrete cosine transform (DCT) to the log of the mel-band amplitudes, as given by Rabiner and Juang [23],

$$\mathrm{MFCC}_l(t_n) = \sum_{m=0}^{M-1} \log[\alpha_m(t_n)] \cos\left[(m+0.5)\frac{\pi l}{M}\right], \quad l = 0, \ldots, L-1 \tag{13}$$

where $\mathrm{MFCC}_l(t_n)$ is the lth MFCC corresponding to the spectrum $\{A_k(t_n)\}$ at frame n, and $L \leq M$ is the number of MFCCs.

Given $\mathrm{MFCC}_l(t_n)$ and $\mathrm{MFCC}'_l(t_n)$ as the lth mel-frequency cepstral coefficients of the resynthesized original and the spectrally altered signal at time t_n, the MFCC error is then defined as an average distance measure based on the two MFCCs treated as vectors,

$$\varepsilon_{\mathrm{mfcce}} = \frac{1}{N}\sum_{n=1}^{N}\sum_{l=1}^{L-1} |\mathrm{MFCC}_l(t_n) - \mathrm{MFCC}'_l(t_n)|^a \tag{14}$$

where a is an arbitrary exponent applied to each amplitude difference. While a is most commonly set to 1 or 2 (see [21]), it may have a different optimum value.

4 RESULTS

4.1 Correspondence of Error Metrics and Discrimination Scores

Each of the error metrics given in Section 3 was calculated to determine its correspondence with the discrimination data. 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 [15], was used to measure how well the data values fit a regression curve and thus measure the correspondence between discrimination scores and a particular error metric. 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 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.

We computed the correspondence using

$$R^2 = \frac{\sum_{i=1}^{I}(d'_i - \bar{d}')^2}{\sum_{i=1}^{I}(d_i - \bar{d})^2} \tag{15}$$

where i corresponds to combinations of error level e and instrument, I is the number of average discriminations (400 in our case), d_i is the ith discrimination score, and d'_i is the ith discrimination score predicted by a regression function, which is an Nth-order polynomial least-squares best fit to the discrimination-versus-metric-error data (for example, see Fig. 4). Note that from Eq. (15) we can see that if $d_i \cong d'_i$ for all i, then $R^2 \cong 1.0$.

For the interpretation of the correspondence results there are two caveats to keep in mind. 1) What is claimed is that if the metric's error value (that is, spectral distance estimate) increases, discrimination should also increase, within the accuracy given by the correspondence. Note that for the error metric to be valid, it is only necessary that the metric and the regression functions be increasing monotonically. 2) These results are only valid for the stimuli tested, and the correspondences and rankings of performance for various metrics that we report here should not be assumed to extend to other sets of stimuli.

Fig. 3 shows the discrimination data plotted versus the error level e with the corresponding overlaid fourth-order regression curve. R^2 in this case is 0.81, or 81%. Fig. 4 shows the discrimination data plotted versus the relative-amplitude spectral error metric [see Eq. (2)] for $a = 1$. Note the improved adherence to the fourth-order regression curve in this case, where R^2 is 0.91, or 91%.

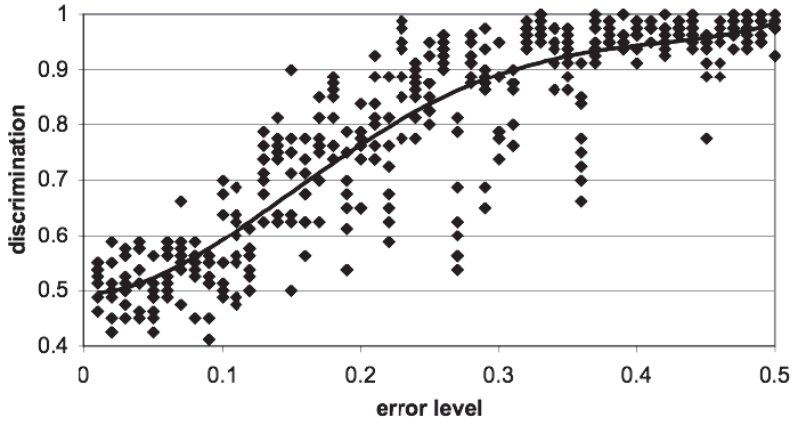


Fig. 3. Average discrimination scores versus error level e with overlaid fourth-order regression curve. Each point is average of 20 subjects times four trial structures. R^2 correspondence 81%.

4.2 Mel-Band Error Metric Correspondence

The harmonic and critical band errors of Figs. 5–8 correspond to Figs. 6–13 in our previous study [1], except that the number of harmonics has been set to 30 for all eight instruments. The resulting differences are relatively minor, except when $a < 0.5$.

