PERCEPTUAL DISTANCE IN TIMBRE SPACE

Hiroko Terasawa[†], *Malcolm Slaney*[†‡], *Jonathan Berger*[†]

Center for Computer Research
in Music and Acoustics (CCRMA)[†]
Department of Music, Stanford University
Stanford, California
{hiroko,brg}@ccrma.stanford.edu

IBM Almaden Research Center[‡]
San Jose, California
malcolm@ieee.org

ABSTRACT

This paper describes a perceptual space for timbre, defines an objective metric that takes into account perceptual orthogonality, and measures the quality of timbre interpolation applicable to perceptually valid timbral sonification. We discuss two timbre representations and measure perceptual judgment. We determined that a timbre space based on Mel-frequency cepstral coefficients (MFCC) is a good model for perceptual timbre space.

1. INTRODUCTION

1.1. Goal and motivation

Timbre is a catch-all term that represents all aspects of a sound, independent of pitch and loudness [1]. A timbre is often described by a combination of subjective perceptual dimensions. While there are many quantitative descriptions of timbre, there is no principled way to synthesize timbres which will lead to a prescribed perception.

The goal of this work is to develop a quantitative mapping between a physical description of a timbre and its percept with the purpose of using timbre in sonification in a meaningful and reliable way.

The need for reliable mappings of data to perceptual space is critical for effective sonification [2]. This study addresses this need in the timbre domain by testing the relationship between perception and physical representations of sound. Our goal is to find a computationally-viable model or representation for timbre that is isomorphic with human perception. We describe this model as a timbre space.

Such a model is vital for timbre-based auditory display. Within the sonification community relative timbre assessment has been used as a basis of data representation including the utilization of vowel space as an intuitive categorical space [3], crystallization sonification [4], and a variety of applications in which traditional musical instrument sounds were used to represent data (for example, [5]). However, the effectiveness of timbre-based sonification is limited by the lack of a generalized representation and a context-free distance metric. In this work, we test a model of timbre space by comparing acoustically derived parameters to comparative perceptual judgments by human subjects.

Before describing our approach to timbre representation, it is worthwhile to compare and contrast previous approaches that describe and measure timbre.

1.2. Timbre descriptions

Although timbre is vital in describing, classifying and categorizing musical, speech and environmental sounds, quantifiable perceptual timbre descriptions are lacking.

Timbral descriptions are often confined to impressionistic adjectival description [6] , [7]. A timbre is 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" and "dull" establish perceptual dimensions that correlate to a combination of parameters including, among others, spectral centroid, spectral flux and attack transience [8].

Speech is a special case regarding timbre. Speech sounds are often categorized by the presence and relationships of their formants, the peaks or resonances in the spectrum. Peterson and Barney [9] plot the location of vowels in this space by noting the typical distribution of formant locations. This approach is useful in understanding how speech is generated, recognized, and categorized. However a perceptual model for speech sounds is not readily extrapolated from this approach.

1.3. 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 [10] [11] 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 [12]. While MFCC coefficients are loosely based on a simple model of auditory perception, their primary benefit is that the different coefficients are statistically in-

dependent so GMMs with diagonal covariance can be used and an MFCC frontend 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 [13], Grey [14] [15] and the subsequent research [8] [16]. It directly measured the distance between two musical sounds. By using multi-dimensional scaling (MDS) the sounds can be represented in a low-dimensional surface (plane or 3d 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 existing sounds, it does not help us 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.

1.4. Timbre Space

Our goal is to create a perceptual space that describes the connection between physical attributes of a timbre and human perception. A good model of timbre perception describes a space of sounds with a number of simple properties and explanations.

The best known perceptual maps involve auditory pitch, auditory loudness, spatial geometry, and color vision. In each case, a relatively simple model connects physical attributes of the sound (mel for pitch, sones for loudness, and the three cones of the visual system) with perceptual judgments. We want to do the same for timbre.

This paper takes a three-step approach. First, we postulate a metric for the quality of a perceptual space, second we describe a mathematical representation of a sound's timbre, finally we measure 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.

A timbre space should be both simple to understand and generate excellent predictions of human perception. Simple is in the eye of the beholder and in this paper we will test two similar signal-processing representations of a sound. 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.

Our test of perceptual parsimony looks at linearity and orthogonality. Linearity suggests that the representation can generate accurate in-between sounds–the perception of an interpolated sound lies half-way in between the original sounds. The vibrone sound in McAdams' work [8] does not fit this criteria since it's MDS representation (McAdams' Figure 1) is not on a line between the original sounds. McAdams' work looked at the linearity of three sounds; in this work we look at 16 sounds at once. Orthogonality says that changes in one parameter do 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.

