Determining the Euclidean Distance Between Two Steady State Sounds

Hiroko Terasawa
CCRMA, Department of Music,
Stanford University.
Stanford, CA, USA
hiroko@ccrma.stanford.edu

Malcolm Slaney
Yahoo! Research / CCRMA
Sunnyvale, CA, USA

Jonathan Berger
CCRMA, Department of Music,
Stanford University
Stanford, CA, USA

ABSTRACT

We describe a perceptual space for timbre, define an objective metric that takes into account perceptual orthogonality and measure the quality of timbre interpolation. We discuss two timbre representations and using these two representations, measure perceived relationships between pairs of sounds on a equivalent range of timbre variety. We determine that a timbre space based on Mel-frequency cepstral coefficients (MFCC) is a good model for a perceptual timbre space.

Keywords

Timbre, perception, MFCC.

INTRODUCTION

Timbre is a key distinguishing feature in the identification, characterization and classification of acoustic signals be they musical, speech or environmental. Paradoxically, this characterization is independent of pitch, loudness and spatial orientation yet incorporative of these attributes. This paradox lies in the ability to comparatively regard commonalities of spectra autonomous of pitch, yet the inherent role that the integration of all other characteristic aspects of sound play in the time-variant factors such as modulation, beating, phase, attack/decay transient, and so on.

Timbre plays a particularly salient role in music perception often assuming a structural function by demarcating segmentation boundaries or delineating patterns. However there is no consistent and principled means of predicting the quality of timbre. Descriptions of timbre are often adjectival and self-referential (for example, describing a bowed sul-tasto violin as 'flautando', an oboe sound as 'nasal', etc.). Existing perceptual models of timbre are not quantifiable.

Our goal is to find a computationally viable model or representation of timbre that is isomorphic with human perception. In this paper we describe a quantitative causality between the percept of timbre and spectral shape and develop a parsimonious model for timbre space based upon this causal relationship.

Timbre descriptions

The descriptive and impressionistic nature of existing timbre descriptions (Hajda, Kendall, Carterette, & Harshberger, 1997), (Krumhansl, 1989) pose an obvious limitation in a music theoretic approach to timbre. Lacking a Euclidean measurement of distance between two timbres, timbre can, at best, be described as a specific point within a multidimensional continuum, with that point defined by a combination of subjective perceptual and physical dimensions. In this approach paired adjectival antonyms such as "bright—dull" or "sharp—not sharp" establish perceptual dimensions (Kendall & Carterette, 1993).

An alternative representation of musical timbre is the tristimulus model (Pollard & Jansson, 1982). This two dimensional space describes harmonic sounds in terms of generalized weightings of harmonics in which three coefficients calculated according to Steven's law, respectively represent the perceived strength of the fundamental, mid-frequency partials and high-frequency partials. The principle benefit of the model is its simplicity resulting in a direct mapping of the coefficients to an adjectival description of percept. The approach, however, is limited by the inability to represent inharmonicity and a rather arbitrary delineation of the three frequency components.

Timbre Distance

Most quantitative approaches to timbre perception describe the distance between two sounds. Popular approaches are based on speech perception, speech recognition, and the perception of musical sounds.

One of the earliest approaches to understand sound perception was undertaken by Harvey Fletcher and his colleagues at Bell Labs at the start of the 20th century. This work (Fletcher, 1934) measured subjects' ability to correctly recognize nonsense words in the presence of filtering and noise. It suggests that wide bands of frequencies provide independent information about the speech sounds that are heard. However this work only applies to speech, only

as part of a recognition task and lacks generalization to describe the underlying acoustic space of any sound.

Speech recognition systems have had great success modeling the acoustic world using Gaussian mixture models (GMMs) to build a probabilistic model of the acoustic spectra that are likely to be found in each type of phoneme. By trial and error, and for statistical reasons, much of the speech-recognition research has settled on Mel-frequency cepstral coefficients (MFCC) as the underlying model of speech sounds (Davis & Mermelstein, 1980). While MFCC coefficients are loosely based on a simple model of auditory perception, their primary benefit is that the different coefficients are statistically independent so GMMs with diagonal covariance can be used and an MFCC front-end produces a working speech recognizer. But MFCC's success in speech recognition is not the same as proving that MFCCs are a good model of perception.