Fig. 5 shows R^2 plotted against exponent a for the linear-amplitude mel-band error metric of Eq. (6). This metric accounts for about 80–87% of the variance when $a \geq 0.15$. The absolute spectral difference (with $a = 1$) is slightly better than the Euclidean spectral distance (with $a = 2$). The best R^2 correspondence is 87% at $a = 0.58$. For $a > 0.6$ the mel-band curve lies very close to the linear-amplitude critical-band and harmonic curves derived in our previous study [1], with the mel-band version being slightly less sensitive to a than the harmonic version.

R^2 versus a for the decibel-amplitude mel-band error of Eq. (7) is shown in Fig. 6. The maximum correspondence peaks at 89% at $a = 0.80$, which is close to the best metric performance of 91% in our previous study [1]. Correspondences for two metrics from the previous study are shown for comparison. The decibel-amplitude critical-band error results are similar to the mel-band results for $0.5 < a < 2.0$, but they deviate outside this range. The decibel-amplitude harmonic error correspondence is considerably worse for $a < 2.0$.

Fig. 7 shows R^2 plotted versus a for the relative-amplitude mel-band error metric of Eq. (8). The maximum correspondence is 90% (at $a = 0.52$), and the curve is quite flat with correspondences above 85% throughout the displayed range. Thus it does not matter much what the value of a is; the results are good in all cases. Also, the relative-amplitude mel-band curve is almost identical to the relative-amplitude critical-band and harmonic curves when $a \geq 0.4$. In general the combination of high peak correspondence with robustness is a major advantage of the relative-amplitude error over the linear and decibel error metrics.

Correspondence curves (not shown) for relative-amplitude mel-band error with dual normalization [Eq. (9)], for relative-amplitude mel-band error with maximum amplitude normalization [Eq. (10)], and for rms relative-amplitude mel-band error metric [Eq. (12)] are almost identical to the correspondence curve of Fig. 7 [Eq. (8)],

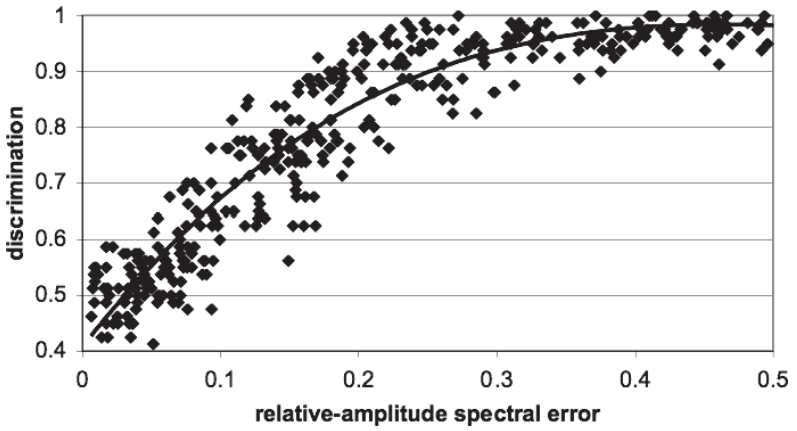


Fig. 4. Average discrimination scores versus relative-amplitude spectral error [Eq. (2)] with overlaid fourth-order regression curve. Each point is average of 20 subjects times four trial structures. R^2 correspondence 91%.

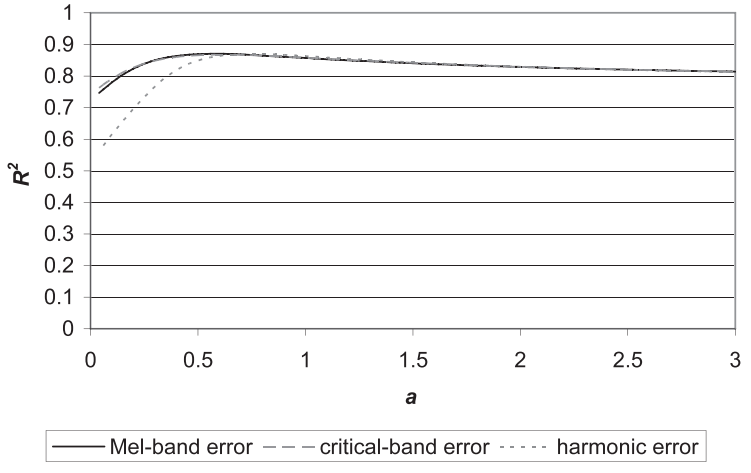


Fig. 5. R^2 versus a for linear-amplitude mel-band error [Eq. (6)]. Maximum correspondence is 87% at $a = 0.58$. Results for linear-amplitude critical-band and harmonic errors are also shown.

having maximum correspondences of 90% for a between 0.50 and 0.58. The latter result indicates that whether the root is taken before [Eq. (8)] or after [Eq. (12)] the summation has no practical effect on the correspondence of the metric to the discrimination data.