This paper describes a procedure to measure the quality of the match between a prototypical timbre representation and human perception. We use this procedure to compare two auditory representations and judge which is a better fit to human perception. In the remainder of this paper we will describe the perceptual representations in Section 2, procedures to synthesize timbres from these representations in Section 3, our experiments in Section 4, the analysis method in Section 5, results in Section 6, and finish with conclusions.

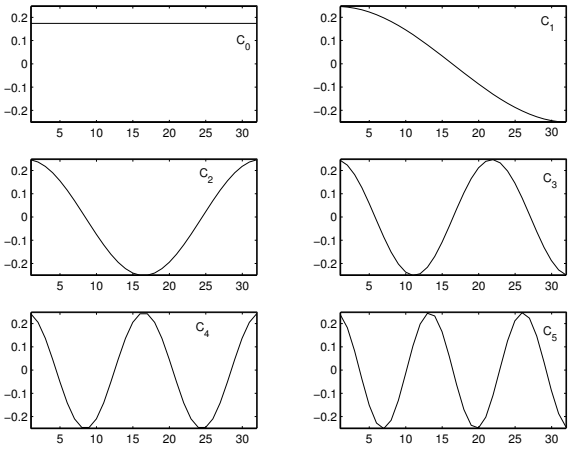

Figure 1: *Kernel functions of DCT. The first coefficient C_0 represents the average power in the spectrum. The second coefficient C_1 represents the broad shape of the spectrum, being equivalent to spectral centroid. Higher-order coefficients, C_2 to C_5, represent finer details of the spectral shape.*

2. REPRESENTATIONS OF THE SOUND

There are many audio representations with different degrees of abstraction, the spectrum being the most common and straightforward form. While a spectrum forms a complete representation of the sound, its arbitrary complexity makes a direct mapping to human perception unknown.

In this work, we study two different and more compact representations of timbre — Linear Frequency Coefficients (LFC) and Mel-frequency Cepstral Coefficients (MFCC) — and measure their ability to model perception. We expect a good representation will map directly to perception, with variations in one parameter's value perceptually orthogonal to changes in another.

MFCC is based on a simple auditory model and is common in the speech recognition world. LFC is a simplification of MFCC that we use for comparison. In both cases the spectral shape of a static sound is represented by a small number of coefficients. The coefficients of LFC and MFCC are ordered. The first coefficient represents the average power in the spectrum. The second coefficient represents the broad shape of the spectrum - roughly the spectral centroid. Higher-order coefficients represent the finer details of the spectral shape. If we use all the coefficients we retain the spectral shape exactly, but we use a handful of coefficients to capture the essence of the spectral shapes. The two representations we study here differ in the details of the spectrum that are removed.

The Discrete Cosine Transform (DCT) of the spectrum is a basic tool in our models [17]. Figure 1 shows the basis function

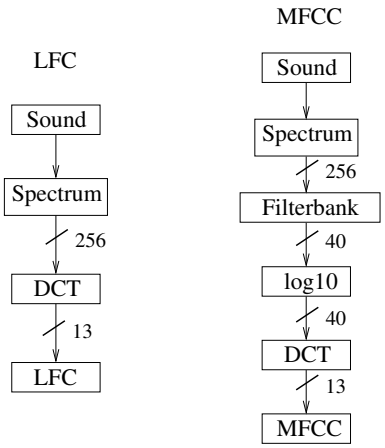

Figure 2: *A summary of the two sound representations studied in this paper. The numbers between the blocks indicate the dimensionality of the data.*

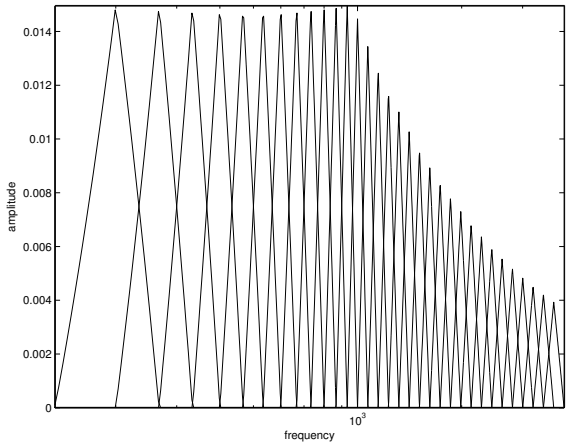

Figure 3: *Frequency responses of 32 filters in the the MFCC filterbank*

used to calculate the DCT. The first few coefficients of the DCT serve to represent the major features of the spectrum.

