In Search of a Perceptual Metric for Timbre: Dissimilarity Judgments among Synthetic Sounds with MFCC-Derived Spectral Envelopes

HIROKO TERASAWA,[1,2] *AES Member,* **JONATHAN BERGER**[3]**, AND SHOJI MAKINO**[1]
(terasawa@tara.tsukuba.ac.jp) (brg@ccrma.stanford.edu) (maki@tara.tsukuba.ac.jp)

[1]*Life Science Center of TARA, University of Tsukuba 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8577, Japan*
[2]*JST, PRESTO (Information Science and Humans) 7 Gobancho, Chiyoda-ku, Tokyo 102-0076, Japan*
[3]*CCRMA, Department of Music, Stanford University 660 Lomita Drive, Stanford, CA 94305, USA*

This paper presents a quantitative metric to describe the multidimensionality of spectral envelope perception, that is, the perception specifically related to the spectral element of timbre. Mel-cepstrum (Mel-frequency cepstral coefficients or MFCCs) is chosen as a hypothetical metric for spectral envelope perception due to its desirable properties of linearity, orthogonality, and multidimensionality. The experimental results confirmed the relevance of Mel-cepstrum to the perceived timbre dissimilarity when the spectral envelopes of complex-tone synthetic sounds were systematically controlled. The first experiment measured the perceived dissimilarity when the stimuli were synthesized by varying only a single coefficient from MFCC. Linear regression analysis proved that each of the 12 MFCCs has a linear correlation with spectral envelope perception. The second experiment measured the perceived dissimilarity when the stimuli were synthesized by varying two of the MFCCs. Multiple regression analysis showed that the perceived dissimilarity can be explained in terms of the Euclidean distance of the MFCC values of the synthetic sounds. The quantitative and perceptual relevance between the MFCCs and spectral centroids is also discussed. These results suggest that MFCCs can be a metric representation of spectral envelope perception, where each of its orthogonal basis functions provides a linear match with human perception.

0 INTRODUCTION

The spectral envelope of a sound is a crucial aspect of timbre perception. In this study, we propose a quantitative model of *spectral envelope perception*, that is, the spectral element in the timbre perception, with a set of orthogonal basis functions. The goal of this work is to develop a quantitative mapping between a physical description of the spectral envelope and its perception, with the purpose of controlling timbre in sonification in a meaningful and reliable way. The model suggests a systematic description of spectral envelope perception whose simplicity may be seen as analogous to the three primary colors in the visual system.

In the earliest studies of timbre perception, Helmholtz speculated that the spectral envelope is the source of the timbre variations [1]. For speech sounds, the formant structure of the overtone series was determined to be the key factor in differentiating vowels [2], [3]. For Western musical-instrument sounds, timbre perception has often been described in terms of the spectral centroid, spectral flux, and attack time [4]–[7]. In addition to these factors, other factors such as amplitude and frequency micromodulations and inharmonicity are also taken into account [8]. Although these descriptive studies can address the relationship between the physical aspects of sound and the perception, more information on the precise shape of the spectral envelope is often needed to synthesize sounds in a controlled way. In other words, although there are multiple layers (i.e., perceptual, cognitive, physical, and social perspectives) in addressing sound quality [9], understanding at one layer does not necessarily lead to the improvement at another layer. Recent studies on morphed instrumental sounds employed the time-varying multiband approach to evaluate the perception of the synthesized timbre, connecting these multiple layers [10]–[12].

A robust quantitative model for timbre perception has been long desired for the control of timbre in sound synthesis, especially in relation to the use of sound in auditory displays of information. To take full advantage of the multidimensionality of timbre in sonification, we need a quantitative, multidimensional description for spectral

envelope perception. Such a model allows reliable mappings of data to perceptual space, which is critical for effective sonification [13]. Many researchers have conceptualized spectral envelope perception by analogy with the visual color system, by finding an orthogonal basis in the spectral shapes of instrumental sounds [14], by proposing the concept of sound color [15], and by visualizing organ sounds as an energy-balance transition across three frequency regions [16].

In this work, we aim for a simple, quantitative, and multidimensional model that can be extended to synthesize perceptually meaningful variations of spectral envelopes. Ideally, such a model will predict the spectral envelope perception in a linear and orthogonal manner; each orthogonal basis should have a quantitative label that can linearly represent the perceived difference, and the perception of a complex spectral envelope could be explained in terms of the superposition of these basis functions.

Seeking such a model for spectral envelope perception, we chose the Mel-cepstrum (also known as Mel-frequency cepstrum coefficients or MFCCs) for the following reasons: (1) MFCCs are constructed by a set of orthogonal basis functions, therefore satisfying the need for an orthogonal model; (2) MFCCs are based on perceptually relevant scalings, which can provide a linear mapping between the numeric description and the perception; and (3) MFCCs have been a powerful front-end tool for many engineering applications, and clarifying the perceptual characteristics of MFCCs by performing psychoacoustic experiments is valuable.

The Mel-cepstrum was originally proposed as "the description of short-term spectra ... in terms of the contribution to the spectrum of each of an orthogonal set of spectrum-shape functions" [17]. The Mel-cepstrum is computed by applying a discrete cosine transform (DCT) to the output of a simple auditory filterbank that roughly resembles critical bands. Unlike other representations of spectral envelope, such as the 1/3-octave-band models or specific loudness, the basis functions of a Mel-cepstrum are mathematically orthogonal. Mermelstein noted that a Mel-cepstrum can constitute a distance metric that reflects the perceptual space of phonemes [18] and examined its efficiency as a front end for automatic speech recognition [19]. Now it is considered to be the classic front-end algorithm for automatic speech recognition [20]. Its application has been extended to timbre-related music information retrieval [21], [22], sound database indexing based on timbre characteristics [23], [24], timbre control for sonification [25], perceptual description of instrumental sound morphing [26], and a proposal that timbre perception be represented in terms of sound color and sound density [27].

Despite such numerous applications, the authors' earlier works were the first to examine the Mel-cepstrum's perceptual characteristics with psychoacoustic experiment procedures [28]–[30], and, before that, the perceptual relevance of MFCCs was demonstrated only by applications. Therefore, it is worthwhile to examining the perceptual characteristics of MFCCs in detail using psychoacoustic experiments. Still, Mel-cepstrum is not the most precise auditory model. Other perceptual models, such as specific loudness [31], the spatiotemporal receptive-field model [32], and the Mellin transform [33] may seem to be better options. However, these models do not consist of orthogonal basis functions, and they are not necessarily a compact algorithm that enables efficient analysis and synthesis of timbre. For these reasons, MFCCs were considered the most suitable for a spectral envelope perception model.

We employed the following framework to test this model. We first synthesized a stimulus set with gradually changing spectral envelopes by varying the Mel-cepstrum values in a stepwise order, while keeping the temporal characteristics constant across the stimuli. The participants listened to the stimuli in pairs and provided dissimilarity ratings. Finally, the relationship between the dissimilarity ratings and the Euclidean distance of the MFCC values was analyzed with a linear regression.

To measure spectral envelope perception, the temporal characteristics of the stimuli must be strongly controlled because the temporal structure has a strong effect on timbre perception. To control this effect, we decided to use the same temporal structure for all of the stimuli. Although it might seem more interesting to employ various kinds of temporal structures in a single experiment, it would not allow us to observe the multidimensionality of spectral envelope perception accurately. In musical instrument timbre studies, Plomp detected three dimensions for spectral envelope perception when he minimized the variation in the temporal structure [14], whereas other researchers detected only a single dimension (spectral centroid) dedicated solely to the spectral envelope, in addition to another spectro-temporal dimension (spectral flux) when they introduced various temporal structures [4]–[7]. Therefore, we decided to maintain a single kind of temporal structure for the entire stimuli set.