An entirely new and quantitative approach to measuring timbre perception started with the work of Wessel (Wessel, 1979), Grey (1975 and 1976) and the subsequent research (McAdams, Winsberg, Donnadieu, De Soete, & Krimphoff, 1995) (Lakatos, 2000). It directly measured the distance between two musical sounds. By using multidimensional scaling (MDS) the sounds can be represented in a low-dimensional surface (plane or 3-D cube) in such a way that the projected locations fit the observed perceptual data as closely as possible. There are two shortcomings with this approach. Most importantly, the axes produced by the MDS algorithm are not labeled. It is up to the imagination of the researcher to look at the position of the sounds, and generate an explanation of what each axis means (for example, sounds are duller/brighter along this direction.) Secondarily, while this approach is descriptive of the tested sounds, it does not help us estimate the percept of sounds which are not tested in the experiment, nor find a sound that has a needed distance from other sounds. For this we need to find and describe a timbre space that matches human perception.

Desired Model for Timbre Space

Perceptual maps exist in the auditory domain for pitch and loudness, as well and for spatial geometry color in the visual domain. In each case, a relatively simple model connects physical attributes (mel for pitch, sones for loudness, and the three cones of the visual system for color) with perceptual judgments. However, such a model does not yet exist for a perceptual timbre space.

Bregman (2001) proposes that a timbre space have two properties; Psychological simplicity (that is as an independent factor in scene analysis), and a straightforward physical definition. “What we need to do,” he writes, “is to develop descriptive terms for timbre, find ways to measure them, and do careful studies of how they affect perceptual grouping.”

A timbre space should be both simple to understand and successfully predict discriminant human perception. Our goal is to create a perceptual space that articulates the physical attributes of a particular timbre with human perception of that timbre. A good model of timbre perception describes a space of sounds with a number of simple properties and explanations.

This paper describes a three-stage approach for establishing a perceptual timbre map. The three steps include (1) postulating a metric for the quality of a perceptual space, (2) describing a mathematical representation of timbre, and (3) measuring the match between representation and perception. The sound representation that provides the simplest and most parsimonious description of timbre perception is the best model for timbre space.

Framework of the Experiment

Our test of perceptual parsimony considers linearity and orthogonality. Linearity suggests that the representation can accurately generate sounds that are perceived as interpolated midway between the original sounds. Orthogonality dictates that changes in one parameter not affect the perception of another parameter. We measure both of these properties of a perceptual space by testing whether the perceptual distance measurements satisfy the Euclidean rule for distance for a range of representation parameters.

When considering the existing timbre studies, one issue is that the time-variant (static) and time-invariant (dynamic) sounds are tested together. Although MDS studies revealed that the difference in spectral centroid contributes to timbral perception, we do not yet know the quantitative causality between the spectral shape of a static sound, and the percept of the timbre. Therefore, we believe, although it may seem conservative, we should start from articulating this quantifiable relationship, keeping the time-variant factors aside. These dynamic factors, which govern a significant part in timbral perception, are yet indescribable without having a perceptually orthogonal spectral shape representation.

In this paper we describe and compare two similar signal-processing representations of a sound in terms of their efficacy in meeting linearity and orthogonality. We know that timbre is a multidimensional quantity and an important metric in this work is that the representation's axis be perceptually orthogonal. This means that changes in one parameter do not affect perception of the other axis.

Preliminary experiments (Terasawa, Slaney & Berger, 2005a, b, c) have determined the MFCC to be a good representation for human perception of timbre for static sounds. These studies compared the MFCC representation to an alternative representation we named the linear frequency coefficients (LFC) representation. The LFC, detailed below shares the statistical properties of the MFCC representation but does not incorporate perceptual weightings. The LFC was used as a strawman in comparison with the MFCC. Both representations were determined to be better representations of timbre perception

tations of timbre perception than the tristimulus model (Terasawa et al., 2005c).

In the previous experiment the stimuli samples used for the LFC and MFCC representations potentially covered unequal areas of perceptual space. The experiment described here aims to deal with this potential discrepancy. Since Taylor's theorem suggests that the accuracy of a linear model is inversely proportional to the size of the neighbor-hood represented, we force the ranges of timbres generated by the LFC to be smaller than that of the MFCC stimuli and, as previously done, compare the representations in terms of their efficacy in representing human perception.

REPRESENTATIONS OF TIMBRE

Parameterization of Spectral Shapes

There are a wide variety of audio representations with differing degrees of abstraction. While a spectrum forms a complete representation of the sound, its arbitrary complexity makes a direct mapping to human perception difficult.