The DCT is often used for compression because of this feature. In speech recognition, it provides a low dimensional representation of the original audio spectrum. The DCT coefficients are uncorrelated from each other, and this statistical independence enables simpler machine learning models when building speech recognition algorithm. However, the statistical independence is not the same as perceptual orthogonality. Two variables are independent when

$$E\{C_i \cdot C_j\} = E\{C_i\} \cdot E\{C_j\} \tag{1}$$

where $E\{x\}$ is the expected value of x. In other words, knowing the value of one coefficient does not provide any information about the expected value of another coefficient.

Orthogonality is a geometrical concept that says that the dot product of one vector and another is zero, and thus the vectors form a Euclidean space. Ideally, a model of perceptual space can be described by orthogonal basis vectors. Both MFCC and LFC representations use a DCT to simplify and smooth a sound's spectrum.

A summary of the audio representations used in this paper is shown in Figure 2. In both cases, the parameter which defines the spectral shape is arbitrarily limited to a vector of 13 coefficients. Our goal is a test to see which sound representation gives the better model of perception. In this paper we vary the $C_3 \times C_6$ or $C_4 \times C_6$ coefficients, and measure the representations fit to perception judgments.

2.1. Spectrum

The spectrum $S(f)$ of an audio signal $s(t)$ is

$$S(f) = \left|\mathtt{FT}(s(t))\right|. \tag{2}$$

We are only modeling static sounds in this work, therefore we can ignore the phase.

2.2. LFC spectrum

The linear frequency coefficients (LFC) C_i' are computed by finding the DCT of the conventional spectrum as follows:

$$C_i' = \mathtt{DCT}(S(f)) \tag{3}$$

where i is the DCT bin number. This is similar to the calculation of the MFCC described in Section 2.3, but without the MFCC's frequency and loudness compression.

2.3. Mel-frequency Cepstrum Coefficients

The Mel-frequency Cepstrum coefficient (MFCC) is the Fourier transform of a spectrum, where both frequency and amplitude are scaled logarithmically. The frequency warping is done according to the critical bands of human hearing. The procedures for obtaining MFCC from a spectrum are illustrated in the figure 2.

A filterbank of 32 channels, with spacing and bandwidth that resemble the auditory system's critical bands, warps the linear frequency.

The frequency response of the filterbank $H_i(f)$ is shown in the figure 3. The triangular window $H_i(f)$ has the passband of 133.3 Hz for the first 13 channels between 0 Hz and 1 kHz, and a wider passband, which grows logarithmically, from the 14th channel as the frequency becomes higher than 1 kHz. The amplitude of each filter is normalized so that each channel has unit power gain.

$$\mathtt{Bandwidth}(H_i) = \begin{cases} 133.3 & (i \leq 13) \\ 1000 \cdot 1.072^{i-13} & (i > 13) \end{cases} \tag{4}$$

We multiply the triangular frequency response of the filters, as shown in Figure 3, and the sound's spectrum. Then the total energy in each channel, F_i is integrated to find the filterbank output.

$$F_i = \int \left|H_i(f) \cdot S(f)\right| df \tag{5}$$

where i is a channel number in the filterbank, and $H_i(f)$ is the filter response of the ith channel.

