THE THIRTEEN COLORS OF TIMBRE

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

CCRMA[†], Department of Music
Stanford University, Stanford, California, USA
{hiroko, brg}@ccrma.stanford.edu

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

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 three timbre representations and measure perceptual judgments. We determine that a timbre space based on Mel-frequency cepstral coefficients (MFCC) is a good model for a perceptual timbre space.

1. INTRODUCTION

Timbre is defined as "that attribute of auditory sensation, in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar" [1]. This paper considers a perceptual space that may be useful in studying the role timbre plays in sound perception. We compare and constrast three different representations for timbre and compare their relevance to perception

Our work has two goals. From a scientific viewpoint, we want to understand how people perceive sound and speech. We want to build a model of sound perception that is as fundamental as the three-color model for vision. From an engineering viewpoint, we want to find a general representation for sound that is a parsimonious description of perception because it could lead to better sound analyzers.

This paper takes a different approach to timbre perception than previous work. The timbre work based on multidimensional scaling [2, 3, 4] start with sounds, measure perceptual distances, and then tries to synthesize a representation or coordinate system which explains the MDS axis. In this work, we start with a coordinate system, synthesize sounds based on this representation, and then measures how well each representation fits our criteria for the optimum perceptual space.

This paper takes a three-step approach. First, we describe a metric for the quality of a perceptual space, second we describe mathematical representations 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.

2. REPRESENTATIONS OF THE TIMBRE

2.1. Parameterization

There are many audio representations with different 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. 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) [5]. The DCT not only removes high-frequency ripples in the spectrum, but serves to decorrelate 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 power to MFCC. We start with a linear-frequency scale and a linear amplitude scale. A 13-D DCT of the normal amplitude spectrum reduces the dimensionality of the spectral space and smooths the spectrum. Both MFCC and LFC use a DCT to reduce the dimensionality and decorrelate 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 from C_0 to C_{12}, where 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.

Pollard's tristimulus model is a popular approach for describing timbre. In the tristrimulus approach any harmonics sound is represented as a 2D point. The tristimulus coefficients are computed as follows. We calculate the loudness of each harmonic N_i according to Zwicker's specific loudness method. Next, the harmonics are devided into three groups: one for the fundamental $i = 1$, and one each for partials $i = 2 \ldots 4$ and $i = 5 \ldots n$. The loudness for each group is computed using Stevens' law that

$$N_i = 0.85 N_{max} + 0.15 \sum_j N_j \qquad (1)$$

where j ranges over the harmonics in set i, N_{max} is the value of the largest partial in the group of the partials. Finally, the loudness of all groups is normalized by the total loudness of three groups: $T1 = N_1/N$, $T2 = N_2^4/N$, $T3 = N_5^n/N$ where $N = N_1 + N_2^4 + N_5^n$. A larger T1 is associated with a "strong fundamental", while a larger T2 means "strong mid-frequency partials" and a larger T3 means "strong high-frequency partials."

2.2. Resynthesis

In this study, we choose a 13-D vector and then synthesize sounds from these coefficients using the inverse transforms of LFC and

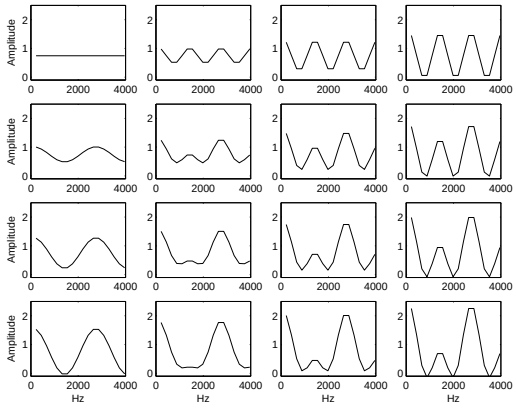

Figure 1: *An array of spectra generated for a 2-D range of LFC coefficients. The column show C_3 ranging from 0 to 0.75, the rows show C_6 ranging from 0 to 0.75. Compare the uniformity of the frequency spacing of the peaks to those of Figure 2.*

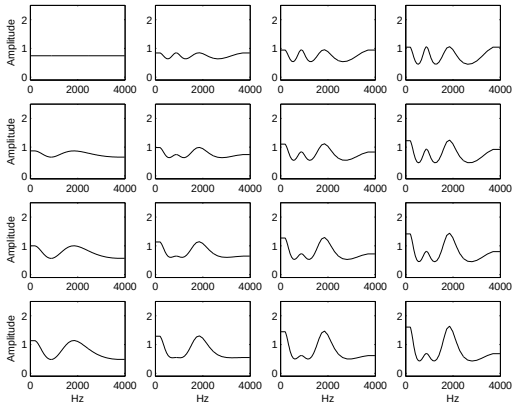

Figure 2: *An array of spectra generated for a 2-D range of MFCC coefficients. The column show C_3 ranging from 0 to 0.75, the rows show C_6 ranging from 0 to 0.75.*