MFCC is well known as a front-end for speech-recognition systems (Davis & Mermelstein, 1980). It uses a filterbank based on the human auditory system: spacing filters in frequency based on the Mel-frequency scale to reshape and resample the frequency axis. A logarithm of each channel models loudness compression. Then a low-dimensional representation is computed using the discrete-cosine transform (DCT) (Blinn 1993). The DCT not only removes high-frequency ripples in the spectrum, but also serves to de-correlate the coefficients. However, this statistical property is not the same as perceptual orthogonality. Generally, based on speech-recognition engineering, a 13-D vector is used to describe speech sounds as a function of time.

LFC is a strawman representation we designed to be similar in representational ability to MFCC. We start with a linear-frequency scale and a linear amplitude scale. A 13-dimensional DCT of the normal amplitude spectrum reduces the dimensionality of the spectral space and smoothes the spectrum. Both MFCC and LFC use a DCT to reduce the dimensionality and de-correlate the coefficients; their difference lies in the frequency and amplitude warping.

In both representations, a static sound is described by a 13-D vector that represents a smoothed version of the original spectrum. The coefficients are labeled as C and C', for MFCC and LFC respectively. The first coefficient from the vector, C_0 or C'_0, represents the average power in the signal (constant in the experiments in this paper), and higher-order coefficients represent spectral shapes with more ripples in the auditory frequency domain. In a later section we show how to convert these 13-D representations into their equivalent spectra, and then back into sound.

Synthesis Method

In this study, we choose a 13-D vector and then synthesize sounds from these coefficients using the inverse transforms of LFC and MFCC. In both representations much information is lost, or equivalently, many different sounds will lead to equivalent coefficients. At each step in the transformation we choose the simplest spectrum.

We reconstruct the smooth spectrum by inverting the LFC and MFCC representations. For LFC, the reconstructed spectrum $\tilde{S}(f)$ is the IDCT of LFC vector C'_i. For MFCC, we first compute the IDCT of the MFCC vector $\tilde{L}_i = IDCT(C_i)$. Then raising ten to that power, $\tilde{F}_i = 10^{\tilde{L}_i}$ is the reconstructed filterbank output for channel i. We then assume that $\tilde{F}_i$ represents the value at the center frequencies of each channel, and render the reconstructed spectrum $\tilde{S}(f)$ by linearly interpolating values between the center frequencies.

Prepared MFCC Stimuli

As it is difficult to fully explore a 13-D space, we first chose discrete pairs of coefficients from 2-D MFCC spaces, and measured our subject's perceptual judgments in these 2-D spaces. Arbitrary pairs were studied to give insight into how the representations behaved. The four pairs studied are $[C_3, C_6]$, $[C_4, C_6]$, $[C_3, C_4]$, and $[C_{11}, C_{12}]$.

When forming the two dimensional subspaces, two of the 13 coefficients are chosen as variables and set to non-zero values, while the others kept constant. For example, the $[C_m, C_n]$ space has the 13-D parameter vector of

$$C = [1,0,\cdots,0,C_m,0,\cdots,0,C_n,0,\cdots,0]. \qquad (1)$$

C_m and C_n are quantized and take one of the following four values, $C_m = [0, M/3, 2M/3, M]$ where M is the maximum value. C_n is varied over four discrete values in the same way as C_m, with the maximum value N. The parameter vector C is interpreted as MFCC for resynthesis. Since we have four levels for each of dimensions C_m and C_n, we form a four by four grid in the 2-D space, resulting in a set of 16 stimuli samples with varying spectral shapes.

Designing LFC Stimuli

It is difficult to directly compare two different types of perceptual spaces such as MFCC and LFC. In general, the sets of sounds will be different and it is hard to ensure that one set of sounds covers no more of the perceptual space than the other. To make this comparison, we generate sounds using the MFCC vectors, transform them into sounds using the inverse algorithm described in Section "Synthesis Method" and then reanalyze the resulting sound using LFC.

Figure 1 (top) shows the case LFC-transformed MFCC space is bigger than the LFC parameter space. In this case, according to the Taylor's theorem, it is expected that LFC fits better to a linear model. If MFCC fits better to a linear model even in this case, it reinforces the probability of MFCC being a better representation of timbre.

