A Timbre Space for Speech

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

Center for Computer Research in Music and Acoustics[†]
Stanford University, Stanford, California, USA

IBM Almaden Research Center[‡]
San Jose, California, USA

{hiroko, malcolm, brg}@ccrma.stanford.edu

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 measure perceptual judgments. We determine 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 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]. By this definition, timbre is a component of all speech sounds. Thus this paper considers a perceptual space that may be useful in studying the role timbre plays in speech 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 and speech perception that is as fundamental as the three-color model for vision. From an engineering viewpoint, we want to find a general representation for speech that is a parsimonious description of perception because it could lead to better speech recognizers.

This paper takes a three-step approach. First, we describe 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.

1.2. Speech Description

Formants frequencies are a common tool for describing speech. Using the frequencies of the first few peaks in the spectrum we can place vowel-like sounds in a low-dimensional diagram and how these formant frequencies change over time. But it is not clear if our brains perceive sounds in terms of formant frequencies. Recent work [2] describes speech perception as a form of spectral template matching. We want to know how to turn the spectrum into a perceptual space.

1.3. Speech Distances

An early approaches to understand sound perception was undertaken by Harvey Fletcher. This work [3] measured subject's ability to correctly recognize nonsense words in the presence of filtering and noise. Confusion matrices [4] suggest a distance between speech sounds: sounds that are easily confusable are close together. However these approaches only apply 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 great success modeling the acoustic world using Mel-frequency cepstral coefficients (MFCC) [5]. MFCC coefficients are statistically independent so Gaussian mixture models (GMM) with diagonal covariance can be used. But there has been no systematic, quantitative study of precisely how well MFCC representations of speech and other acoustic signals match the perceptual representation of the signals. Perceptual studies served as the inspiration for MFCC, but this does not mean that the implementation is an accurate model of perception.

A different approach to measuring timbre perception, which directly measures the perceptual distance between two sounds, was proposed by Wessel [6], Grey [7] and others [8] [9]. By using multi-dimensional scaling (MDS) the sounds can be represented in a low-dimensional space 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 axis produced by the MDS algorithm are not labeled. Secondarily, while this approach is descriptive of existing sounds, it does not help us interpolate between sounds. For this we need to find and describe a timbre space that matches human perception.

1.4. Principles of Timbre Space

We want a representation of sound, more general than formants, independent of pitch and loudness, that may be fundamental to future speech-perception and -recognition research.

A parsimonious description of timbre must have the following three properties. First, it must be consistent with perception—it should accurately predict the perceptual distance between two sounds. Second, it must be simple—we judge simplicity by requiring that the underlying representation's axis are orthogonal. Third, it must describe a linear space—we want to be able to interpolate and describe in-between sounds using a straight line.

In this work, we tested two spectral-shape representations of timbre: MFCC and a strawman we call linear frequency coefficients (LFC). In each case we synthesize diverse timbres from these representations, and measure the match between the representation coefficients and the perceptual judgements. We mea-

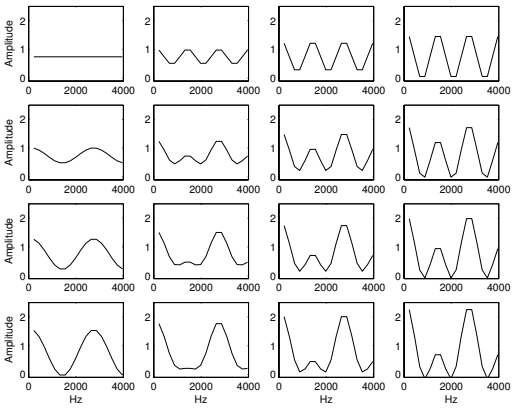

Figure 1: *An array of spectra generated for a 2-D range of LFC coefficients. The column show C_3 ranging from 0 to 0.75, the rows show C_6 ranging from 0 to 0.75. Compare the uniformity of the frequency spacing of the peaks to those of Figure 2.*

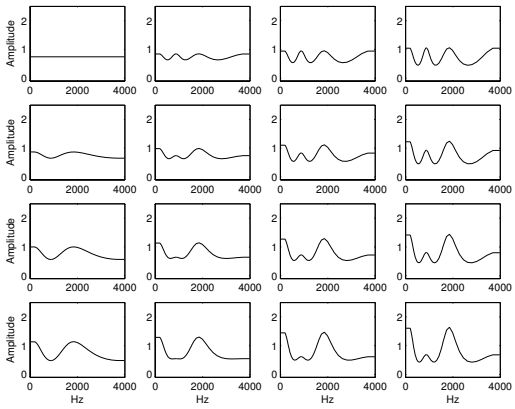

Figure 2: *An array of spectra generated for a 2-D range of MFCC coefficients. The column show C_3 ranging from 0 to 0.75, the rows show C_6 ranging from 0 to 0.75.*

sure the parsimony of the representation by assuming a linear model and fitting the data to a Euclidean model. The best description of timbre space fits a Euclidean model.