On the other hand, R^2 for the maximum relative-amplitude mel-band error metric of Eq. (11), as shown in Fig. 8, is noticeably inferior to the correspondences shown in Figs. 5–7, especially for $a < 1.0$. The maximum correspondence is 83% at $a = 1.88$.

4.3 MFCC Error Metric Correspondence

Fig. 9 shows R^2 versus a for the MFCC error metric of Eq. (14). The curve is very nearly as good as the relative mel-band error for $a < 1.5$, with a peak correspondence of 89% at $a = 1.32$, but degrades for higher values of a.

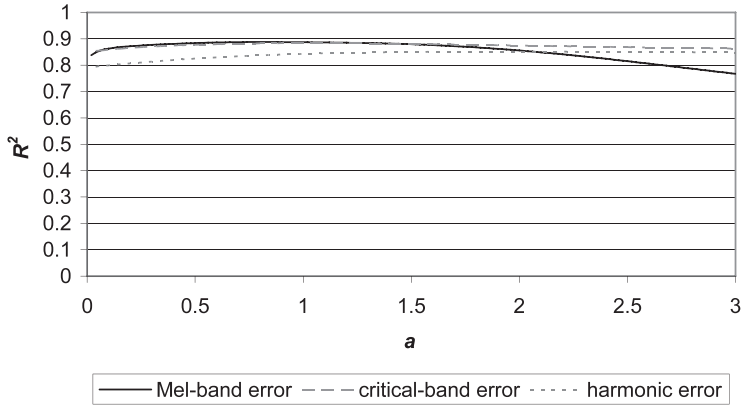


Fig. 6. R^2 versus a for decibel-amplitude mel-band error [Eq. (7)]. Maximum correspondence is 89% at $a = 0.80$. Results for decibel-amplitude critical-band and harmonic errors are also shown.

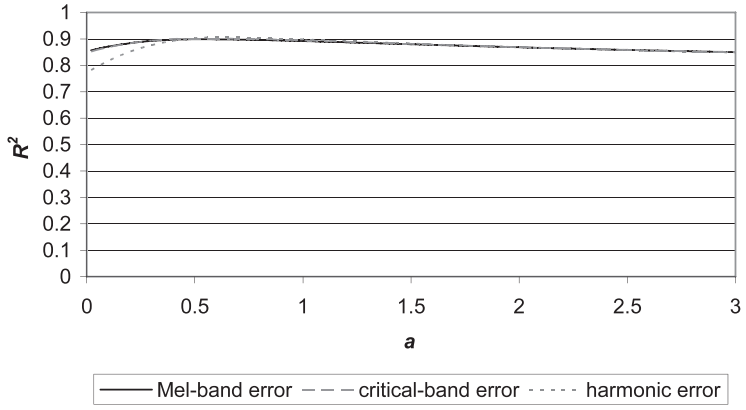


Fig. 7. R^2 versus a for relative-amplitude mel-band error [Eq. (8)]. Maximum correspondence is 90% at $a = 0.52$. Results for relative-amplitude critical-band and harmonic errors are also shown.

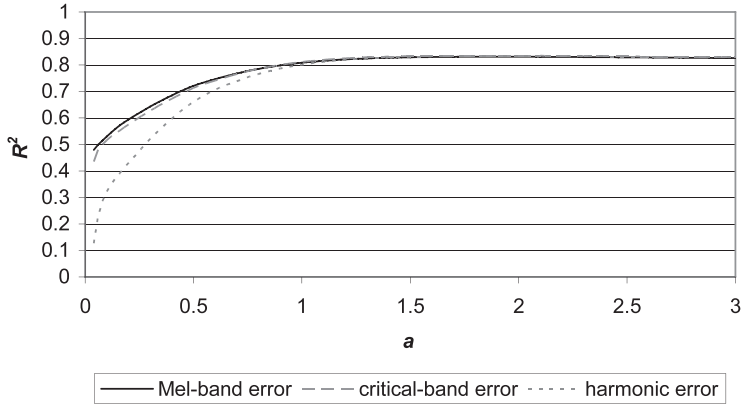


Fig. 8. R^2 versus a for maximum relative-amplitude mel-band error [Eq. (11)]. Maximum correspondence is 83% at $a = 1.88$. Results for maximum relative-amplitude critical-band and harmonic errors are also shown.

4.4 Sensitivity Analyses

4.4.1 Effects of Instruments

Fig. 10 superimposes the curves of R^2 versus a for the relative-amplitude mel-band error metric for the eight instruments (bassoon, clarinet, flute, horn, oboe, saxophone, trumpet, and violin). Inspection of Fig. 10 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 saxophone, flute, and oboe ($p < 0.001$). 2) This was followed by R^2 for bassoon, trumpet, clarinet, and violin. 3) R^2 values for bassoon and trumpet were not significantly different from each other ($p > 0.5$). 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 saxophone, flute, and horn—the extreme cases. The analyses described were repeated for the rms relative-amplitude mel-band error metric and similar trends were obtained.