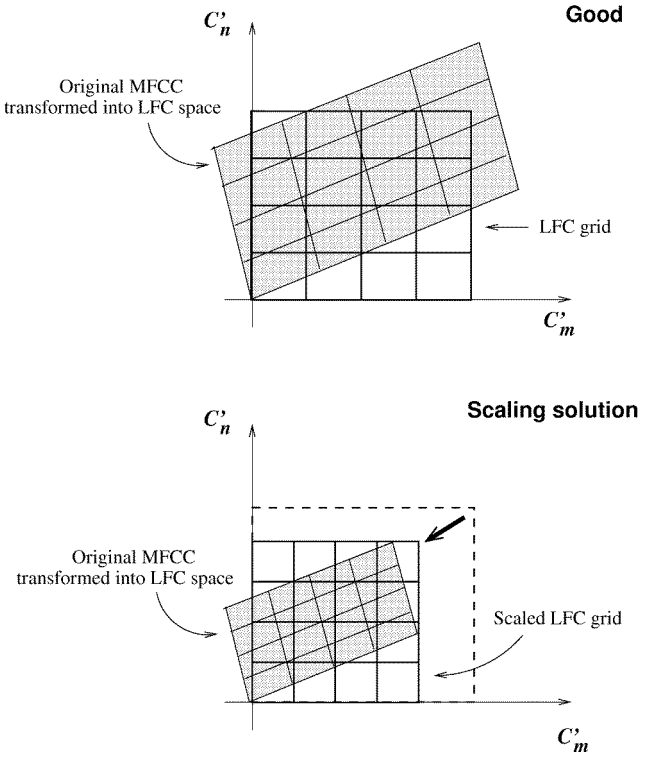

Figure 1. Corresponding MFCC space and LFC space. Top: comparison of two spaces is considered legitimate when corresponding regions between the two representations overlap (MFCC space is transformed into LFC space.) Bottom: When the transformed MFCC space has much smaller region than LFC parameter space (dotted rectangular), LFC parameter space is rescaled to match with the size of LFC-transformed MFCC space (solid grid.)

In our previous work, LFC and MFCC sounds covered very different regions. In that case, it is arguable that the good performance of MFCC timbre representation might have come from the fact it covered less timbral space than LFC. Therefore, this time, we want to do a very conservative test, which forces MFCC being bigger size than LFC by scaling it. This idea is shown as scaling solution in Figure 1 (bottom).

In this work we transform one set of sounds, created on a grid in MFCC space, into the LFC space. These 16 MFCC sounds will not form a regular grid on a two-dimensional plane in LFC space—they form a 2-D manifold. For this reason, we use Principal Component Analysis to find the largest two-dimensional LFC space that describes the sounds, and ignore the other dimensions. We then scale the LFC coefficients so that they are no bigger than the transformed MFCC dimensions, as shown in Figure 1 (bottom). This is a very conservative test—we have thrown out many dimensions of variations, so that we can guarantee that the LFC space is no bigger than MFCC.

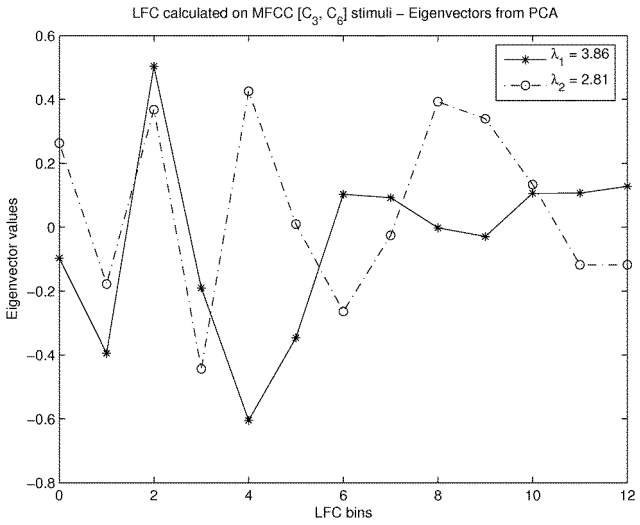

Figure 2. Eigenvectors of PCA—LFC-transformed MFCC [C_3, C_6] stimuli. The first two eigenvectors are plotted with their eigenvalues in the legend. It is visible that LFC C'_2 and C'_4 deliver most of the energy. These two dimensions are chosen to form a corresponding 2-D LFC space.