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

2.3. Prepared Stimuli

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

Two of the 13 coefficients are chosen as variables and set to non-zero values. For example, in the $[C_3, C_6]$ space, the parameter vector is $[C_3, C_6] = [1, 0, 0, C_3, 0, 0, C_6, 0, 0, 0, 0, 0, 0]$. The values of these parmeters are varied over the set $C = [0, 0.25, 0.5, 0.75]$. The vector is interpreted as LFC or MFCC for resynthesis.

2.4. Representation comparison

Any point in LFC or MFCC space is a sound. Figures 1 and 2 show 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.

We can analyze any sound that we synthesize in LFC and MFCC space with the tristimulus model and represent it as a point in a 2D tristimulus space. The result of this analysis for the [C_4,C_6] spaces are shown in Figure 3. Here each of the spectra we will use in our experiment are plotted as a point in the T2-T3 tristimulus space. The mapping between the rectangular sampling grid in MFCC space and each sound's point in tristimulus space is especially non-linear.

2.5. Additive FM synthesis

The voice-like stimuli used in this study are synthesized from the spectrum derived in Section 2.2 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%. Using the reconstructed spectral shape $\tilde{S}(f)$, with the harmonics number n, the synthesized sound is

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

3. EXPERIMENT

We measured the distance for several sets of timbre parameters by asking subjects for their subjective evaluation of the difference

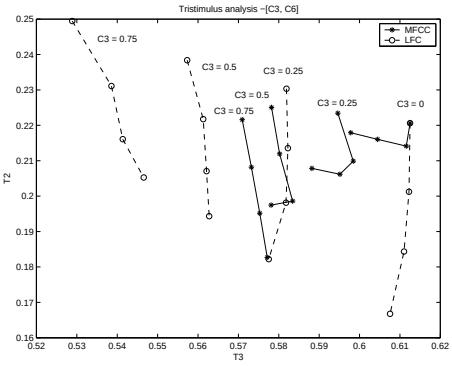

Figure 3: *Tristimulus plots of stimuli. C6 starts at 0 at the top of each line and increases to 0.75.*

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.

Ten 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.

4. 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.

4.1. Individual Euclidean model fitting

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

$$d^2 = ax^2 + by^2 \tag{3}$$

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:

$$d_{res} = \frac{1}{16} \sum_{x,\ y} \left| \ d - \hat{d} \ \right| \tag{4}$$

where $\hat{d}$ is the estimated distance by the linear regression model. Figure 4 shows the measured perceptual distances for one subject and the estimated Euclidean model.

4.2. Integrating the individual timbre space of the subjects

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

$$\bar{d}_{res} = \frac{1}{N} \sum_{i=1}^{N} d_{res,i} \tag{5}$$

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

$$\sigma_{Mean} = \frac{\sqrt{\sum_{i=1}^{N} |d_{res,i} - \bar{d}_{res}|^2}}{N} \tag{6}$$

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.

5. RESULTS

Figure 5 compares the quality of the two perceptual spaces, LFC versus MFCC, when tested with five different 2-D sets of parameters. On average, either timbre space predicts the perceptual judgment with a mean error of 1 point on a 10-point scale. In all cases, the MFCC representation forms a better model of timbre space than the simplified LFC representation. In other words, the 13 colors in the MFCC representation allows for more accurate timbre interpolation and creates a model where the parameter axis are orthogonal than the other representations we tested.

For most pairs of dimensions within a representation, the model error is relatively constant. This result suggests that these pairs of dimensions form an orthogonal perceptual model of timbre. This is true even for a range of dimensions as close as C_3 and C_4 and as wide as C_3 and C_{12}. But quite notably, the model error jumps dramatically when we studied C_{11} and C_{12} dimensions. Since C_3 and C_{12} proved to be a good model, evaluated by interpolation and orthogonality, this suggests that the perceptual model is still linear for higher-order dimensions. But when C_{11} and C_{12} are paired the model error goes up, suggesting that these two dimensions are not as orthogonal as the others.