In designing the temporal structure of the stimuli, we wanted to create tones with a distinct quality that helped the participants make reliable judgments. For this purpose, the stimuli are desirably sustained and have the fewest random factors. The simplest design that satisfies this criterion is obviously the addition of sinusoids in a harmonic series. But this design has an unwanted effect: when the spectral envelope is manipulated, the amplified partials are perceived as obtrusive and separated from the other partials.

To avoid this perceptual segregation, we added a vibrato-like frequency modulation to all the harmonics, so that all of the partials contribute to a unified tone thanks to the "common fate" effect [34]. With this vibrato, the synthesized sounds exhibited a voice-like quality that is more natural than sinusoid beeps. Because parameter-mapping sonification can sound unpleasant [35], such naturalness is valuable. As already shown in voice-based sonification projects, voice-like qualities often facilitate the comprehension of data [36], [37]. However, stimuli with vibrato may be unacceptable for the experiment because vibrato might influence spectral envelope perception due to its dramatic musical effect, which is particular to Western operatic singing. But, in fact, adding vibrato to a voice does not

change the perceived vowel [38], and people can distinguish subtle changes in the spectral envelope of the tones with vibrato [39]. This means that adding vibrato does not interfere with the perception of the spectral envelope and that, therefore, the use of vibrato for the experiment stimuli is acceptable. Furthermore, we expect that the inclusion of vibrato implies a musical setting and encourages the participants to engage in "musical listening" with greater attention to timbre.

Using these stimuli, we conducted two experiments in the experimental framework described above: the first was designed to test the perceptual effect when modifying a single dimension from MFCC, and the second to test the orthogonality of the timbre space using two dimensions from MFCC. We used linear regression to analyze our data because we were explicitly investigating the relationship between MFCC and subjective ratings, rather than exploring unknown dimensions that could be discovered with the multidimensional scaling (MDS) method.

This paper aims to show (1) that there is a linear relationship between each of the Mel-cepstrum orthogonal functions and the perceived timbre dissimilarity, (2) that the multidimensionality of complex spectral envelope perception can be explained in terms of the Euclidean distance of the orthogonal function coefficients, and (3) that the widely used Mel-cepstrum can form a valid representation of spectral envelope perception. However, the multidimensionality of spectral envelope perception beyond two dimensions and the temporal aspect of timbre perception remain outside the scope of this study.

In the following sections, we describe the method we used to synthesize the stimuli while varying the MFCC values in a controlled way. We describe our two experiments on spectral envelope perception and their result followed by a discussion and our conclusion.

1 MFCC-BASED SOUND SYNTHESIS

1.1 Mel-Cepstrum

The MFCC is the DCT of a modified spectrum, in which its frequency and amplitude are scaled logarithmically. Of the various implementations that exist, the Mel-cepstrum algorithm from Auditory Toolbox [40] was employed. The spectrum is first processed with a filterbank of 32 channels, which roughly approximate the spacing and bandwidth of the auditory system's critical bands. The frequency response of the filterbank $H_i(f)$ is shown in Fig. 1, and the passband of each triangular window $H_i(f)$ is shown in Eq. (1). The amplitude of each filter is normalized so that each channel has unit power gain.

$$Bandwidth\,(H_i) = \begin{cases} 200.0 & (i = 1) \\ 133.3 & (1 < i \leq 13). \\ 1000 \cdot 1.072^{i-13} & (i > 13) \end{cases} \quad (1)$$

The filterbank, whose triangular frequency response is shown in Fig. 1, is applied to the sound in the frequency domain, and provides the filterbank output, F_i:

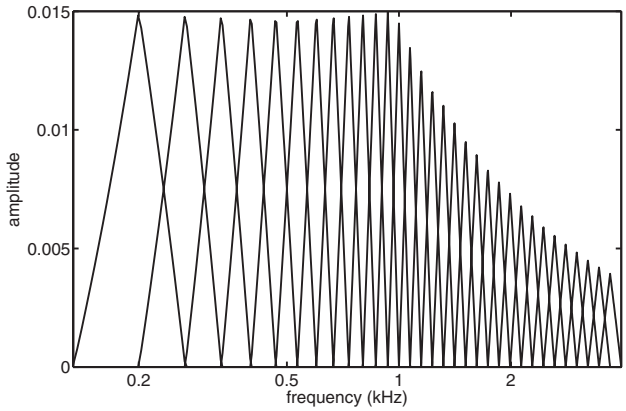


Fig. 1. Frequency response of the filterbank used for the MFCC. The sound spectrum is first processed with this filterbank, which roughly approximates the characteristics of auditory critical bands. Taking the lower coefficients from the DCT of this filterbank output yields MFCC.

$$F_i = \int_{f=f_{i\text{-low}}}^{f_{i\text{-high}}} H_i(f) \cdot S(f)\,\mathrm{d}f, \quad (2)$$

where i is the channel number in the filterbank, f is the frequency, $H_i(f)$ is the filter response of the ith channel, and $S(f)$ is the absolute value of the discrete Fourier transform of a signal. $f_{i\text{-low}}$ and $f_{i\text{-high}}$ denote the lowest and highest frequency bins, respectively, of the passband of the ith channel filter.

The MFCCs, C_i, are computed by taking the DCT of the log-scaled filterbank output:

$$L_i = \log_{10}(F_i), \quad (3)$$

$$C_n = w_n \sum_{i=1}^{I} L_i \cdot \cos \frac{\pi(2i-1)\cdot(n-1)}{2I}, \quad (4)$$

where $w_0 = 1/\sqrt{I}$, $w_n = \sqrt{2/I}$ for $1 \leq n \leq N-1$. I and N represent the total number of filters and the total number of Mel-cepstrum coefficients, respectively. Taking 13 lower coefficients from C_n, the set of coefficients from C_0 to C_{12} is called the MFCC which summarizes the spectral envelope.

1.2 Sound Synthesis

The sound synthesis for the stimuli has two stages: (1) the spectral envelope is created by the pseudo-inverse transform of the Mel-cepstrum, and (2) an additive synthesis of sinusoids is performed using the spectral envelope generated earlier.

1.2.1 Pseudo-Inversion of MFCC

As described above, the MFCC takes only the 13 lower coefficients, and therefore it is a lossy transform from a spectrum. The inversion of the MFCC is not possible in a strict sense. This section describes the pseudo-inversion

of the MFCC, which generates a smooth spectral envelope from a given Mel-cepstrum.

The generation of the spectral envelope starts with a given array of Mel-cepstrum coefficients C_n, which is an array of 13 coefficients. The reconstruction of the spectral shape from the MFCC starts with the inverse discrete cosine transform (IDCT) and amplitude scaling:

$$\tilde{L}_i = \sum_{n=1}^{N} w_n \cdot C_n \cdot \cos \frac{\pi(2i-1)\cdot(n-1)}{2I}, \tag{5}$$

$$\tilde{F}_i = 10^{\tilde{L}_i}. \tag{6}$$

In this pseudo-inversion, the reconstructed filterbank output $\tilde{F}_i$ is considered to represent the value of the reconstructed spectral envelope $\tilde{S}(f)$ at the center frequency of each channel from the filter bank,

$$\tilde{S}(f_i) = \tilde{F}_i, \tag{7}$$

where f_i is the center frequency of the ith auditory filter. Therefore, to obtain a reconstruction of the entire spectrum, $\tilde{S}(f)$, a linear interpolation was applied to the values between the center frequencies $\tilde{S}(f_i)$.

1.2.2 Additive Synthesis

The voice-like stimuli used in this study are synthesized using additive sinusoidal synthesis. The reconstructed spectral envelope $\tilde{S}(f)$ determines the amplitude of each sinusoid. A slight amount of vibrato is added to give some coherence and life to the resulting sound.