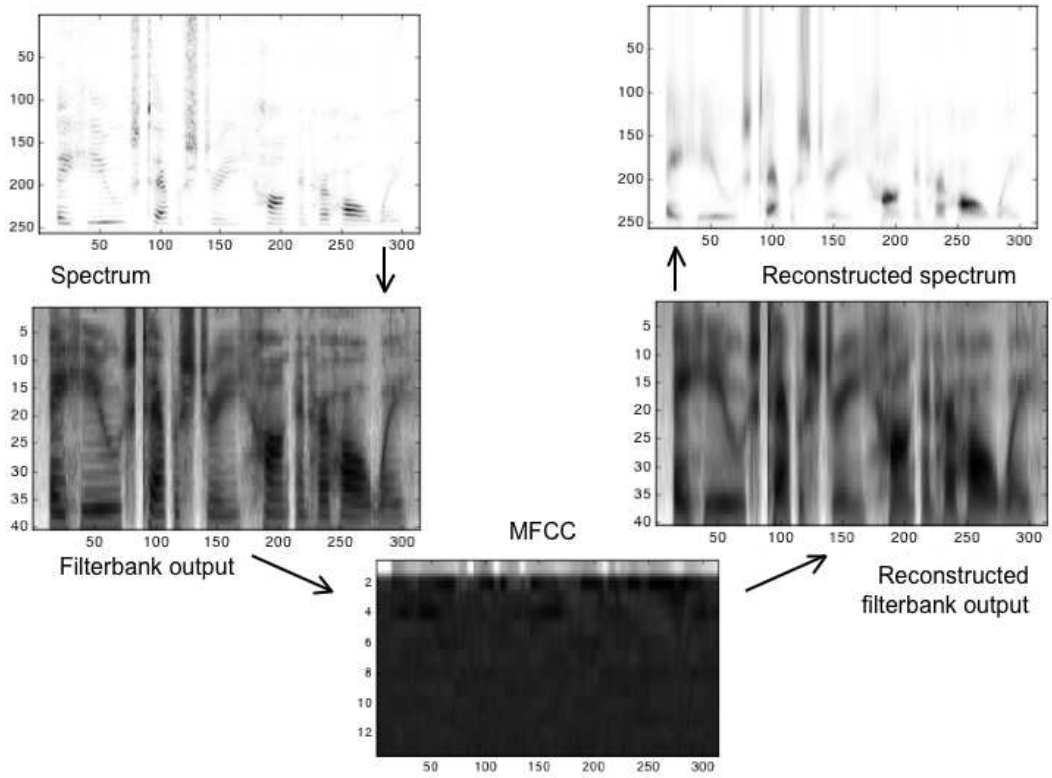

Figure 4: *The process to obtain a reconstructed spectrum: First, the spectrum is fed into the filterbank. The MFCC output is the lower-order coefficients of the DCT on the filterbank output. The IDCT of the MFCC is a smooth version of the filterbank output, discarding the fundamental frequency and its harmonics. The reconstructed spectrum has a smooth spectral shape.*

The Mel-frequency Cepstral coefficients, C_i are computed by taking the DCT of the log-scaled filterbank output.

$$L_i = \log_{10}(F_i) \tag{6}$$

$$C_i = \texttt{DCT}(L_i) \tag{7}$$

Figure 4 shows the computational process, and pictures of the resulting data, at a number of points in the calculation. Note as the signal is processed much of the pitch information, the horizontal striations, disappear. Pitch is artificially added back in when we synthesize from the low-dimensional representations.

2.4. Representation comparison

In both cases, LFC and MFCC, we represent a timbre as a low dimensional vector. Any point in this multidimensional space is a sound, which for visual purposes we can display as a spectrogram. Figures 5 and 6 show an array of points in this space 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 set to zero, and $C_0 = 1$, the spectrum is flat. As the value of C_3 is increased, going down the first column, there is a growing bump in the spectrum at DC and in the mid-frequencies. As the value of C_6 is increased, going across the first row, three bumps increase in size. Notice that the spectral peaks are equi-spaced in frequency for LFC.

3. SYNTHESIS

We test our perceptual models by synthesizing sounds and measuring subjects' perceptions of relative timbral distance. This section describes the steps needed to invert the representation described in the previous section. To make the sounds more life-like, we present the desired timbre with a pitch and vibrato in the vocal range.

3.1. Inverse transform of DCT and MFCC

We synthesize the stimuli by inverting the procedures described in the previous section: We start the synthesis from a given array of, for example, 13 coefficients, which could be interpreted as either LFC or MFCC, depending on the representations.

The reconstruction of the spectral shape from the LFC C_i' is simply the inverse transform of the DCT

$$\tilde{S} = \texttt{IDCT}(C_i'). \tag{8}$$

The reconstruction of the spectral shape from the MFCC starts with the inversion of the DCT and amplitude scaling

$$\tilde{L}_i = \texttt{IDCT}(C_i) \tag{9}$$

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

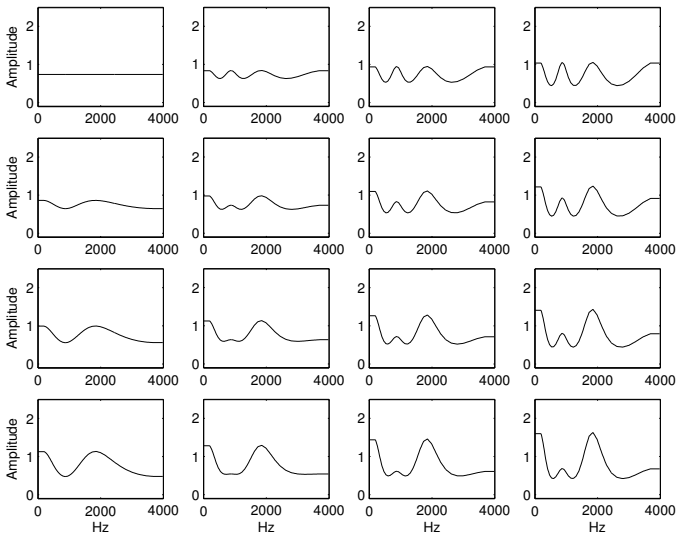

Figure 5: *An array of spectrograms generated for a range of MFCC coefficients. The columns show C_6 ranging from 0 to 0.75, the rows show C_3 ranging from 0 to 0.75.*