The Euclidean models does an excellent job of predicting the perceptual judgements. The variance of the residuals was 6.8 units^2 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 6 shows model residuals based on the tristimulus parameters. In this comparison, the same sounds and perceptual judgements used in the LFCC/MFCC comparison are reevaluated using their tristimulus parameters. In general, in a pair-wise comparison, the LFC and MFCC models have a smaller residual error (Figure 5) than the timbre representation based on the tristimulus model (Figure 6).

Arguably, the tristimulus approach is at a disadvantage with our tests because our reference sound is not at (0,0) in the tristimulus space. We can compensate for this artifact by using a richer Euclidean model that includes a offset ($d^2 = ax^2 + by^2 + c$). When we did this, the tristimulus model residuals went down, as one would expect since there is one more free parameters, but still

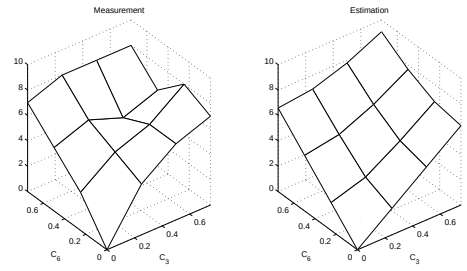

Figure 4: *Plots of perceptual distances, a) measured b) idealized model, for one subject.*

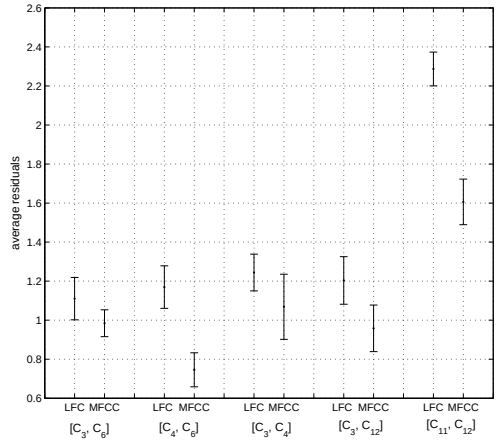

Figure 5: *Model residuals and standard errors comparing MFCC and LFC for five sets of axis.*

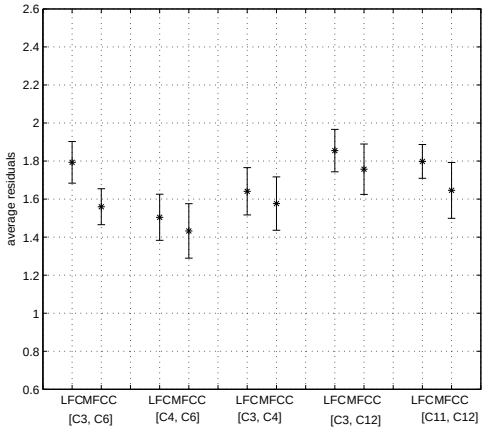

Figure 6: *Model residuals and standard errors for 10 sets of sounds as represented by the tristimulus model.*

the residuals did not fall to the same level as the MFCC model shown in Figure 5.

6. CONCLUSIONS

In this paper we have articulated a set of criteria for evaluating a timbre space, described three 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 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.

We can not make a comparison of LFC and MFCC spaces just based on Figure 6. As far as the tristimulus model is concerned, the 16 sounds we used to test our models are arbitrary sounds. Further, we expect the Euclidean model to hold whether the sounds are on a rectangular grid, as they are in the LFC and MFCC case, or randomly positioned as they are in the tristimulus space. We expect all these models to only work in a local region—these initial models are bound to fail at extreme points in timbre space. The sounds that we synthesized from LFC space cover a wider range of the tristimulus space than the MFCC-generated sounds. This different could account for the difference in performance between MFCC-generated sounds and LFC-generated sounds in Figure 6.

Our comparison of LFC and MFCC representations is based on two separate sets of sounds that are synthesized from 2D manifolds in LFC and MFCC space. Since these two sets of sounds cover different extents of the timbre space, we can not be sure that their relative difference in our tests are caused by the different coverage of timbre space. We are evaluating how we can perform the tests we describe here, sampled in arbitrary 13-D timbre spaces.

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

Finally, we have not begun to understand the contextual differences involved in timbre for sound perception [7]. However, this work addresses the underlying representational issues.

7. 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, Stephen McAdams, Dan Ellis and Tom Rossing.

8. REFERENCES

[1] B.C.J.Moore. *An introduction to the psychology of hearing, fifth ed.* Academic Press, 2003.

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

[3] 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.

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

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

[6] H.F.Pollard, E.V.Jansson. "A Tristimulus Method for the Specification of Musical Timbre" *Acustica*, 51, pp. 162–171, 1982.

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