In the synthesis, a harmonic series is prepared, and the level of each harmonic is weighted based on the desired smooth spectral shape. The pitch, or fundamental frequency f_0, is set at 200 Hz, with the frequency of the vibrato v_0 set at 4 Hz and the sampling rate at 8 kHz.

Using the reconstructed spectral shape $\tilde{S}(f)$, the additive synthesis of the sound is accomplished as follows:

$$s(t) = \sum_{q=1}^{Q} \tilde{S}(f_{\text{inst}}(q,t)) \cdot \sin(2\pi q f_0 t + 1 - q\cos 2\pi v_0 t), \tag{8}$$

where q specifies the qth harmonic of the harmonic series. The total number of harmonics Q is 19, and all the harmonics stay under the Nyquist frequency of 4 kHz. The amplitude of each harmonic is determined by using a lookup table of $\tilde{S}(f)$ and the instantaneous frequency f_{inst}, which is defined as follows:

$$f_{\text{inst}}(q,t) = q f_0 + q v_0 \cdot \sin 2\pi v_0 t. \tag{9}$$

The fundamental frequency $f_0 = 200$ (Hz) is determined from the range of 180–230 Hz (the fundamental frequency of the female voice), so that the MFCC of the resulting sound maintains the intended stepwise or grid structure the best.

The duration of the resulting sound s is 0.75 s. For the first 30 ms of the sound, its amplitude is linearly fading in, and for the last 30 ms, its amplitude is linearly fading out. All the stimuli are scaled with an identical scaling coefficient. The specific loudness [31] of all the stimuli showed a very small variance, and their loudness was considered to be fairly similar within the stimuli set. For all of the 144 stimuli synthesized for this study, 123 stimuli scored under 3%, 10 stimuli scored 3–6%, and 7 stimuli scored 6–8% loudness deviations when compared with the mean loudness of all the stimuli.

2 EXPERIMENT 1: SPECTRAL ENVELOPE PERCEPTION OF SINGLE-DIMENSIONAL MFCC FUNCTION

2.1 Scope

This experiment considers the linear relationship between spectral envelope perception and each coefficient from the Mel-cepstrum, namely, a single function from the orthogonal set of spectral envelope functions. Following the sound-synthesis method described in the previous section, when a coefficient from Mel-cepstrum changes gradually in a linear manner while the other coefficients are kept constant, the spectral envelope of the resulting sound holds a similar overall shape, but the humps of the envelope change their amplitudes exponentially. In the experiment, it was examined whether the Mel-cepstrum can linearly represent the spectral envelope perception, and all 12 coefficients from Mel-cepstrum were tested based on this framework. The experiment was granted the approval for human-subject research by the Stanford University Institutional Review Board.

2.2 Method

2.2.1 Participants

Twenty-five participants (graduate students and staff members from the Center for Computer Research in Music and Acoustics at Stanford University) volunteered for the experiment. The participants were aged 20–35 years old, and had a musical background (majoring or minoring in music in college and graduate school), and/or an audio engineering background (enrolled in a music technology degree program). They all described themselves as having normal hearing. We conducted a pilot study with Japanese engineering students, and confirmed that the experimental results did not depend significantly on the participant group.

2.2.2 Stimuli

Twelve sets of synthesized sounds were prepared. The set n is associated with the MFCC coefficient C_n, the stimuli set 1 consists of the stimuli with C_1 varied, and the stimuli set 2 consists of the stimuli with C_2 varied, and so on. Although C_n is increased from zero to one with five levels, namely, $C_n = 0, 0.25, 0.5, 0.75, 1.0$, to form a stepwise structure, the other coefficients are kept constant, that is, $C_0 = 1$ and all the other coefficients are set at zero.

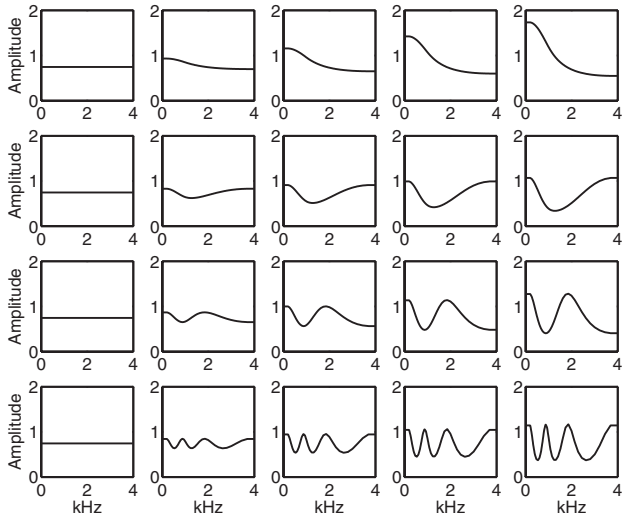


Fig. 2. Spectral envelopes generated by varying a single Mel-cepstrum coefficient. The first row shows the spectral envelopes when C_1 from MFCC was varied from 0 to 1 with five steps (0, 0.25, 0.5, 0.75, and 1.0). The second, third, and fourth rows correspond, respectively, to cases where C_2, C_3, and C_6 from MFCC were varied in the same manner.

For example, stimuli set 4 consists of five stimuli based on the following parameter arrangement:

$$C = [1, 0, 0, 0, C_4, 0, ..., 0], \tag{10}$$

where C_4 is varied with five levels:

$$C_4 = [0, 0.25, 0.5, 0.75, 1.0]. \tag{11}$$

Fig. 2 illustrates the idea of varying a single coefficient of MFCC, and the resulting set of the spectral envelopes for the cases of varying C_1,C_2, C_3, and C_6.

2.2.3 Procedure

The experiment had 12 sections, one for each of the 12 sets of stimuli. Each section consisted of a practice phase and an experimental phase.

The task of the participants was to listen to a pair of stimuli that were played in sequence with a short intervening silence, and to rate the perceived timbre dissimilarity of the presented pair. They rated the perceived dissimilarity on a scale of 0 to 10, with 0 indicating that the presented pair of sounds were identical, and 10 indicating that they were the most different within the section.

The participants pressed the "Play" button of the experiment GUI to play a sound, and reported the dissimilarity rating using a slider on the GUI. To facilitate the judgment, the pair with the largest spectral envelope difference in the section (i.e., the pair of stimuli with the lowest and highest, $C_n = 0$ and $C_n = 1$, is assumed to have a perceived dissimilarity of 10) was presented as a reference pair throughout the practice and experimental phases. Participants were allowed to listen to the test pair and the reference pair as many times as they wanted, but were advised not to repeat this too many times before making their final decision on scaling and proceeding to the next pair.

In the practice phase, five sample pairs were presented for rating. In the experimental phase, 25 pairs per section (all the possible pairs from five stimuli) were presented in a random order. The order of presenting the sections was also randomized. The participants were allowed to take a break as they wished.

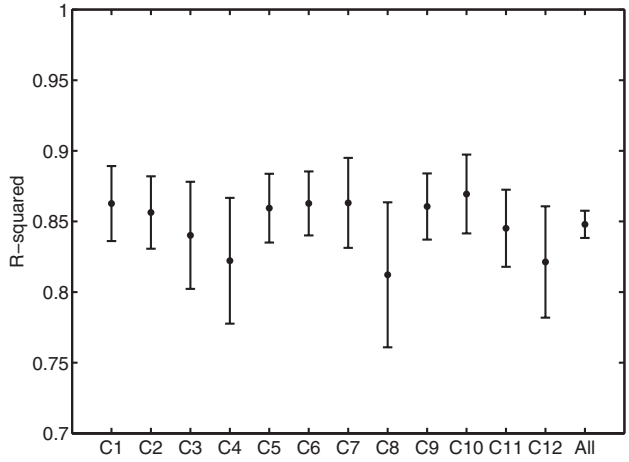


Fig. 3. Coefficients of determination (R^2) from the linear regression analysis of Experiment 1 with 95 % confidence intervals for each of the 12 Mel-cepstrum coefficients, C_n, and for the average of all the coefficients.