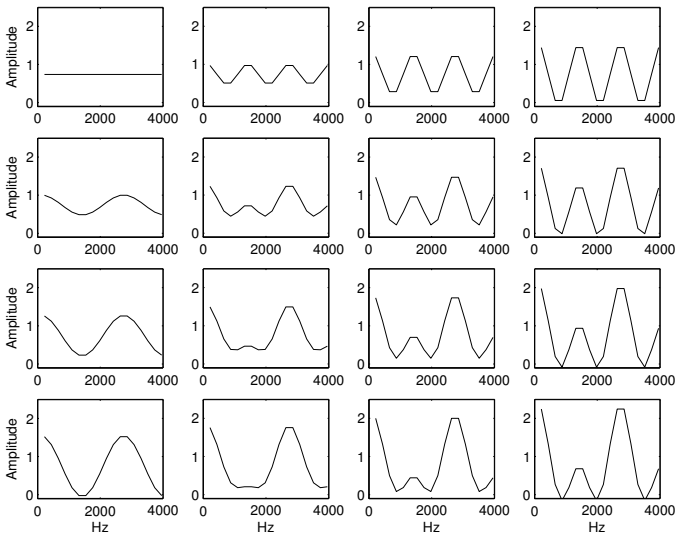

Figure 6: *An array of spectrograms generated for a range of LFC coefficints. The columns show C_6 ranging from 0 to 0.75, the rows show C_3 ranging from 0 to 0.75. Unlike Figure 5, the spectral bumps are spaced linearly in frequency.*

We assume that the reconstructed filterbank output $\tilde{F}_i$ represents the value of the reconstructed spectrum $\tilde{S}(f)$ at the center frequency of each filter bank,

$$\tilde{S}(cf_i) = \tilde{F}_i \tag{11}$$

where cf_i is the center frequency of the ith auditory filter. Therefore in order to obtain the reconstruction of the entire spectrum, $\tilde{S}(f)$, we linearly interpolate the values between the center frequencies $\tilde{S}(cf_i)$. The inversions are summarized in the figure 7.

3.2. Additive synthesis

The voice-like stimuli used in this study are synthesized using additive synthesis of frequency modulated sinusoids. In effect, we start with a flat and infinite harmonic series, and set the level of each harmonic based on the desired smooth spectral shape. The pitch, or fundamental frequency f_0, is set to 220 Hz, with the frequency of the vibrato v_0 set to 6 Hz, and the amplitude of the modulation V set 6 %.

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

$$s = \sum_n \tilde{S}(n \cdot f_0) \cdot \sin(2\pi n f_0 t + V(1 - \cos 2\pi n v_0 t)) \tag{12}$$

where n is the harmonics number.

3.3. Prepared Stimuli

We prepared four sets of stimuli, as shown in the table 1, to test the perceptual judgments of our subjects. For the first group of the stimuli, the [C_3, C_6] entities of the 13 coefficients are set to non-zero value, and for the second group, the [C_4, C_6] entities are set to non-zero value giving the following sets of parameters.

$$[C_3, C_6] = [1, 0, 0, C_3, 0, 0, C_6, 0, 0, 0, 0, 0, 0] \tag{13}$$
$$[C_4, C_6] = [1, 0, 0, 0, C_4, 0, C_6, 0, 0, 0, 0, 0, 0] \tag{14}$$

The values of C_3, C_4 and C_6 are varied over the set

$$C = [0, 0.25, 0.5, 0.75]. \tag{15}$$

These are the same parameter values shown in Figures 5 and 6 The arrays of the coefficients are interpreted as LFC or MFCC, and provide two sets of stimuli for representation, as shown in the table 1.