4.4.2 Effects of Order of Regression Fit

Based on the assumption in our previous work that the discrimination versus error data would follow a curve with two inflection points [1], we decided to use a fourth-order regression fit. However, we also tried replacing the fourth-order regression function with third and fifth-order polynomials. Inspection of Fig. 11 indicates that there is only a very slight observable difference among the three regressions for the relative-amplitude mel-band error metric [Eq. (8)]. Further examination of the R^2 data indicates that the differences are less than 1%. The analyses were repeated for the rms relative-amplitude mel-band error metric [Eq. (12)], and similar results were obtained.

4.5 Effect of Varying the Number of Terms

We decided it would be interesting to see how varying the number of terms in two of the new metric formulas would affect correspondence. We also compared this with the relative-amplitude spectral error metric [see Eq. (2)], which in our previous study [1] yielded the best correspondence. Fig. 12 shows R^2 for this metric plotted versus the number of harmonics K for exponents $a = 0.5$, 1.0, 1.5, 2.0, 2.5, and 3.0. As K increases, the correspondence increases to 0.8 independent of a for

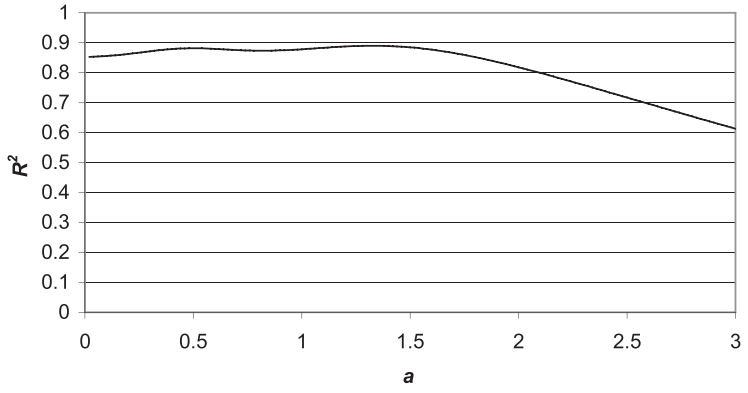


Fig. 9. R^2 versus a for MFCC error [Eq. (14)]. Maximum correspondence is 89% at $a = 1.32$.

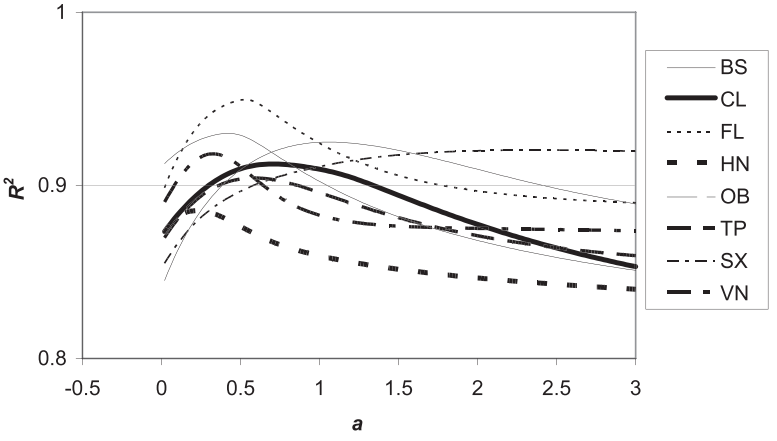


Fig. 10. R^2 versus a for relative-amplitude mel-band error [Eq. (8)] for each of the eight instruments (bassoon, clarinet, flute, horn, oboe, saxophone, trumpet, and violin).

only three harmonics and then begins to diverge, reaching about 0.85 for $0.5 \leq a \leq 1.0$ for five harmonics; it eventually converges to a maximum of 0.91 for $K > 20$.

Using the relative-amplitude mel-band error defined by Eq. (8) results in the graphs of Fig. 13, where R^2 is plotted versus the number of mel bands M for the same range of a values. The overall behavior is similar to that of the harmonic metric, but the rise for low values of M is much steeper and rises to about 0.85 for 10 bands, again for $0.5 \leq a \leq 1.0$, converging to a maximum of about 0.90 for $a = 0.5$ and $M > 22$.