2.3 Linear Regression Analysis

The dissimilarity judgments were analyzed using simple linear regression [41], with absolute C_n differences as the independent variable, and their reported perceived dissimilarities as the dependent variable. The coefficient of determination R^2 represents the goodness of fit in the linear regression analysis. The linear regression analysis was individually applied for each section and each participant, because it is anticipated that every listener could respond differently to the stimuli sets, which would result in the deviation of the regression coefficients. With a quantile–quantile plot, the R^2 values formed a straight line except for a very few outliers with low R^2 values, showing that the distribution of the R^2 values is close to normal.

After the linear regression, the R^2 values for one section from all the participants were averaged to find the mean degree of fit (mean R^2) of each section. The mean R^2 among the participants was used to judge the linear relationship between the C_n distance and the perceived dissimilarity.

The mean R^2 and the corresponding confidence interval are plotted in Fig. 3. The mean R^2 for all the responses was 85%, with the confidence intervals for all the sections overlapped. This means that all of the coefficients, from C_1 to C_{12}, have a linear correlation with the perception of sound color with a statistically equivalent degree of fit, when an experiment is performed on an individual coefficient independent of other coefficients.

3 EXPERIMENT 2: SPECTRAL ENVELOPE PERCEPTION OF TWO-DIMENSIONAL MFCC SUBSPACE

3.1 Scope

This experiment tested the spectral envelope perception of the two-dimensional MFCC subspace. The stimuli set was synthesized by varying two coefficients from the Mel-cepstrum, say C_n and C_m, to form a two-dimensional subspace. The subjective response to the stimuli set was tested based on the Euclidean space hypothesis, namely, that each coefficient functions as an orthogonal basis when estimating the spectral envelope perception. As it is not realistic to test all of the 144 two-dimensional subspaces, five two-dimensional subspaces were chosen for testing. The experiment was approved for human subject research by the Stanford University Institutional Review Board.

3.2 Method

3.2.1 Participants

Nineteen participants, who were audio engineers, administrative staff members, visiting composers, and artists from the Banff Centre, Alberta, Canada, volunteered for this experiment. The participants were aged 25–40 years old, and they had a strong interest in music, with many of them having received professional training in music and/or audio engineering. All of them described themselves as normal-hearing.

3.2.2 Stimuli

Five sets of synthesized sounds were prepared that were associated with the five different kinds of two-dimensional subspaces. The five subspaces were made by varying [C_1, C_3], [C_3, C_4], [C_3, C_6], [C_3, C_{12}], and [C_{11}, C_{12}], respectively. For each set, the coefficients in question were independently varied over four levels ($C_n = 0$, 0.25, 0.5, 0.75, and $C_m = 0$, 0.25, 0.5, 0.75) to form a grid-like structure; the other coefficients were kept constant, that is, $C_0 = 1$ and all other coefficients were set at zero. By varying two coefficients independently, over four levels, each set had 16 synthesized sounds.

For example, the first set made of the subspace [C_1, C_3] consists of the 16 sounds based on the following parameter arrangement:

$$C = [1, C_1, 0, C_3, 0, ..., 0], \tag{12}$$

where C_1 and C_3 were varied over four levels, creating a grid with two variables.

The subspaces were chosen with the intention of testing the spaces made of: nonadjacent low to middle coefficients ([C_1, C_3] and [C_3, C_6]); two adjacent low coefficients ([C_3, C_4]); low and high coefficients ([C_3, C_{12}]); and two adjacent high coefficients ([C_{11}, C_{12}]).

Fig. 4 shows an example of the generated spectral envelopes for this experiment.

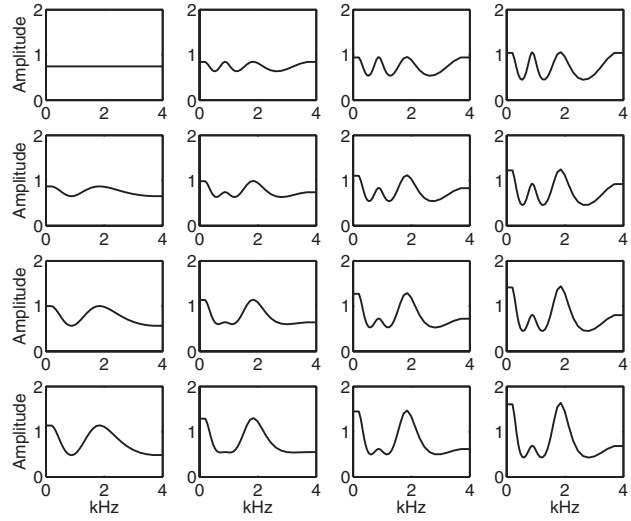


Fig. 4. Spectral envelopes generated by varying two Mel-cepstrum coefficients. The horizontal direction (left to right) corresponds to incrementing C_6 from 0 to 0.75 in four steps (0, 0.25, 0.5, and 0.75), and the vertical direction (top to bottom) corresponds to incrementing C_3 from 0 to 0.75 in four steps. For example, the top-left subplot shows the spectral envelope when $C_6 = C_3 = 0$, and the bottom-right subplot is when $C_6 = C_3 = 0.75$.

3.2.3 Procedure

There are 16 stimuli sounds per one subspace, making 256 possible stimulus pairs. Because testing all the pairs would take too much time and exhaust the participants, it was necessary to reduce the number of the stimulus pairs in the experiment. The strategies for reducing the test pairs were (1) test *either* AB or BA ordering when measuring the perceived difference of stimuli A and B, instead of measuring the perception for *both* AB and BA; and (2) test only some interesting pairs instead of testing all the possible combinations of stimulus pairs. We adopted the first strategy, and the actual order for a stimulus pair in the experiment was randomly selected from AB and BA ordering. However, the selection of ordering for each stimulus pair was not varied across the participants.

To employ the first strategy, it was necessary to evaluate whether the ordering of the stimuli had a significant effect on the perceived dissimilarity of the spectral envelope. To compare the AB responses and BA responses, equivalence testing was conducted based on confidence intervals [42]. First, regression analyses with AB order and BA order were separately conducted for each section and each participant. Then the difference between the R^2 values of AB and BA order regressions for each section was calculated. After that, for each section, the mean and the confidence intervals for the R^2 differences were calculated across participants. The confidence intervals of the differences for each section were 2–3.5%, falling into the predefined 5% minimum difference range. This reveals that the regression analyses based on AB responses and BA responses were statistically equivalent. Because of this equivalency, it was decided that presenting only one of two possible directions of a stimulus pair was sufficient.

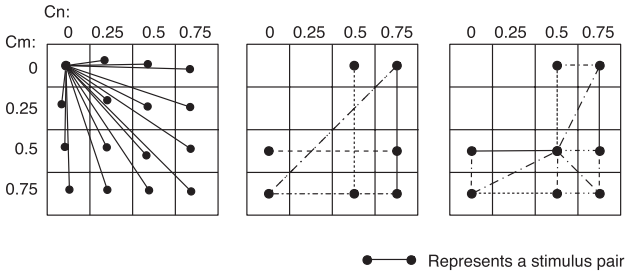


Fig. 5. Selection of the test pairs for the two-dimensional MFCC subspace experiment. Left: 16 pairs to examine distances from the origin. Middle: 5 pairs to examine large distances. Right: 13 pairs to examine some shorter parallel and symmetric distances.

Even after halving the number of stimulus pairs, there were still too many and further reduction was needed. Therefore, some pairs were chosen to represent large and small distances with some geometric order in the parameter subspace. Within each subspace, the test pairs were selected with the following interests, resulting in the total of 34 test pairs per section:

- From the zero of the space $C_n = C_m = 0$ to all the nodal points of the grid on the parameter subspace (16 pairs);
- Other large distances (5 pairs);
- Some shorter parallel and symmetric distances to test if they have similar perceived dissimilarities (13 pairs).