4. EXPERIMENT

4.1. Procedures

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

The sounds were presented in pairs, where the first is the reference sound and the second is the trial sound to be evaluated, with no pause between the paired stimuli. The reference sound was

Table 1: prepared stimuli

Set	Num. of stimuli	Variable	Representation
1	16	[C_3, C_6]	LFC
2	16	[C_3, C_6]	MFCC
3	16	[C_4, C_6]	LFC
4	16	[C_4, C_6]	MFCC

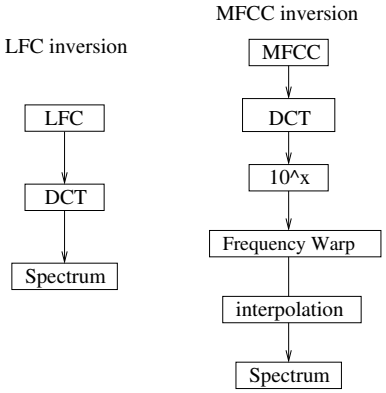

Figure 7: *Inverse transforms of LFC and MFCC*

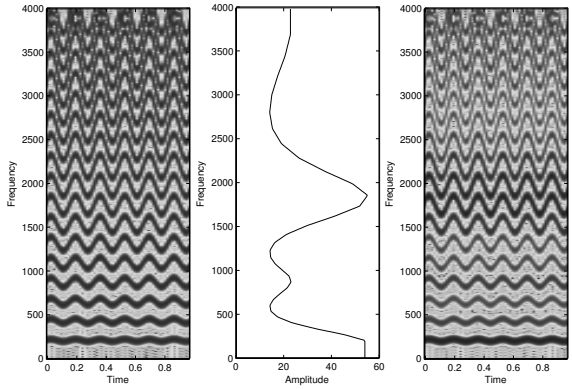

Figure 8: *Additive synthesis of the stimuli: The left spectrogram shows the harmonic series with pitch vibrato. The middle is the spectral shape of the desired sound. The right spectrogram shows the resulting spectrum, in which each harmonics is weighted according to the spectral shape shown in the middle.*

kept identical throughout the entire experiment: It has a flat spectrum, all the 13 coefficients are zero except C_0 (i.e. $[C_3, C_6]$ or $[C_4, C_6] = [0, 0]$.)

The second element of each pair, the trial sound, was varied in each presentation pair. It was synthesized using the discrete parameter values shown in Eq. 15.

In order to provide the time for the evaluation, three seconds of silence was given between pairs. The pairs of sounds were presented to the subjects in a random order.

For each of the four sets of sounds we played five trials to help the subjects understand the types and range of sounds that appear on the main experiment. The first of the example pairs has the widest range of difference between the reference sound and the trial sound ($[0.75, 0.75]$). For the second example, the trial sound is identical to the reference sound ($[0, 0]$). The subsequent three example pairs presented sounds that fell between the extremes presented in the first two sample pairs.

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 difference between pair elements on a scale of one to ten, where one is identical and ten is very different.

4.2. Instruction to the subjects

The following instruction was given to the subjects before the experiment.

You will hear a set of pairs of sounds. Each pair is presented without pause between the paired sounds. The first element of the pair will always be the same sound in all pairs throughout the experiment. The second element of each pair may change in each set presented. Your task is to rate each pair of sounds in terms of degree of similarity on a scale of one to ten where one is identical and ten is very different.

This session has four sections total.

Before each section, you'll hear five example pairs. The first pair will present the widest range of difference between the reference (first) sound and the second element of the pair. The second is the closest to identical that will be heard in the section. The other three example pairs will present sounds that fall between the extremes presented in the first two sample pairs.

4.3. Experiment setup

Ten students, aging 20 – 35 years old, participated in the experiment. The stimuli were presented to the subject using a headset in a quiet office environment.

5. ANALYSIS METHOD

We have two stages in the analysis procedures.

In the first stage, we fit the individual distance judgments to a simple Euclidean model. The residual — the difference between the subject's distance estimates and the model's predictions — is used to evaluate the performance of the representations (LFC and MFCC) on each subject.

The second stage evaluates the average performance of the representation across all subjects. For each of four tests (LFC, MFCC combined with [C_3, C_6] or [C_4, C_6]), we computed the mean of the residuals by ten people, and its standard error. The averaged residuals and the standard errors are used to evaluate the representation and to judge the quality of the perceptual space.

5.1. Individual Euclidean model fitting

We test the quality of a representation (see Section 1.4) by fitting the auditory parameters and the subject's perceptual distance data to a Euclidean distance model. If the subjective data fits a Euclidean model then we say that the representation is a good model of human perception.