For a fairest comparison, we want to find a 2-D LFC space that is smaller, in a perceptual space, than the corresponding MFCC space. We do this in three steps. First we represent the test MFCC sounds with the LFC algorithm. Second, we find the two LFC dimensions that have the greatest variation. Third, we select and scale these two LFC dimensions so that the maximum extent is equivalent to the maximum extent of the LFC-transformed MFCC sounds.

The MFCC stimuli sounds are analyzed with the LFC algorithm, providing LFC vector C''. After analyzing all the 16 MFCC stimuli samples, we operate a principal component analysis on 16 LFC vectors C''.

The procedures of a principal component analysis are as follows (Duda, Gart, & Stork, 2001). The 13-dimensional mean vector and the 13 by 13 covariance matrix are computed for the full data set of 16 vectors of length 13. The eigenvectors and the eigenvalues are computed, and then sorted according to decreasing eigenvalues. Call these sorted eigenvectors $\mathbf{e_1}$ with eigenvalue λ_1, $\mathbf{e_2}$ with eigenvalue λ_2, and so on. Our MFCC stimuli are resolved into two-dimensional LFC subspaces, having two large eigenvalues.

We observe $\mathbf{e_1}$ in order to determine which coefficients of the LFC vector carry most of the energy, and choose two largest coefficients C''_m and C''_n from $\mathbf{e_1}$ in order to form a two-dimensional LFC space. Once we determine the dimensions, we go back to the C'' sample vectors and observe the coefficient with the largest deviation from zero out of 16 samples, and define the parameter range M' and N' as follows.

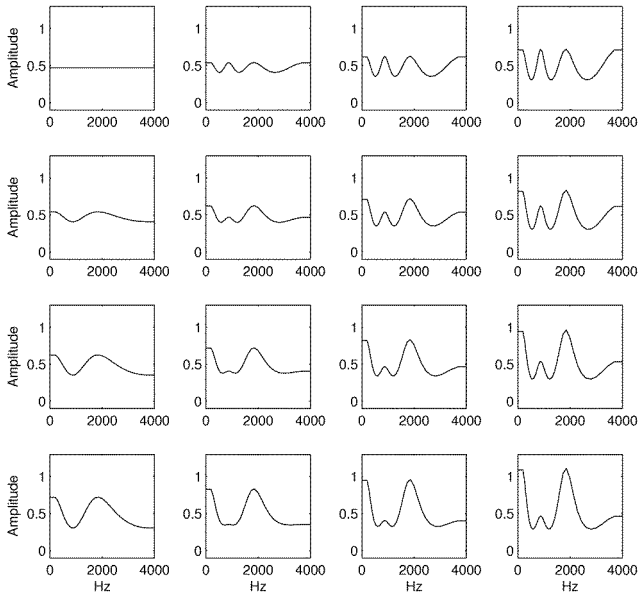

Figure 3. An array of spectra generated for a 2-D range of MFCC coefficients. The column show C_3 ranging from 0 to 0.75, the rows show C_6 ranging from 0 to 0.75.

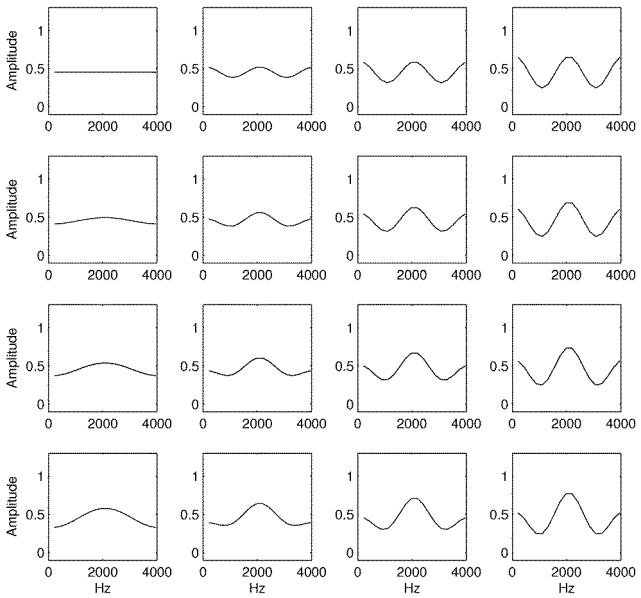

Figure 4. An array of spectra generated for a 2-D range of LFC coefficients. The column shows C'_2 ranging from 0 to -0.20, the rows show C'_4 ranging from 0 to 0.32.