The final configuration of the test pairs is presented in Fig. 5 .

The participants' task was to listen to the paired stimuli, which were played in sequence with a short intervening silence, and to rate the perceived timbre dissimilarity of the presented pair using a 0 to 10 scale. Here 0 indicates that the paired stimuli were identical, and 10 indicates that the perceived dissimilarity between the paired stimuli was the largest in the section.

The participants reported the dissimilarity rating using a slider on the experiment's GUI. To facilitate the judgment, the pair with the greatest spectral envelope difference in the section is presented as a reference pair throughout the practice and experimental phases, assuming that the pair of stimuli with the lowest and highest, $C_n = C_m = 0$ and $C_n = C_m = 0.75$, would have a perceived dissimilarity of 10 within the stimuli set. Participants were allowed to listen to the test pair and the reference pair as many times as they wanted, but they were advised not to repeat this too many times before making their final decision on scaling and proceeding to the next pair.

In the practice phase, five sample pairs were presented for rating. In the experimental phase, 34 pairs per section were presented in a random order. The order of presenting the sections was also randomized. The participants were allowed to take breaks as they wished.

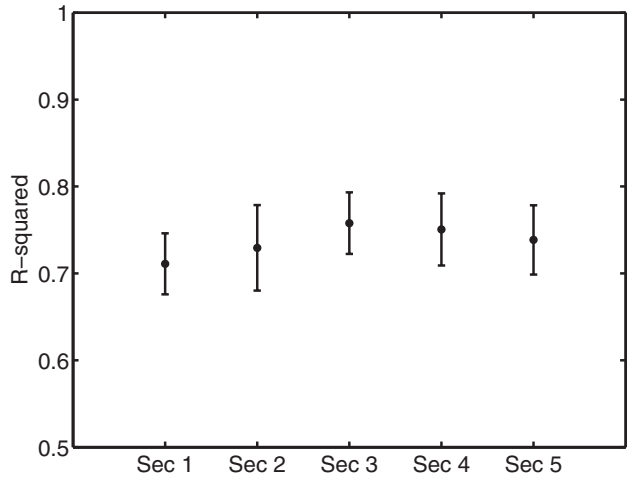


Fig. 6. Coefficient of determination (R^2) from the regression analysis of the two-dimensional sound color experiment with 95% confidence interval. Sections 1–5 represent the tests on subspaces $[C_1, C_3]$, $[C_3, C_4]$, $[C_3, C_6]$, $[C_3, C_{12}]$, and $[C_{11}, C_{12}]$, respectively.

3.3 Linear Regression Analysis

The dissimilarity judgments were analyzed using linear regression. The orthogonality of the two-dimensional subspaces was tested with a Euclidean distance-based model: the independent variable is the Euclidean distance of the MFCC between the paired stimuli, and the dependent variable is the subjective dissimilarity rating:

$$d^2 = ax^2 + by^2, \tag{13}$$

where d is the perceptual distance that subjects reported in the experiment, x and y are the respective differences between the C_n and C_m values of the paired stimuli. This model reflects the idea that the perceptual distance should be described in terms of the Euclidean distance of the spectral-envelope description vectors. The standard least-squares estimation is used with the linear regression analysis. The coefficient of determination, R^2, represents the goodness of fit in the linear regression analysis.

Individual linear regression for each section and each participant was applied first, and the R^2 values of one section from all the participants were then averaged to find the mean degree of fit (mean R^2) of each section. The mean R^2 among the participants is used to determine whether the perceived dissimilarity reflects the Euclidean space model.

The mean R^2 and the corresponding 95% confidence interval are plotted in Fig. 6. The mean R^2 of all the responses was 74% with the confidence intervals for all the sections overlapping. This means that all of the five subspaces demonstrate a similar degree of fit to a Euclidean model of two-dimensional sound color perception regardless of the various choices of coordinates from the MFCC space.

Fig. 7 shows the regression coefficients [i.e., a and b from Eq. (13)] for each of the two variables from the regression analysis for all five sections. The mean regression coefficients were consistently higher for the lower one of the two MFCC variables, which means that lower Mel-cepstrum

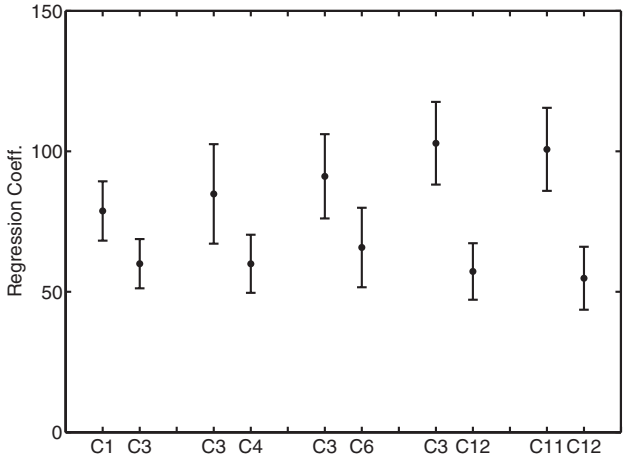


Fig. 7. Regression coefficients from regression analysis of the two-dimensional sound color experiment. The first two points on the left represent the regression coefficients for each dimension of the [C_1, C_3] subspace, followed by the regression coefficients for the subspaces of [C_3, C_4], [C_3, C_6], [C_3, C_{12}], and [C_{11}, C_{12}].

coefficients are perceptually more significant. Although the confidence intervals overlap for the lower-order MFCCs, and not for the higher-order MFCCs, this trend as regards the mean regression coefficients is consistent across all the MFCC subspace arrangements. This can be interpreted as indicating that the degree of contribution of the MFCCs is similar in the low- to mid-order MFCCs with a slightly decreasing trend, and for higher-order MFCCs, the degree of contribution drops more quickly and significantly.

4 DISCUSSION

4.1 Representing the Spectral Envelope Perception with MFCC

This section integrates the two experiments and discusses whether an MFCC can be a fair representation for spectral envelope perception. To summarize Experiment 1, it was shown that every orthogonal basis from the MFCC is linearly correlated to spectral envelope perception with an average degree of fit of 85%. This holds true for every single coefficient from the 12 dimensions in the MFCC vector, meaning that each of the coefficients is directly associated with spectral envelope perception. Experiment 2 tested the association between spectral envelope perception and two-dimensional MFCC subspace. The Euclidean distance in the MFCC explains the spectral envelope perception with an average degree of fit of 74%. Five different arrangements of two-dimensional subspaces were selected, and all the arrangements showed a similar degree of fit to the Euclidean distance model. An examination of the regression coefficients demonstrated that lower MFCC coefficients had a stronger effect in the perceived sound color space. These findings suggest that the MFCC can satisfy the desired characteristics of the spectral envelope perception model described in "Introduction".

The limitation of this experiment is that it only measured the responses to single-dimensional and two-dimensional MFCC subspaces. However, for further dimensionality, Beauchamp reported that the full dimensional MFCC can represent the timbre perception of musical instrument sounds with a comparable precision to the Mel-band or harmonics-based representations [43]. Other successful applications such as automatic speech recognition [20] or music information retrieval [21] suggest that the MFCC can efficiently retrieve timbre-related information such as vowels, consonants, and types of musical instruments. The recent work by Alluri and Toviainen reports that the polyphonic timbre of excerpts from musical works may not be necessarily well described using an MFCC[44]. However, because the scope of this experiment was the perception of musically organized mixtures of complex instrumental sounds, this finding does not deny the capability of the MFCC to represent the spectral envelope perception.