2. Representations of the sound

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) [10]. 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 initial stages of 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. Next we describe how we convert these 13-D representations into their equivalent spectra, and then back into sound.

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 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. Representation comparison

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

3. Synthesis

3.1. 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

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

3.2. Prepared Stimuli

As it is difficult to fully explore a 13-D space, we chose discrete pairs of coefficients, and measured subject's perceptual judgements in these 2-D spaces. Two such spaces are shown in Figures 2 and 1. 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.

4. 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.

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.

5. 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.

5.1. 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

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

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:

$$d_{res} = \frac{1}{16} \sum_{x,\ y} \left| \ d - \hat{d} \ \right| \tag{3}$$

where $\hat{d}$ is the estimated distance by the linear regression model. Figure 3 shows the measured perceptual distances for one subject and the estimated Euclidean model.

5.2. 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

$$\bar{d}_{res} = \frac{1}{N} \sum_{i=1}^{N} d_{res,i} \tag{4}$$

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

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} |d_{res,i} - \bar{d}_{res}|^2}{N-1}} \tag{5}$$

$$\sigma_{Mean} = \frac{\sigma}{\sqrt{N}} \tag{6}$$

By comparing the standard error σ_{Mean} of each representation, we decide which representation is a better model of human perception.

6. Results

Figure 4 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 MFCC representation allows for more accurate timbre interpolation and creates a model where the parameter axis are orthogonal.

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.

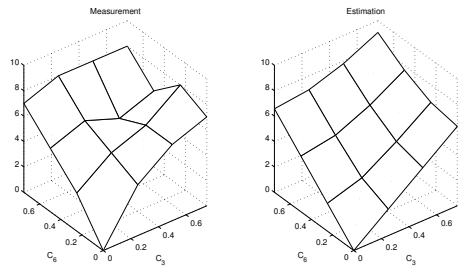

Figure 3: *Plots of perceptual distances, a) measured b) idealized model, for one subject*

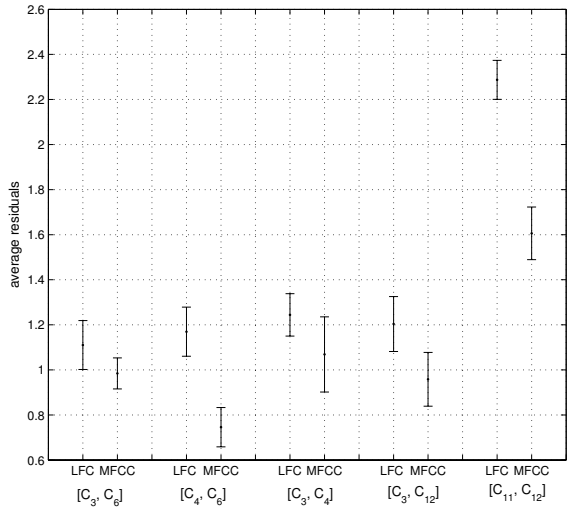

Figure 4: *Model residuals and standard errors: From left, [C_3, C_6], [C_4, C_6],[C_3, C_4], [C_3, C_{12}], [C_{11}, C_{12}]*

We also evaluated the magnitude of each dimension's effect on the resulting perceptual judgements. This is quantified by the size of the a and b parameters in the fit to Equation 2. In our initial study, these quantities are similar in size, for all tested dimensions, suggesting that all 13 dimensions are equally important in the perceptual distance judgement.

The Euclidean models did 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.

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 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 the small sense, the results described here are not surprising. MFCC has enjoyed well-deserved success as a means of representing sounds in speech-recognition systems. But this paper makes two contribution to our understanding of speech perception. First, we have measured the connection between the MFCC representation and perception—MFCC is a good model of perceptual distance. Second, we have established a procedure for testing new timbre and speech representations and comparing their results to perception. With this test we hope to find even better representations of audio signals to help us understand speech perception and build better speech recognizers.

Most importantly, the timbre representations we tested here are static; speech is 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 and individual differences involved in timbre for speech perception [11]. However, this work addresses the underlying representational issues.

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, Daniel Levitin, Karon MacLean, Stephen McAdams, Dan Ellis, Peter Assman, Steven Greenberg and John Lazzaro. Thanks to Grace Leslie for helping with the data analysis.

9. References

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

[2] M.Ito, J.Tsuchida, and M.Yano. "On the effectiveness of whole spectral shape for vowel perception." *Journal of Acoustical Societry of America* 110(2) pp. 1141–1149, 2001.

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

[4] G.A.Miller, P.Nicely. "An analysis of perceptual confusions among some English consonants." *Journal of Acoustical Society of America* 27, pp. 338–352, 1955.

[5] 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

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

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

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

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

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