A statistical model of timbre perception

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

CCRMA[†], Department of Music
Stanford University, Stanford, California, USA
Yahoo! Research[‡]
Sunnyvale, 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 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.

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 two different representations for timbre and compare their relevance to perception.

We aim for a timbre descriptor which is a parsimonious model of human timbre perception. 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 perspective, a compact description could provide the basis for improved sound analysis methods.

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, synthesizes a representation or coordinate system, and then explain 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.

Besides the perceptual studies of timbre, there have been statistical approaches in audio signal processing, using various features to capture the timbral quality. Features such as spectral centroid, or Mel-frequency Cepstral Coefficients (MFCC) have been used to model spectral shapes. However the quantitative causality between these features and the timbre perception is still unknown, and such knowledge is needed for optimizing countless features to perceptually essential ones, hopefully resolving the curse of dimensionality.

This is a part of a series of investigations. Our previous works showed that MFCC is a good representation for human perception of timbre compared to an alternative representation. This alternative representation is called linear frequency coefficients (LFC), and as described in the later section, has the same statistical property as MFCC except the perceptual weightings. However, it is possible that the LFC test stimuli covered a larger perceptual space than MFCC test stimuli used in the previous experiments. When a representation covers more space than the other, it may not be a fair comparison. Taylor's theorem suggests that a linear model is more accurate in a smaller neighborhood. In this experiment, we force the range of timbres covered by LFC stimuli to be smaller than MFCC stimuli, compare the representations, so that we can judge their relevance to timbre perception.

In this introduction, we describe the type of timbre metric needed to evaluate the quality of a perceptual space. In the later sections, we describe mathematical representations of a sound's timbre, and then 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 of spectral shape

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 [5]. 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) [6]. 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 ability 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 spec-

trum. 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.

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. Prepared 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 judgements 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, $C_m = [0, \frac{1}{3}M, \frac{2}{3}M, M]$ 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 2D space, resulting in a set of 16 stimuli samples with varying spectral shapes.

2.4. Designing LFC stimuli space

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 2.2, 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

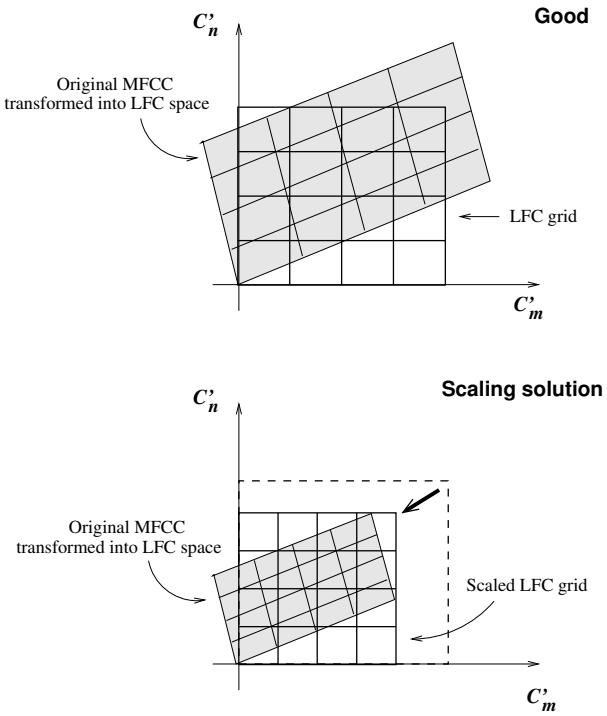

Figure 1: *Corresponding MFCC space and LFC space. Top: comparison of two spaces is fair when covering similar region in one representation (MFCC space is transformed into LFC space.) Bottom: When the transformed MFCC space has much smaller region than LFC parameter space (dotted rectangular), LFC parameter space is rescaled to match with the size of LFC-transformed MFCC space (solid grid).*

reinforces the probability of MFCC being a better representation of timbre.

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.

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

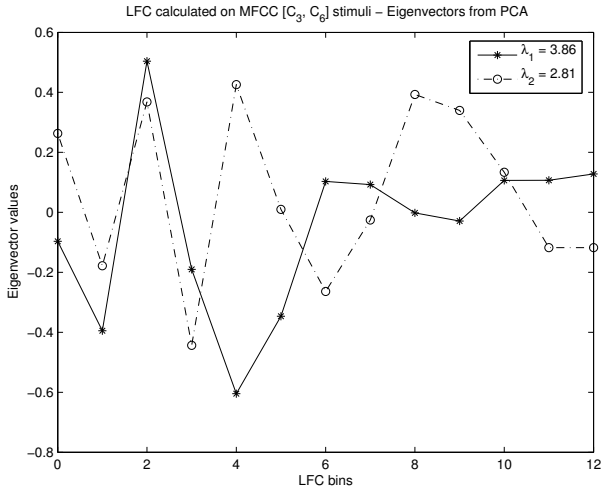

Figure 2: *Eigenvectors of PCA – LFC-transformed MFCC* $[C_3, C_6]$ *stimuli. The first two eigenvectors with their eigenvalues in the legend. It is visible that LFC* C'_2 *and* C'_4 *deliver most of the energy. These two dimensions are chosen to form a corresponding 2-D LFC space.*