Previous works and applications have demonstrated that the MFCC is a useful description for timbre-related information, but did not show how each of the MFCC components contributes to the overall performance of the whole MFCC system. The experiments in this study showed that each of the coefficients linearly correlates to the spectral envelope perception and that there is a linear mapping between the perceived dissimilarity of the spectral envelope and the Euclidean distance in a two-dimensional MFCC subspace. These findings, along with Beauchamp's full-dimensional MFCC study, suggest that the MFCC can be a fair representation of spectral envelope perception, and that spectral envelope perception can be fully described in terms of the Euclidean space constituted by MFCCs.

4.2 Associating the Spectral Centroid and an MFCC

This section discusses the relationship between an MFCC and the spectral centroid in representing the spectral envelope perception. A spectral centroid has a clear, strong correlation with the perceived brightness of sound [45], which is an important factor in timbre perception [6].

First, to compare the spectral centroid with the MFCC, the linear regression analysis of Experiment 1 was conducted using the spectral centroid of stimuli as an independent variable. The results were almost identical and statistically equivalent to Fig. 3. To investigate this effect, the spectral centroid for each of the stimuli used in Experiment 1 was calculated, which is shown in Fig. 8. This illustrates that when a single dimension of the MFCC is manipulated, the resulting stimuli have a linear increase/decrease in the spectral centroid. The C_1 stimuli had lower centroids while C_1 was increasing from 0 to 1, and the C_2 stimuli had higher centroids while C_2 was increasing, but with a smaller coefficient (less slope), and so on. In summary, lower MFCC coefficients have a stronger correlation to the spectral centroid, and the correlation is *negative* for odd-numbered MFCC dimensions (the spectral centroid decreases while C_n increases, where n is an odd number), and *positive* for even-numbered MFCC dimensions (the spectral

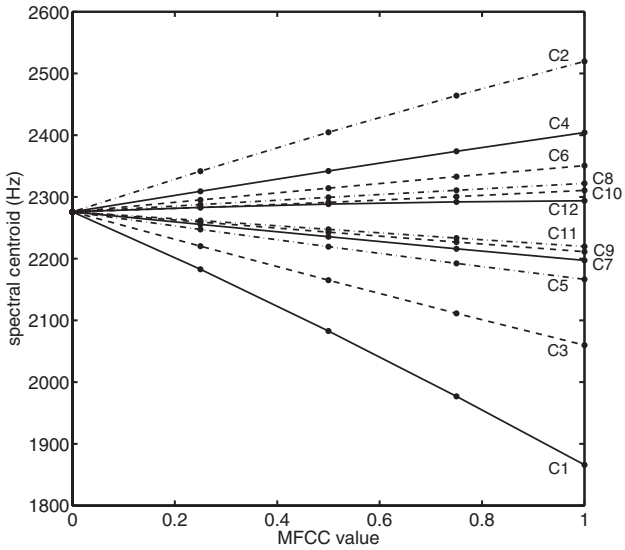


Fig. 8. Spectral centroid of the stimuli used for Experiment 1, when a single coefficient from the Mel-cepstrum was varied from 0 to 1 in five steps.

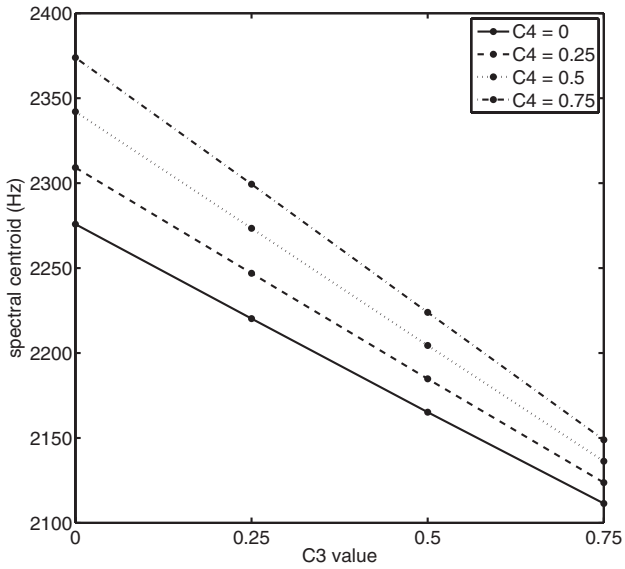


Fig. 9. Spectral centroid of the stimuli used for Experiment 2, Section 2, when two coefficients from the Mel-cepstrum, C_3 and C_4, were varied from 0 to 0.75 in four steps.

centroid increases while C_n increases, where n is an even number).

This is not a coincidence based on the trend in spectral envelopes generated for this experiment as shown in Fig. 2. The spectral envelopes generated by varying C_1 have a hump around the low-frequency range, which corresponds to the cosine wave at $\omega = 0$, and a dip around the Nyquist frequency, which corresponds to $\omega = \pi/2$. As C_1 increases, the magnitude of the hump becomes higher.

The concentrated energy around the low-frequency region corresponds to the fact that the spectral centroids are lower while the value of C_1 increases. Now, if the spectral envelopes are generated by varying C_2, there are two humps at the lowest frequency and the Nyquist frequency that correspond to $\omega = 0$ and $\omega = \pi$. Another hump at the Nyquist frequency makes the spectral centroid higher, whereas increasing the value of C_2 increases the spectral centroid. The same trends are conserved for odd- and even-numbered MFCC coefficients. With higher orders of MFCC, the basis function has its humps more sparsely distributed over the spectrum, which results in a weaker correlation between the MFCC and the spectral centroid (i.e., the slope of the line in Fig. 8 becomes more shallow as n increases).

Furthermore, the results from Experiment 2 show that the lower-order Mel-cepstrum coefficient is perceptually more important. As shown in Fig. 9, the linear relationship between the MFCC and spectral centroid is consistent in the stimuli set for Experiment 2. The low coefficient's strong association with the spectral centroid can explain this effect. Because of the correlation between the spectral centroid and MFCC in the stimuli for Experiment 2, the result of the regression analysis based on the spectral centroid was very similar to Fig. 6, except for Section 1. For Section 1, the R^2 of the spectral-centroid-based regression was 84%, scoring it 13% above the R^2 of the MFCC-based regression, without overlapping confidence intervals. This could be explained in terms of the coefficient choice of C_1 and C_3, which have a strong correlation with the spectral centroid in the same direction, and therefore are easily confused. For Sections 2–5, the R^2 of the MFCC-based regression was consistently higher by 2–5% than the R^2 of spectral-centroid-based regression, with overlapping confidence intervals.

The above-mentioned characteristics can be dependent on the specific MFCC implementation, and the pseudo-inversion of the MFCC used in this experiment. Depending on how the MFCC and its inversion are implemented, it could have different kinds of relationships to the spectral centroid. The relevance between the MFCC and spectral centroid present in this experiment may be generalized with further mathematical rationalization. If it is mathematically promised that higher Mel-cepstrum coefficients have a weaker correlation with the spectral centroid resulting in the reduced perceptual significance, it may explain the efficiency of the common practice, which uses only 12 or 13 lower coefficients from the MFCC for automatic speech recognition or music information retrieval.

However, there was a trend in the spectral centroids in the MFCC-based stimuli set for both experiments, and our results do not conflict with the previously reported characteristics of the spectral centroid in relation to the timbre perception. Both Experiments 1 and 2 suggest that an MFCC-based description holds a similar degree of linearity in predicting spectral envelope perception to a spectral-centroid-based description. Yet the spectral centroid is essentially a single-dimensional descriptor and does not describe the complex shapes of the spectral envelope itself. Two sounds with different spectral envelopes could have the same spectral-centroid value, but be represented with different Mel-cepstrum values. The multidimensional

Mel-cepstrum delivers more information about the spectral envelope than the spectral centroid.