Based on the metric of Eq. (14), R^2 is plotted versus the number of mel-frequency cepstral coefficients (MFCCs) in Fig. 14. Compared to the mel-band-based metrics this shows much greater divergence, depending on a. However, R^2 is quite independent of a for $0.5 \leq a \leq$

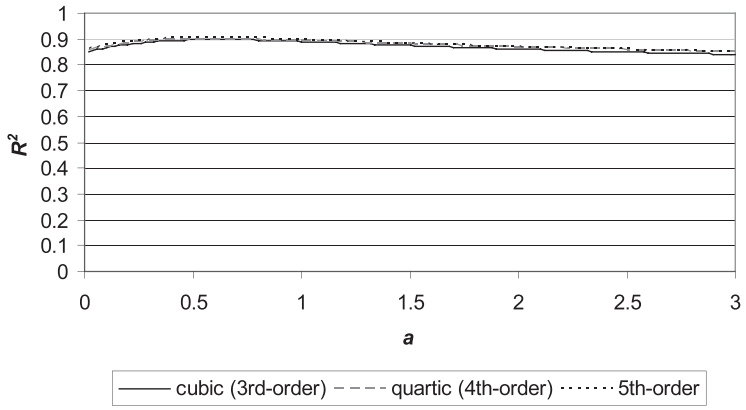


Fig. 11. R^2 versus a for relative-amplitude mel-band error fitted to data collected from all instruments using third, fourth, and fifth-order regression.

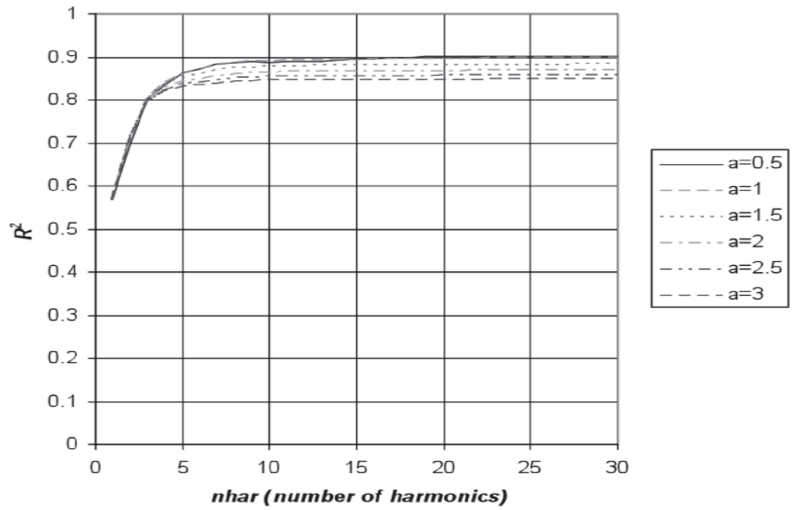


Fig. 12. R^2 correspondence versus number of harmonics for relative-amplitude spectral error metric [Eq. (2)] for several values of exponent a.

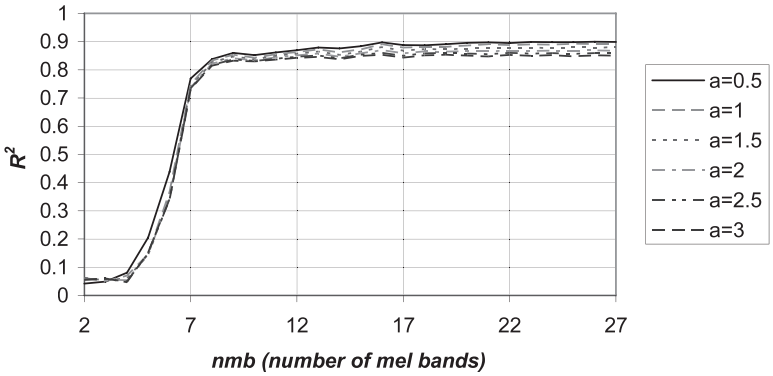


Fig. 13. R^2 correspondence versus number of mel bands for relative-amplitude mel-band error metric [Eq. (8)] for several values of exponent a.

1.5. It takes about 10 coefficients to reach $R^2 = 0.85$, and a maximum of 0.89 is reached for $M > 17$.

When comparing the three cases illustrated in Figs. 12–14 it can be seen that correspondences of 85% are achieved with five terms for harmonics, ten for mel bands, and ten for MFCCs.

5 DISCUSSION

It seemed likely that the mel-band and MFCC error metrics discussed in Section 3 would yield similar or possibly better correspondences with the discrimination data than the error metrics explored in our previous study [1]. Indeed we found very striking similarities in their correspondences, but no improvement in terms of the maximum correspondence was achieved.

Several of the mel-based 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 mel-band error [Eq. (8)] explains 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. Our previous study [1] also found that similarly formulated harmonic-based and critical-band-based error metrics performed best, and results for two of these metrics are included in Table 1 for comparison.