For a two-dimensional test as we performed, the Euclidean model says

$$d^2 = ax^2 + by^2 \tag{16}$$

where d is the perceptual distance that subjects reported in the experiment, x is one entity from the 13 coefficients (C_3 or C_4) and y is another entity from the coefficients (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 each subject's perceptual data to a Euclidean model. Using

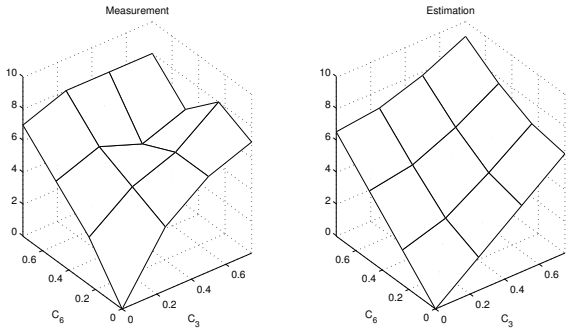

Figure 9: *Plots of perceptual distances, a) measured b) idealized model, for one subject*

matrix notation, Equation (16) is written as follows

$$\mathbf{d^2} = [\, \mathbf{x^2}\, \mathbf{y^2} \,] \cdot [\, a\, b \,] \tag{17}$$

where $\mathbf{d^2}, \mathbf{x^2}, \mathbf{y^2} \in \mathbf{R}^{16}$ (16 is the number of stimuli in a set). The estimation of the regression model is done by the least squares method:

$$[\, \hat{a}\, \hat{b} \,] = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \cdot \mathbf{d^2} \tag{18}$$

where $\mathbf{X} = \left[\mathbf{x^2}\, \mathbf{y^2}\right]$. The left inverse $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is known as pseudo-inverse of the matrix, which guarantees the minimum-error linear estimate.

After estimating the $\hat{a}$ and $\hat{b}$, the estimated perceptual distance for any two timbre coefficients x and y is given as follows

$$\hat{d} = \sqrt{\hat{a}x^2 + \hat{b}y^2}. \tag{19}$$

An example of the measured and estimated perceptual distance for all 16 sounds in one test is shown in Figure 9.

The residual of the linear estimation is:

$$d_{res} = \frac{1}{16}\sum_{x,\,y} \left|\, d - \hat{d} \,\right| \tag{20}$$

This residual, d_{res} is computed for each section of the individual subject, and used to evaluate which audio representation (LFC or MFCC) better fits the Euclidean model.

5.2. Integrating the individual timbre space of the subjects

After computing the residuals for individual subjects, the mean of the residuals across subjects is calculated for each representation

$$\bar{d}_{res} = \frac{1}{N}\sum_{i=1}^{N} d_{res,i} \tag{21}$$

where N is the number of subjects. The standard error is calculated as follows.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} |d_{res,i} - \bar{d}_{res}|^2}{N-1}} \tag{22}$$

$$\sigma_{Mean} = \frac{\sigma}{\sqrt{N}} \tag{23}$$

By comparing the standard error σ_{Mean} of each representation, we evaluate if the representation is a good model of human perception.

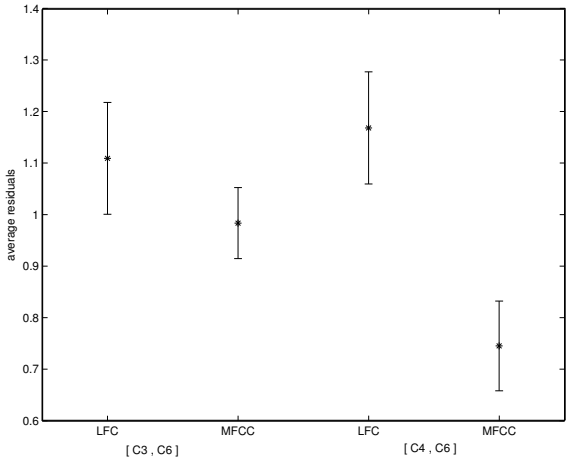

Figure 10: *Model residuals and standard errors: From left, (a) LFC, [C_3, C_6], (b) MFCC [C_3, C_6], (c) LFC [C_4, C_6] (d) MFCC [C_4, C_6].*

6. RESULTS

Figure 10 compares the quality of the two representations of perceptual space — LFC versus MFCC — when compared with two different sets of parameters. By this test, the MFCC representation forms a better model of timbre space than the simplified LFC representation. In other words, the MFCC representation allows for more accurate timbre interpolation and creates a model where the parameter axes are orthogonal.

Figure 10 uses each model's average residual to compare the two representations for two different pairs of dimensions. On average, either timbre space predicts the perceptual judgment with a mean error of 1 point on a 10-point scale. More precisely, the variance of the residuals was 6.8 units for the LFC model (on a 10-point scale) and 3.9 for the MFCC model. In both cases, the models were able to account for 66 % of the variance of the original distance judgements. Figure 11 shows the histograms of the experiment data and the residuals for both LFC and MFCC.