5 CONCLUSION

On the basis of desirable properties for modeling spectral envelope perception (linearity, orthogonality, and multidimensionality), Mel-frequency cepstral coefficients (MFCCs) were chosen as a hypothetical metric for modeling spectral envelope perception. Quantitative data from two experiments illustrate the linear relationship between the subjective perception of vowel-like synthetic sounds and the MFCC.

The first experiment tested the linear mapping between spectral envelope perception and all 12 Mel-cepstrum coefficients. Each Mel-cepstrum coefficient showed a linear relationship to the subjective judgment at a statistically equivalent level to any other coefficient. On average, the MFCC explains 85% of spectral envelope perception when a single coefficient from the MFCC is varied in an isolated manner from all the other coefficients.

In the second experiment, two Mel-cepstrum coefficients were simultaneously varied to form a stimulus set in a two-dimensional MFCC subspace, and the relevant spectral envelope perception was tested. A total of five subspaces were tested, and all five exhibited a linear relationship between the perceived dissimilarity and the Euclidean distance of the MFCC at a statistically equivalent level. A subjective dissimilarity rating showed an average correlation of 74% with the Euclidean distance between the Mel-cepstrum coefficients of the tested stimulus pair. In addition, the observation of regression coefficients demonstrated that lower-order Mel-cepstrum coefficients influence spectral envelope perception more strongly.

The use of MFCCs to describe spectral envelope perception was further discussed. Such a representation can be useful not only in analyzing audio signals, but also in controlling the timbre in synthesized sounds. The correlation between the MFCC and the spectral centroid was also discussed, although such a correlation can be specific to our experimental conditions, and further mathematical investigation is needed.

These experiments examined the MFCC model at low dimensionality. Much work remains to be done in understanding how MFCC variation across the entire 12 dimensions might relate to human sound perception. An interesting approach is currently being employed by Horner and coworkers, who are taking their previous experimental data on timbre morphing of instrumental sounds [10, 11] and reanalyzing it using MFCC [26], [43]. Their approach using instrumental sounds will provide a good complement to the approach taken here.

6 ACKNOWLEDGMENT

We thank Malcolm Slaney for his contributions in establishing this research, and for his generous support in the preparation of this article. We also thank Jim Beauchamp, Andrew Horner, Michael Hall, and Tony Stockman for their helpful comments. This work was supported by France–Stanford Center for Interdisciplinary Studies, The Banff Centre, AES Educational Foundation, and JST-PRESTO.

7 REFERENCES

[1] H. Helmholtz, *On the Sensation of Tone (translation by Alexander John Ellis)*, pp. 64–65 (Dover Publications, Mineola, NY, Original German Edition in 1863, English translation in 1954).

[2] J. B. Allen, "How do humans process and recognize speech?," *IEEE Trans. Speech Audio Process.*, vol. 2, pp. 567–577 (1994 Oct.).

[3] G. E. Peterson and H. L. Barney, "Control methods used in a study of the vowels," *J Acoust Soc Am.*, vol. 24, no. 2, pp. 175–184 (1952).

[4] J. Grey, "Multidimensional perceptual scaling of musical timbres," *J. Acoust. Soc. Am.*, vol. 61, no. 5, pp. 1270–1277 (1977).

[5] D. L. Wessel, "Timbre space as a musical control structure," *Comput. Music J.*, vol. 3, no. 2, pp. 45–52 (1979).

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

[7] S. Lakatos, "A common perceptual space for harmonic and percussive timbres," *Percept. Psychophys.*, vol. 62, no. 7, pp. 1426–1439 (2000).

[8] J. W. Beauchamp, "Perceptually correlated parameters of musical instrument tones," *Arch.Acoust.*, vol. 36, no. 2, pp. 225–238 (2012).

[9] J. Blauert and U. Jekosch, "A layer model of sound quality," *J. Audio Eng. Soc.*, vol. 60, no. 1/2, pp. 4–12 (2012).

[10] 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.).

[11] A. B. Horner, J. W. Beauchamp, and R. H. Y. So, "Detection of time-varying harmonic amplitude alterations due to spectral interpolations between musical instrument tones," *J. Acoust. Soc. Am.*, vol. 125, no. 1, pp. 492–502 (2009).

[12] M. Hall and J. Beauchamp, "Clarifying spectral and temporal dimensions of musical instrument timbre," *Acoust. Can. J. Can. Acoust. Assoc.*, vol. 37, no. 1, pp. 3–22 (2009).

[13] S. Barrass, "A perceptual framework for the auditory display of scientific data," *ACM Trans. Appl. Percept.*, vol. 2, no. 4, pp. 389–402 (2005).

[14] R. Plomp, *Aspects of Tone Sensation: A Psychophysical Study*, ch. 6 (Timbre of Complex Tones), pp. 85–110 (Academic Press, New York, 1976).

[15] W. Slawson, *Sound Color*, pp. 3–21 (University of California Press, Berkeley, CA, 1985).

[16] H. F. Pollard and E. V. Jansson, "A tristimulus method for the specification of musical timbre," *Acustica*, vol. 51, pp. 162–171 (1982).

[17] J. S. Bridle and M. D. Brown, "An experimental automatic word-recognition system: Interim report," JSRU Report 1003, Joint Speech Research Unit, 1974.

[18] P. Mermelstein, "Distance measures for speech recognition, psychological and instrumental," in *Pattern Recognition and Artificial Intelligence* (C. H. Chen, ed.), pp. 374–388 (Academic Press, New York, 1976).

[19] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Speech Audio Process.*, vol. ASSP-28, pp. 357–366 (1980 Aug.).

[20] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, pp. 183–190 (Prentice Hall, Upper Saddle River, NJ, 1993).

[21] G. D. Poli and P. Prandoni, "Sonological models for timbre characterization.," *J. New Music Res.*, vol. 26, pp. 170–197 (1997).

[22] J.-J. Aucouturier, *Ten Experiments on the Modelling of Polyphonic Timbre*. Ph.D. thesis (University of Paris 6, Paris, France, 2006).

[23] S. Heise, M. Hlatky, and J. Loviscach, "Aurally and visually enhanced audio search with soundtorch," in *ACM CHI 2009 Extended Abstracts*, pp. 3241–3246 (2009 Apr.).

[24] N. Osaka, Y. Saito, S. Ishitsuka, and Y. Yoshioka, "An electronic timbre dictionary and 3d timbre display," in *Proc. 2009 Int. Computer Music Conference*, pp. 9–12 (2009).

[25] M. Hoffman and P. R. Cook, "Feature-based synthesis for sonification and psychoacoustic research," in *Proc. 12th Int. Conf. Auditory Display, London, UK.*, pp. 254–257 (2006).

[26] A. B. Horner, J. W. Beauchamp, and R. H. Y. So, "Evaluation of mel-band and mfcc-based error metrics for correspondence to discrimination of spectrally altered musical instrument sounds," *J. Audio Eng. Soc.*, vol. 59, no. 5, pp. 290–303 (2011).

[27] H. Terasawa, *A Hybrid Model for Timbre Perception: Quantitative Representations of Sound Color and Density*. Ph.D. thesis (Stanford University, Stanford, CA, Stanford, CA, 2009).

[28] H. Terasawa, M. Slaney, and J. Berger, "Perceptual distance in timbre space," in *Proc. ICAD 05 - Eleventh Meeting of the International Conference on Auditory Display*, pp. 61–68 (2005).

[29] H. Terasawa, M. Slaney, and J. Berger, "A timbre space for speech," in *Proc. Interspeech 2005–Eurospeech*, pp. 1729–1732, 2005.

[30] H. Terasawa, M. Slaney, and J. Berger, "The thirteen colors of timbre," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 323–326 (2005).

[31] E. Zwicker and H. Fastl, *Psychoacoustics – Facts and Models*, pp. 223–226 (Springer, Berlin, 1999).