$$M' = \arg\max_{\mathbf{C''_m}} \left(\left| \mathbf{C''_m} \right| \right) \quad (2)$$

$$N' = \arg\max_{\mathbf{C''_n}} \left(\left| \mathbf{C''_n} \right| \right) \quad (3)$$

In the above equations, $\mathbf{C''_m}$ and $\mathbf{C''_n}$ consist 16 elements of C''_m and C''_n from 16 sample vectors of C''. In order to form a new four by four grid, M' and N' become the maximum values of new parameter space for the LFC stimuli in the $[C'_m, C'_n]$ space. The parameter vector C' for LFC stimuli is defined in the same way as in Eq. (1), while C'_m and C'_n are varied over four discrete levels and the others are kept constant.

After designing this four by four parameter grid in LFC space, the parameter vector C' is interpreted as LFC for resynthesis, resulting in comparable 16 LFC stimuli sounds. Table 1 shows the tested pairs of MFCC and relevant LFC stimuli, and the maximum values in coefficients.

Table 1. Corresponding MFCC and LFC spaces for our test. The LFC spaces are designed to be no bigger than the corresponding MFCC space.

MFCC		LFC	
$[C_m, C_n]$	$[M, N]$	$[C'_m, C'_n]$	$[M', N']$
$[C_3, C_6]$	[0.75, 0.75]	$[C'_2, C'_4]$	[-0.20, 0.32]
$[C_4, C_6]$	[0.75, 0.75]	$[C'_3, C'_4]$	[-0.29, 0.17]
$[C_3, C_4]$	[0.75, 0.75]	$[C'_2, C'_3]$	[-0.20, -0.21]
$[C_{11}, C_{12}]$	[0.75, 0.75]	$[C'_5, C'_6]$	[-0.13, -0.12]

Representation Comparison

Any point in LFC or MFCC space is a sound. Figure 3 shows an array of spectra as we vary the C_3 and C_6 components of the vector, keeping all other coefficients but the C_0 component equal to zero. With both C_3 and C_6 coefficients set to zero, and $C_0 = 1$, the spectrum is flat. As the value of C_3 increases, going down the columns, there is a growing bump in the spectrum at DC and in the mid-frequencies. As the value of C_6 increases, going across rows, three bumps increase in size. Figure 4 shows an array of spectra for the corresponding LFC stimuli set that we test this time.

Additive Synthesis

The voice-like stimuli used in this study are synthesized from the spectrum derived by MFCC and LFC inversion using a source-filter model of speech. The source is an impulse train with the desired pitch. The filtering was implemented using additive synthesis. The amplitude of each harmonic component is scaled based on the desired spectral shape. The pitch, or fundamental frequency, f_0, is 220 Hz, the frequency of the vibrato v_0 is 6 Hz, and the amplitude of the modulation V is 6%. The synthesized sound is

$$s = \sum_n \tilde{S}\left(n \cdot f_0\right) \cdot \sin\left(2\pi n f_0 t + V(1 - \cos(2\pi n v_0 t))\right) \quad (4)$$

using the reconstructed spectral shape $\tilde{S}(f)$, with the harmonics number n.

EXPERIMENT

We measured the distance for several sets of timbre parameters by asking subjects for their subjective evaluation of the difference between two sounds in the prospective representation.

A stimulus consisted of two sounds, where the first is a reference sound and the second is a trial sound, with no pause between the paired sounds. The reference sound was kept identical through the entire experiment. It has a flat spectrum; all the 13 coefficients are zero except C_0 (i.e. $[C_m, C_n] = [0,0]$.) The second element of each pair, the trial sound, was varied in each presentation pair.

For each of the ten sets of sounds we played five examples to help the subjects understand the types and range of sounds that appear on the main experiment. In the main experiment, a distance measurement is recorded after playing a subject a pair of sounds. The subject was asked to rate the degree of similarity between pair elements on a scale of one to ten, where one is identical and ten is very different. The 16 stimuli in a set were presented to the subjects in a random order.

Twelve students with ages between 20—35 years old participated in the experiment. The stimuli were presented to the subject using a headset in a quiet office environment.

ANALYSIS METHOD

There are two steps in the analysis procedures. In the first step, we fit the individual distance judgments to a simple Euclidean model. We compute the residual from the model to evaluate the performance of the representations (LFC and MFCC) on each subject. In the second step, we computed the mean of the residuals and its standard error for each of ten sets in order to evaluate the representation.