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[7]. The 13-dimensional mean vector and the 13×13 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 eigenvalue. Call these sorted eigenvectors $\mathbf{e_1}$ with eigenvalues λ_1, $\mathbf{e_2}$ with eigenvalues 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

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

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

where $\mathbf{C''}_m$ 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 2.4 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	
$[C_m, C_n]$	$[M, N]$	$[C'_m, C'_n]$	$[M', N']$
$[C_3, C_6]$	$[0.75, 0.75]$	$[C'_2, C'_4]$	$[-0.20, 0.32]$
$[C_4, C_6]$	$[0.75, 0.75]$	$[C'_3, C'_4]$	$[-0.29, 0.17]$
$[C_3, C_4]$	$[0.75, 0.75]$	$[C'_2, C'_3]$	$[-0.20, -0.21]$
$[C_{11}, C_{12}]$	$[0.75, 0.75]$	$[C'_5, C'_6]$	$[-0.13, -0.12]$

2.5. 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 corresponding stimuli set that we test this time.

2.6. Additive 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

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

3. 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.

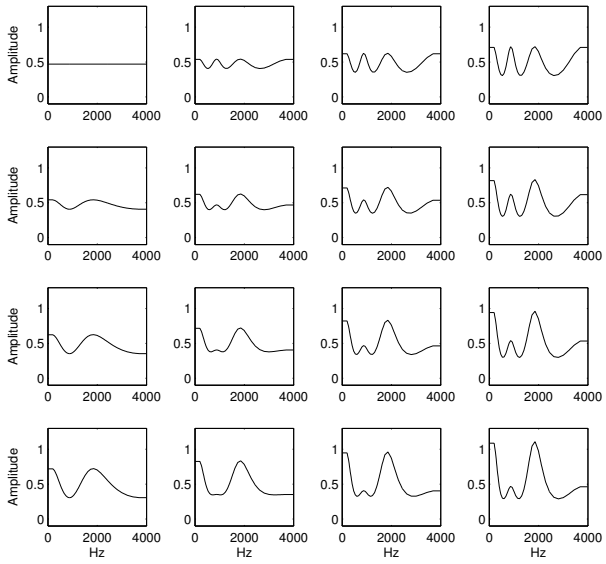

Figure 3: *An array of spectra generated for a 2-D range of MFCC coefficients. The column show C_3 ranging from 0 to 0.75, the rows show C_6 ranging from 0 to 0.75.*

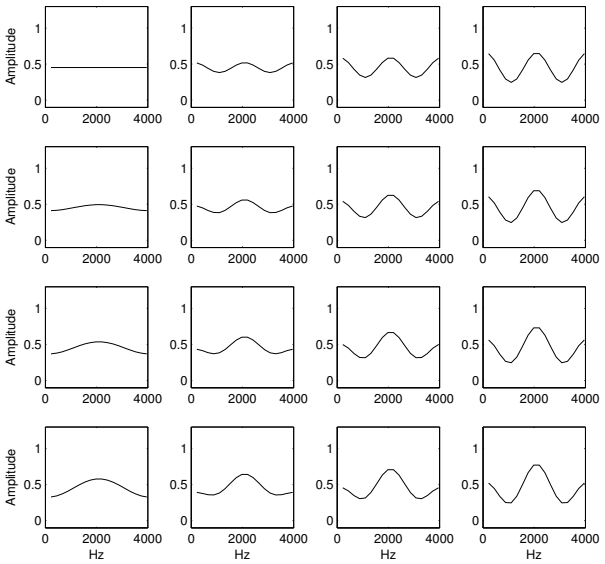

Figure 4: *An array of spectra generated for a 2-D range of LFC coefficients. The column show C_2' ranging from 0 to -0.20, the rows show C_4' ranging from 0 to 0.32.*

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

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

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.

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

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

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

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

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.

5. 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 2.4. 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.

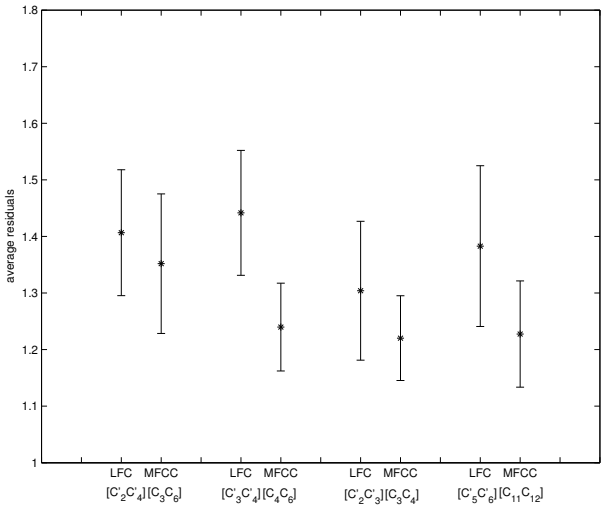

Figure 5: *Model residuals and standard errors comparing MFCC and LFC for four sets of corresponding subspaces.*

6. 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 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; 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 [8]. However, this work addresses the underlying representational issues.

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

[7] R.O.Duda, P.E.Gart and D.G.Stork. *Pattern Classification, second ed.* Wiley-Interscience, pp. 114–117, 2001.

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