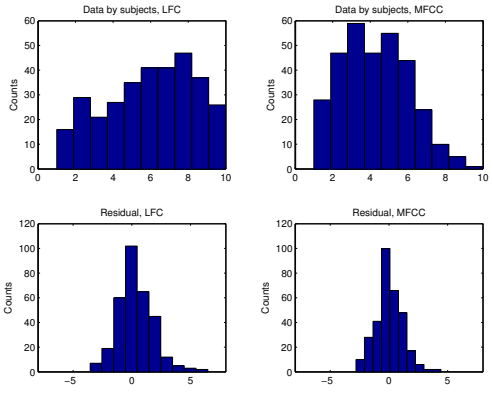

Figure 11: *Data histograms. The top row shows the histogram of the subject's perceptual data, across all coefficients and subjects. The bottom row shows the histogram of the model residuals.*

7. 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 an objective criteria that describes 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 that MFCC is an orthogonal model of perception.

Clearly timbre perception is a highly non-linear process. 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. We are not saying, that MFCC is the best perceptual model: It is the vest model we have tested to date. This paper is the first step towards a complete model of timbre perception.

Most importantly, the timbre representations we tested here are static. 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.

Like the MDS work that precedes this paper, our test based on distance judgments can not discern the principle axes involved in timbre perception. Initial statistical tests suggest that the cepstral coefficients we tested are equally important to timbre judgment. But any rotation of these axis will produce the same distances. In vector-space terms, the DCT calculate both a rotation and a subspace of the original spectral data. Unlike the previous MDS work, the axis we tested here are defined in mathematical terms and thus are amenable to direct synthesis.

Finally, we have not begun to understand the contextual and individual differences involved in timbre perception [18]. Native American-English speakers have a hard time hearing nasalized vowels in the French language, native Japanese speakers do not hear the difference between "l" and "r." Highly-trained audiophiles cringe over acoustic differences that most of us don't hear. Our work aims to understand the common principles of timbre perception.

8. ACKNOWLEDGEMENTS

The initial studies for this work were done as part of the 2004 Telluride Neuromorphic Workshop. We appreciate the thoughtful discussions we have had with Shihab Shamma and Daniel Levitin.

9. REFERENCES

[1] AJM.Houtsma, "Pitch and Timbre: Definition, Meaning and Use." *Journal of New Music Research,* 2, 1997.

[2] S.Barrass, "A Perceptual Framework for the Auditory Display of Scientific Data." *Proceedings of the Second International Conference on Auditory Display ICAD '94*, Santa Fe Institute, New Mexico, 1994.

[3] R.Cassidy, J.Berger, K.Lee, M.Magioni and R.Coifman. "Auditory display of hyperspectral colon tissue images using vocal synthesis models." *Proceedings of the 2004 International Conference on Auditory Display,* Sydney, Australia, 2004.

[4] T.Hermann, and H.Ritter, "Crystallization Sonification of High Dimensional Datasets", *Proceedings of the 2002 International Conference on Auditory Display*, Kyoto, Japan, 2002.

[5] J.Flowers, L.Whitwer, D.Grafel and C.Kotan, "Sonification of Daily Weather Records: Issues of Perception, Attention and memory in Design Choices." *International Conference on Auditory Display,* Espoo, Finland, 2001.

[6] J.M.Hajda, R.A.Kendall, E.C.Carterette, and M.L.Harshberger, "Methodological issues in timbre research." *The Perception and Cognition of Music* Delige & Sloboda (eds.),. pp. 253–306, L. Erlbaum, London, 1997.

[7] C.L.Krumhansl, "Why is musical timbre so hard to understand?" *Structure and perception of electroacoustic sound and music.* pp. 43–54, Excerpta Medica, New York, 1989.

[8] 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." *Psychological Research*, 58, pp. 177–192, 1995.

[9] G.E.Peterson, and H.L.Barney, "Control methods used in a study of the vowels." *J. Acoust. Soc. Am.* 24, pp. 175–184, 1952.

[10] H.Fletcher, "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, pp. 59–69. 1934 .

[11] J.B.Allen, "How do humans process and recognize speech?" *IEEE Trans. on Speech and Audio Proc.*, 2(4) pp. 567–577, October 1994.

[12] S.B.Davis, P.Mermelstein. "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 pp. 357–366, 1980

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

[14] J.M.Grey. "An exploration of musical timbre." PhD dissertation, Stanford University, 1975.

[15] J.Grey, "Multidimensional Scaling of Musical Timbres." *Journal of the Acoustical Society of America* 61(5): pp. 1270–1277, 1976.

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

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

[18] D.C.Dennett, "Quining Qualia", *Consciousness in Modern Science* Eds. A.Marcel, and E.Bisiach, Oxford University Press, Oxford, 1988.