Individual Euclidean Model Fitting

For a two-dimensional test as performed, the Euclidean model predicts the perceptual distance, *d*, which subjects reported in the experiment

$$d^2 = ax^2 + by^2 \tag{5}$$

where *x* is one of the 13 coefficients (e.g. C_3) and *y* is another coefficient (e.g. C_6). Note that this is a linear equation in the known quantities d^2, x^2 and y^2. Multidimensional linear regression is used in order to test the fit of perceptual data to a Euclidean model. The estimation of the regression model is done by the least squares method, using the left inverse (pseudo-inverse) of the matrix, which guarantees the minimum-error linear estimate. The residual of the linear estimation is:

$$d_{res} = \frac{1}{16}\sum_{x,y}\left|d - \hat{d}\right| \tag{6}$$

where $\hat{d}$ is the estimated distance by the linear regression model.

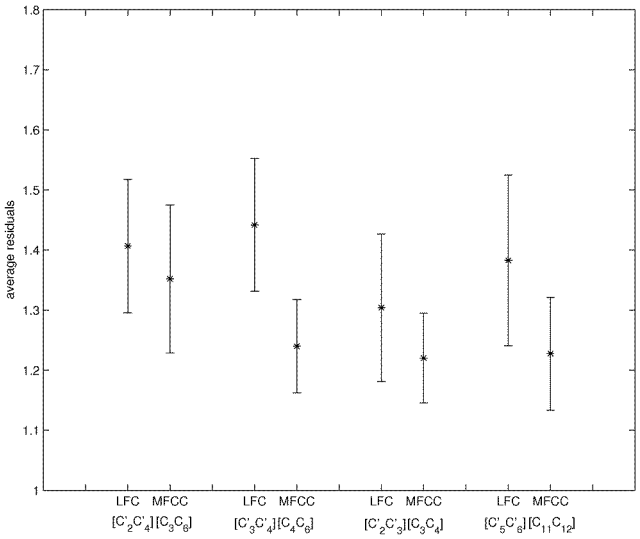

Figure 5. Model residuals and standard errors comparing MFCC and LFC for four sets of corresponding subspaces

Integrating the Individual Timbre Spaces

Given the model residuals for individual subjects, the mean of the residuals is calculated for each representation

$$\bar{d}_{res} = \frac{1}{N}\sum_{i=1}^{N} d_{res,i} \tag{7}$$

where *N* is the number of subjects. The standard error σ_{Mean} is calculated as follows:

$$\sigma_{Mean} = \frac{\sqrt{\sum_{i=1}^{N}\left|d_{res,i} - \bar{d}_{res}\right|^2}}{N} \tag{8}$$

By comparing the mean of the residuals and the standard error of each representation, we decide which representation is a better model of human perception.

RESULTS

Figure 5 compares LFC and MFCC in terms of each representation's ability to model a human's perception of timbre space. Each adjacent LFC and MFCC subspaces, e.g. [C'_2, C'_4] and [C_3, C_6], [C'_3, C'_4] and [C_4, C_6], and so on, are the corresponding sets of sounds with relevant spectral changes. On average, either timbre space predicts the perceptual judgment with a mean error of 1.32 point on a 10-point scale. In all cases, the MFCC representation performs as a better model for timbre space perception than the LFC representation, although the difference between the first pair of subspaces [C'_2, C'_4] and [C_3, C_6] is smaller than the other pairs.

In this experiment, we designed LFC parameter space so that LFC perceptual space would have similar or more linearity than MFCC, as described in Section "Designing LFC Stimuli." The timbral spaces covered by LFC stimuli are strictly constrained to be smaller than that of MFCC stimuli. As a result, the spectral deviations for the LFC

stimuli are smaller than MFCC parameter settings, providing an advantage to LFC stimuli. The LFC model covers smaller spectral region, and is more likely to behave linearly according to Taylor's theorem. Yet we observe that MFCC performs better than LFC with consistency and robustness, which suggests that MFCC is the better representation for human timbre perception.

CONCLUSIONS

In this paper we have articulated a set of criteria for evaluating a timbre space, described two representations of timbre, measured subject's perceptual distance judgments, and found that a model for timbre based on the MFCC representation accounts for 66% of the perceptual variance.