Three forms of relative-error normalization [Eqs. (8)–(10)] gave excellent performance, as did the rms relative-amplitude mel-band error [Eq. (12)]. The best results for linear-amplitude mel-band error, decibel-amplitude mel-band error, and MFCC error [Eq. (14)] were about as good as the relative mel-band error, but less robust in terms of sensitivity to the power a.

The mel-band-based error metrics did not improve on the critical-band metrics, and they were clearly worse than the decibel-amplitude critical-band error metric for larger values of a. Like the critical-band-based errors, the mel-band errors were much less sensitive to changes in the power a than errors based on harmonics. With one exception the mel-band-based errors were better for absolute differences ($a = 1$) than the Euclidean distances ($a = 2$). This was also true for the MFCC-based error.

When comparing the sensitivities of correspondences to a reduced number of terms used in the metric expressions for the harmonic [Eq. (2)] and MFCC [Eq. (14)] cases [Eq. (2)], we expected that for a given R^2 value MFCCs, which are designed to capture the overall structures of spectral envelopes in an efficient way, would require fewer terms than harmonics. Instead the results turned out the other way around, with only five harmonics required for 85% correspondence, as opposed to ten MFCCs for the same R^2. The other result, that it took ten mel bands to achieve 85%, is easily explained from Fig. 2, where it can be seen that ten mel bands encompasses four harmonics. The MFCCs, on the other hand, are all broadband (in our case encompassing 30 harmonics, or 9333 Hz), with each successive coefficient corresponding to increasing spectral detail. This suggests that the listeners were focusing on the first few harmonics rather than on the total spectrum when discriminating between similar musical sounds. Again, the caveat is that these results apply only to the 311.1-Hz fundamental frequency (and probably higher F_0 values) and might be quite different for lower fundamentals.

Nevertheless MFCCs have been favored for experiments in music information retrieval and instrument recognition [3], [6], [8]. The authors conjecture that this is because MFCCs based on the short-time Fourier transform work well when F_0 is highly time variant. For that case, to isolate harmonics it would be necessary to employ a general-purpose F_0 versus time detector. It is an open question whether isolating harmonics of time-

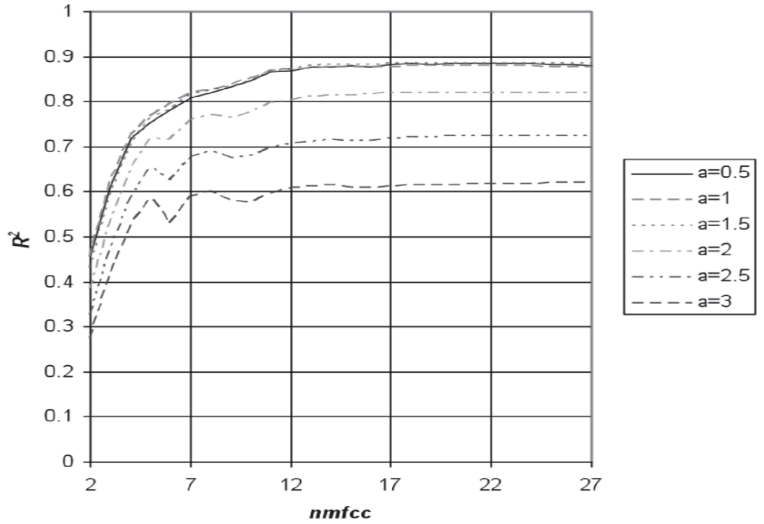


Fig. 14. R^2 correspondence versus number of mel-frequency cepstral coefficients for MFCC error metric [Eq. (14)] for several values of exponent a.

varying F_0 sounds will yield an advantage over using MFCCs in these applications.

6 CONCLUSIONS

All of the mel-band error metrics tested were found to have at least reasonable ranges of correspondence (80% or more) with the discrimination data, and several had excellent correspondences of approximately 90%. Three types of relative-amplitude mel-band error metrics and an rms relative-amplitude mel-band error metric achieved the best correspondences. These correspondences were almost identical to those of comparable critical-band errors, and, in general, they were more robust than harmonic-based error with respect to variations in the metric exponent a.

Absolute spectral differences ($a = 1$) outperformed Euclidean spectral differences ($a = 2$) on all metrics but one. This was especially true of the MFCC metric, with $a = 1.0$ about optimal. Though the peak correspondences of the critical-band error metrics were fractionally better than their mel-based counterparts, the differences were very small.

When the number of terms used for computing the metrics was reduced from the maxima (30 for harmonics, 27 for mel bands, and 27 for MFCCs) it was found that to achieve correspondences of 85%, the number of terms needed were five for harmonics, ten for mel bands, and ten for MFCCs (see Figs. 12–14). While the agreement between ten mels and five harmonics is easily explainable in terms of their overlap, as shown in Fig. 2, the need for more MFCCs than harmonics seems to indicate that listeners focus on the lower harmonics when discriminating similar musical sounds. This observation agrees with the results of Gunawan and Sen, who found that discrimination thresholds are governed by the first few harmonics [24].