[32] S. Shamma, "Speech processing in the auditory system," *J. Acoust. Soc. Am.*, vol. 78, no. 5, pp. 1612–1632, 1985.

[33] T. Irino and R. D. Patterson, "Segregating information about the size and shape of the vocal tract using a time-domain auditory model: The Stabilised Wavelet-Mellin Transform," *Speech Commun.*, vol. 36, pp. 181–203, 2002.

[34] A. Bregman, *Auditory Scene Analysis*, 2nd ed (MIT Press, Cambridge, MA, 2001).

[35] S. Barrass and G. Walker, "Using sonification," *Multimedia Syst.*, vol. 7, pp. 23–31 (1999).

[36] T. Hermann, G. Baier, U. Stephani, and H. Ritter, "Vocal sonification of pathologic EEG features," in *Proc. Int. Conf. Auditory Display (ICAD 2006)*, pp. 158–163 (2006).

[37] R. Cassidy, J. Berger, K. Lee, M. Maggioni, and R. R. Coifman, "Auditory display of hyperspectral colon tissue images using vocal synthesis models," in *Proc. Int. Conf. Auditory Display (ICAD 2004)*, pp. 1–8 (2004).

[38] J. Sundberg, "Vibrato and vowel identification," *Arch. Acoust.*, vol. 2, pp. 257–266 (1977).

[39] S. McAdams and X. Rodet, "The role of FM-induced AM in dynamic spectral profile analysis," in *Basic Issues in Hearing* (H. Duifhuis, J. Horst, and H. Wit, eds.), pp. 359–369 (Academic Press, London; San Diego, CA, 1988).

[40] M. Slaney, "Auditory toolbox version 2," Tech. Rep. 1998-010, Interval Research, 1998.

[41] W. Mendenhall and T. Sincich, *Statistics for Engineering and the Sciences*, pp. 531–698 (Prentice Hall, Upper Saddle River, NJ, 1995).

[42] J. Rogers, K. Howard, and J. Vessey, "Using significance tests to evaluate equivalence between two experimental groups," *Psychological Bulletin*, vol. 113, no. 3, pp. 553–565 (1993).

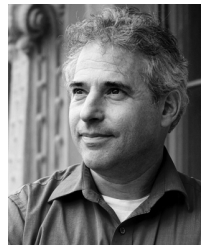
[43] J. W. Beauchamp, H. Terasawa, and A. B. Horner, "Predicting perceptual differences between musical sounds: A comparison of Mel-band and MFCC based metric results to previous harmonic-based results," in *Proc. Soc. Music Perception and Cognition 2009 Biennial Conference*, p. 82 (2009).

[44] V. Alluri and P. Toiviainen, "Exploring perceptual and acoustical correlates of polyphonic timbre," *Music Percept.*, vol. 27, no. 3, pp. 223–241 (2009).

[45] E. Schubert and J. Wolfe, "Does timbral brightness scale with frequency and spectral centroid?," *Acta Acust. United Acust.*, vol. 92, pp. 820–825 (2006).

THE AUTHORS

Hiroko Terasawa



Jonathan Berger

Shoji Makino

Hiroko Terasawa received B.E. and M.E. degrees in Electrical Engineering from the University of Electro-Communications, Japan, and M.A. and Ph.D. degrees in Music from Center for Computer Research in Music and Acoustics (CCRMA), Stanford University, the United States. She is the recipient of the Centennial TA Award from Stanford University (2006), the Artist in Residence at Cité Internationale des Arts (2007), the second place of the Best Student Paper Award in Musical Acoustics at the 156th ASA Meeting (2008), the John M. Eargle Memorial Award from AES Educational Foundation (2008), the Super Creator Award from ITPA Mitoh Program (2009), and the JST-PRESTO Research Grant (2011). Her research interests include timbre perception modeling and timbre-based data sonification. She is now a researcher at University of Tsukuba and JST PRESTO, and a lecturer on electronic music at Tokyo University of the Arts.

●

Jonathan Berger, The Denning Provostial Professor in Music at CCRMA, Stanford University, is a composer and researcher. He has composed orchestral music as well as chamber, vocal, and electro-acoustic and intermedia works. Berger was the 2010 Composer in Residence at the Spoleto USA Festival, which commissioned a chamber work for soprano Dawn Upshaw and piano quintet. He is currently working on a chamber opera commissioned by the Andrew Mellon Foundation. Other major commissions and fellowships include the National Endowment for the Arts (a work for string quartet, voice, and computer in 1984, soloist collaborations for piano, 1994, and for cello, 1996, and a composers fellowship for a piano concerto in 1997); The Rockefeller Foundation (work for computer-tracked dancer, live electronics, and chamber ensemble); and The Morse and Mellon Foundations (symphonic and chamber music). Berger received prizes and commissions from the Bourges Festival, WDR, the Banff Centre for the Arts, Chamber Music America, Chamber Music Denver, the Hudson Valley Chamber Circle, The Connecticut Commission on the Arts, The Jerusalem Foundation, and others. Bergers recording of chamber music for strings, Miracles and Mud, was released by Naxos on their American Masters series in 2008. His violin concerto, Jiyeh, is soon to be released by Harmonia Mundis Eloquentia label. Bergers research in music perception and cognition focuses on the formulation and processing of musical expectations, and the use of music and sound to represent complex information for diagnostic and analytical purposes. He has authored and co-authored over seventy publications in music theory, computer music, sonification, audio signal processing, and music cognition. Before joining the faculty at Stanford he taught at Yale where he was the founding director of Yale University's Center for Studies in Music Technology. Berger was the founding co-director of the Stanford Institute for Creativity and the Arts (SICA) and, codirected the Universitys Arts Initiative.

●

Shoji Makino received B.E., M.E., and Ph.D. degrees from Tohoku University, Japan, in 1979, 1981, and 1993, respectively. He joined NTT in 1981. He is now a Professor at University of Tsukuba. His research interests include adaptive filtering technologies, the realization of acoustic echo cancellation, blind source separation of convolutive mixtures of speech, and acoustic signal processing for speech and audio applications. He received the ICA Unsupervised Learning Pioneer Award in 2006, the IEEE MLSP Competition Award in 2007, the TELECOM System Technology Award in 2004, the Achievement Award of the Institute of Electronics, Information, and Communication Engineers (IEICE) in 1997, and the Outstanding Technological Development Award of the Acoustical Society of Japan (ASJ) in 1995, the Paper Award of the IEICE in 2005 and 2002, the Paper Award of the ASJ in 2005 and 2002. He is the author or co-author of more than 200 articles in journals and conference proceedings and is responsible for more than 150 patents. He was a Keynote Speaker at ICA2007, a Tutorial speaker at ICASSP2007, and a Tutorial speaker at INTERSPEECH2011. He has served on IEEE SPS Awards Board (2006–2008) and IEEE SPS Conference Board (2002–2004). He is a member of the James L. Flanagan Speech and Audio Processing Award Committee. He was an Associate Editor of the IEEE Transactions on Speech and Audio Processing (2002–2005) and is an Associate Editor of the EURASIP Journal on Advances in Signal Processing. He is a member of SPS Audio and Electroacoustics Technical Committee and the Chair of the Blind Signal Processing Technical Committee of the IEEE Circuits and Systems Society. He was the Vice President of the Engineering Sciences Society of the IEICE (2007–2008), and the Chair of the Engineering Acoustics Technical Committee of the IEICE (2006–2008). He is a member of the International IWAENC Standing committee and a member of the International ICA Steering Committee. He was the General Chair of WASPAA2007, the General Chair of IWAENC2003, the Organizing Chair of ICA2003, and is the designated Plenary Chair of ICASSP2012. Dr. Makino is an IEEE SPS Distinguished Lecturer (2009–2010), an IEEE Fellow, an IEICE Fellow, a council member of the ASJ, and a member of EURASIP.