This result is interesting because we have shown objective criteria that describe the quality of a timbre space, and established that MFCC parameters are a good perceptual representation for static sounds. Previous work has demonstrated that MFCC (and other DCT-based models) produce representations that are statistically independent. This work suggests that the auditory system is organized around these statistical independences and that MFCC is a perceptually orthogonal space. The procedure described in this paper does not give a closed-form solution to the timbre-space problem. All we can do is test a representation and see if it is parsimonious with perceptual judgments. This paper is the first step towards a complete model of timbre perception.

In this work, we constrained LFC stimuli to have smaller deviation than MFCC, in order to insure the tested stimuli stay in a corresponding group of timbres. The parameter for the LFC was carefully constrained using a statistical approach so that LFC perceptual space is similarly, or even more likely, to be linear when compared to MFCC space. The experiment, however, proved that MFCC is still a better representation, which is orthogonal to our perception, even in this disadvantageous experiment condition for MFCC.

Most importantly, the timbre representations we tested here are static; musical sounds are not. Many timbre models find that onset time, for example, is an important component of timbre perception. But the criteria (linearity and orthogonality) we described here are important as we add features to the timbre space.

Although we have not begun to understand the contextual effects on timbre perception (Dennett, 1988) by addressing the underlying representational issues we hope that this will enable future research.

ACKNOWLEDGMENTS

The authors wish to thank Stephen McAdams, Daniel Levitin and Albert Bregman for the critical discussions that helped shape this work.

REFERENCES

Blinn, J. F. (1993) Jim Blinn's Corner: What's the Deal with the DCT? *IEEE Computer Graphics & Applications, July 1993,* 78–83.

Bregman, A. (2001) *Auditory Scene Analysis, second ed*. Cambridge: MIT press.

Davis, S. B. and Mermelstein, P. (1980) Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing, vol ASSP–28, No.4,* 357–366.

Dennett, D. C., (1988) Quining Qualia. In A. Marcel, and E. Bisiach (Eds.), *Consciousness in Modern Science* (pp. 42–77). New York: Oxford University Press.

Duda, R. O., Gart, P. E. and Stork, D. G. (2001) *Pattern Classification, second ed.* New York: Wiley-Interscience, 114–117.

Fletcher, H. (1934) Loudness, pitch and the timbre of musical tones and their relation to the intensity, the frequency and the overtone structure. *Journal of the Acoustical Society of America 6*, 59–69.

Grey, J. M. (1975) *An exploration of musical timbre*. Doctoral dissertation, Stanford University, Stanford.

Grey, J. M. (1976) Multidimensional Scaling of Musical Timbres. *Journal of the Acoustical Society of America, 61(5)*, 1270–1277.

Hajda, J. M., Kendall, R. A., Carterette, E. C. and Harshberger, M. L. (1997) Methodological issues in timbre research. In I. Deliege & J. Sloboda (eds.), *The Perception and Cognition of Music* (pp. 253–306), Hove, East Sussex : Psychology Press.

Kendall, R. A. and Carterette, E. C. (1993) Verbal attributes of simultaneous wind instrument timbres: I. von Bismarck's adjectives. *Music Perception, 10(4),* 445-468.

Krumhansl, C. L., (1989) Why is musical timbre so hard to understand? In S. Nielzen, and O. Olsson (Eds.), *Structure and perception of electroacoustic sound and music* (pp. 43–54). New York: Excerpta Medica.

Lakatos, S. (2000) A common perceptual space for harmonic and percussive timbres. *Perception & Psychophysics, 62 (7),* 1426–1439.

McAdams, S., Winsberg, W., Donnadieu, S., De Soete, G., and Krimphoff. J. (1995) Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes. *Psychological Research*, *58,* 177–192.

Pollard, H. F. and Jansson. E. V. (1982) A Tristimulus Method for the Specification of Musical Timbre *Acustica*, *51,* 162–171,

Terasawa, H., Slaney, M. and Berger, J. (2005a) Perceptual Distance in Timbre Space. *Proceedings of ICAD 05 - Eleventh Meeting of the International Conference on Auditory Display, Limerick, Ireland.*

Terasawa, H., Slaney, M., and Berger, J. (2005b) A Timbre Space for Speech. *Proceedings of Interspeech 2005 – Eurospeech, Lisbon, Portugal,* 1729-1732.

Terasawa, H., Slaney, M., and Berger, J. (2005c) The Thirteen Colors of Timbre. *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, Mohonk, NY.*

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