In summary, for the single-tone case the authors were not able to find that mel-band-based or MFCC-based metrics offer any significant advantage over harmonic-based metrics for estimating the perception of spectral differences. It may be that these metrics will prove superior for stimuli with highly variable pitch or with several simultaneous pitches, and a study to explore this question is highly recommended for the future.

7 ACKNOWLEDGMENT

The authors would like to thank Simon Cheuk-Wai Wun for his excellent listening test design and Jenny Lim for managing the listening test. They would also like to thank Hiroko Terasawa, Mert Bay, Richard Lyon, and Dan Ellis for their discussions on mel frequency and mel-frequency cepstral coefficients and the anonymous reviewers for their helpful comments on the manuscript. This work was supported in part by the Hong Kong Research Grant Council under project 613508.

8 REFERENCES

[1] A. B. Horner, J. W. Beauchamp, and R. H. Y. So, "A Search for Best Error Metrics to Predict Discrimination of Original and Spectrally Altered Musical Instrument Sounds," *J. Audio Eng. Soc.,* vol. 54, pp. 140–156 (2006 Mar.).

[2] E. Zwicker and E. Terhardt, "Analytical Expressions for Critical-Band Rate and Critical Bandwidth as a

Table 1. Maximum R^2 for various mel-band metrics and MFCC-based error metric, and range of parameter a over which R^2 is within 5% of maximum R^2 value.

Error Metric	Maximum R^2	a Value of Lower Bound	a Value of Maximum R^2	a Value of Upper Bound
Linear-amplitude mel-band error	0.87	0.22	0.58	2.08
Decibel-amplitude mel-band error	0.89	0.02	0.80	2.16
Relative-amplitude mel-band error with simple normalization	0.90	0.02	0.52	2.76
Relative-amplitude mel-band error with dual normalization	0.90	0.02	0.58	2.60
Relative-amplitude mel-band error with maximum normalization	0.90	0.02	0.56	2.36
Maximum relative-amplitude mel-band error	0.83	0.84	1.88	3.00+
Rms relative-amplitude mel-band error	0.90	0.02	0.50	2.90
MFCC error	0.89	0.02	1.32	1.84
Relative-amplitude (harmonic) spectral error with simple normalization	0.91	0.30	0.64	2.24
Relative-amplitude critical-band error with simple normalization	0.90	0.06	0.54	2.52

Function of Frequency," *J. Acoust. Soc. Am.,* vol. 68, pp. 1523–1524 (1980).

[3] G. De Poli and P. Prandoni, "Sonological Models for Timbre Characterization." *J. New Music Res.,* vol. 26, pp. 170–197 (1997).

[4] B. Logan, "Mel Frequency Cepstral Coefficients for Music Modeling," *Proc. Int. Symp. on Music Information Retrieval (ISMIR)* (2000).

[5] J. J. Aucouturier and M. Sandler, "Segmentation of Musical Signals Using Hidden Markov Models," presented at the 110th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts),* vol. 49, pp. 541, 542 (2001 June), convention paper 5379.

[6] W. D'haes and X. Rodet, "Discrete Cepstrum Coefficients as Perceptual Features," in *Proc. 2003 Int. Computer Music Conf.* (Singapore, 2003), pp. 235–238.

[7] C. W. Weng, C. Y. Lin, and J. S. R. Jang, "Music Instrument Identification Using MFCC: Erhu as an Example," in *Proc. 9th Int. Conf. of the Asia Pacific Society for Ethnomusicology* (Phnom Penh, Cambodia, 2004), pp. 42–43.

[8] H. Terasawa, M. Slaney, and J. Berger, "Perceptual Distance in Timbre Space," in *Proc. 11th Mtg. of the Int. Conf. on Auditory Display* (Limerick, Ireland, 2005), pp. 61–68.

[9] W. Brent, "Perceptually Based Pitch Scales in Cepstral Techniques for Percussive Timbre Identification," in *Proc. 2009 Int. Computer Music Conf.* (2009), pp. 121–124.

[10] A. B. Horner, J. W. Beauchamp, and L. Haken, "Genetic Algorithms and Their Application to FM Matching Synthesis," *Computer Music J.,* vol. 17, no. 4, pp. 17–29 (1993).

[11] A. B. Horner, J. W. Beauchamp, and L. Haken, "Methods for Multiple Wavetable Synthesis of Musical Instrument Tones," *J. Audio Eng. Soc.,* vol. 41, pp. 336–356 (1993 May).

[12] A. B. Horner and J. W. Beauchamp, "Piecewise Linear Approximations of Additive Synthesis Envelopes: A Comparison of Various Methods," *Computer Music J.,* vol. 20, no. 2, pp. 72–95 (1996).

[13] R. Plomp, "Timbre as a Multidimensional Attribute of Complex Tones," in *Frequency Analysis and Periodicity Detection in Hearing,* R. Plomp and G. F. Smoorenburg, Eds. (Sijthoff, Eliden, The Netherlands, 1970), pp. 405–408.

[14] A. B. Horner, J. W. Beauchamp, and R. H. Y. So, "Detection of Random Alterations to Time-Varying Musical Instrument Spectra," *J. Acoust. Soc. Am.,* vol. 116, pp. 1800–1810 (2004).

[15] E. J. Pedhazur, *Multiple Regression in Behavioral Research* (Holt, Rinehart, and Winston, New York, 1982), chap. 3.

[16] J. W. Beauchamp, "Unix Workstation Software for Analysis, Graphics, Modification, and Synthesis of Musical Sounds," presented at the 94th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts),* vol. 4, p. 387 (1993 May), preprint 3479.

[17] J. W. Beauchamp, "Analysis and Synthesis of Musical Instrument Sounds," in *Analysis, Synthesis, and Perception of Musical Sounds: The Sound of Music,* J. W. Beauchamp, Ed. (Springer, New York, 2007), pp. 1–89.

[18] B. C. J. Moore, B. R. Glasberg, and T. Baer, "A Model for the Prediction of Thresholds, Loudness, and Partial Loudness," *J. Audio Eng. Soc.,* vol. 45, pp. 224–240 (1997 Apr.).

[19] S. S. Stevens and J. Volkmann, "The Relation of Pitch to Frequency: A Revised Scale," *Am. J. Psychol.,* vol. 53, pp. 329–353 (1940).

[20] S. Umesh, L. Cohen, and D. Nelson, "Fitting the Mel Scale," in *Proc. 1999 IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP '99),* pp. 217–220.

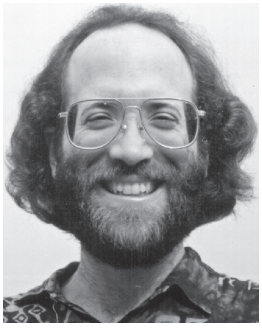
[21] D. O'Shaughnessy, *Speech Communications: Human and Machine,* 2nd ed. (IEEE Press, New York, 2000), pp. 128, 214.

[22] S. McAdams, J. W. Beauchamp, and S. Meneguzzi, "Discrimination of Musical Instrument Sounds Resynthesized with Simplified Spectrotemporal Parameters," *J. Acoust. Soc. Am.,* vol. 105, pp. 882–897 (1999).

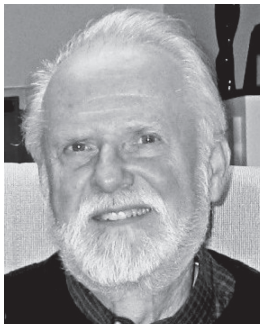
[23] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition* (Prentice-Hall, Englewood Cliffs, NJ, 1993), pp. 189–190.

[24] D. Gunawan and D. Sen, "Spectral Envelope Sensitivity of Musical Instrument Sounds," *J. Acoust. Soc. Am.,* vol. 123, pp. 500–506 (2008).

THE AUTHORS



A. Horner



J. Beauchamp

R So

Andrew Horner received a Ph.D. degree in computer science from the University of Illinois at Urbana-Champaign.

Dr. Horner is a professor in the Department of Computer Science at the Hong Kong University of Science and Technology. His research interests include music synthesis, musical acoustics of Asian instruments, peak factor reduction in musical signals, music in mobile phones, and spectral discrimination.

●

James Beauchamp received Bachelor of Science and Master of Science degrees in electrical engineering from the University of Michigan in 1960 and 1961, respectively, and a Ph.D. degree in electrical engineering from the University of Illinois at Urbana-Champaign (UIUC) in 1965.

In 1965 he joined the electrical engineering faculty at UIUC. From 1968 to 1969 he was a research associate at the Stanford University Artificial Intelligence Laboratory, Stanford, CA. In 1969 he returned to UIUC and assumed a joint appointment in music and electrical and computer engineering. While on the UIUC faculty he taught courses in musical acoustics, computer music, and audio engineering. In 1988 he was a visiting scholar at the Center for Computer Research in Music and Acoustics at Stanford (CCRMA). During 1994–1995 he was a visiting researcher at the Institut de Recherche et Coordination in Acoustique/Musique (IRCAM) in Paris, France. In 1997 he retired from UIUC but has continued his affiliation with it as professor emeritus. His current research interests are in sound analysis algorithms, sound synthesis models, musical timbre perception, automatic pitch detection, and musical sound source separation.

Dr. Beauchamp is a Fellow of the Audio Engineering Society and of the Acoustical Society of America.

●

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.