Chapter 6

FEATURES FOR AUDIO CLASSIFICATION

Jeroen Breebaart and Martin F. McKinney

Abstract Four audio feature sets are evaluated in their ability to differentiate five audio classes: popular music, classical music, speech, background noise and crowd noise. The feature sets include low-level signal properties, mel-frequency spectral coefficients, and two new sets based on perceptual models of hearing. The temporal behavior of the features is analyzed and parameterized and these parameters are included as additional features. Using a standard Gaussian framework for classification, results show that the temporal behavior of features is important for automatic audio classification. In addition, classification is better, on average, if based on features from models of auditory perception rather than on standard features.

Keywords Audio classification, automatic content analysis

6.1 Introduction

Developments in Internet and broadcast technology enable users to enjoy large amounts of multimedia content. With this rapidly increasing amount of data, users require automatic methods to filter, process and store incoming data. Examples of applications in this field are automatic setting of audio equalization (e.g., bass and treble) in a playback system, automatic setting of lighting to correspond with the mood of the music (or vice versa), automatic cutting, segmenting, labeling, and storage of audio, and automatic playlist generation based on music similarity or some other user specified criteria. Some of these functions will be aided by attached *metadata*, which provides information about the content. However, due to the fact that metadata is not always provided, and because local processing power has increased tremendously, interest in *local* automatic multimedia analysis has increased. A major challenge in this field is the automatic classification of audio. During the last decade, several authors have proposed algorithms to classify incoming audio data based on different algorithms [Davis & Mermelstein, 1980; Wold et al., 1996; Spina &

W. Verhaegh et al. (eds.), Algorithms in Ambient Intelligence, 113-129

Zue, 1996; Scheirer & Slaney, 1997; Spina & Zue, 1997; Scheirer, 1998; Zhang et al., 1998; Wang et al., 2000a; Wang et al., 2000b; Zhang & Kuo, 2001; Li et al., 2001]. Most of these proposed systems combine two processing stages. The first stage analyzes the incoming waveform and extracts certain parameters (features) from it. The feature extraction process usually involves a large information reduction. The second stage performs a classification based on the extracted features.

A variety of signal features have been proposed for general audio classification. A large portion of these features consists of low-level signal features, which include parameters such as the zero-crossing rate, the signal bandwidth, the spectral centroid, and signal energy [Davis & Mermelstein, 1980; Wold et al., 1996; Scheirer & Slaney, 1997; Scheirer, 1998; Wang et al., 2000a; Wang et al., 2000b]. Usually, both the averages and the variances of these signal properties are included in the feature set. A second important feature set which is inherited from automatic speech recognizers consists of mel-frequency cepstral coefficients (MFCC). This parametric description of the spectral envelope has the advantage of being level-independent and of yielding low mutual correlations between different features for both speech [Hermansky & Malayath, 1998] and music [Logan, 2000]. Classification based on a set of features that are uncorrelated is typically easier than that based on features with correlations.

Both low-level signal properties and MFCC have been used for general audio classification schemes of varying complexity. The simplest audio classification tasks involve the discrimination between music and speech. Typical classification results of up to 95% correct have been reported [Toonen Dekkers & Aarts, 1995; Scheirer & Slaney, 1997; Lu & Hankinson, 1998]. The performance of classification schemes usually decreases if more audio classes are present [Zhang et al., 1998; Zhang & Kuo, 2001]. Hence, the use of features with high discriminative power becomes an issue. In this respect, the MFCC feature set seems to be a powerful signal parameterization that outperforms low-level signal properties. Typical audio classes that have been used include clean speech, speech with music, noisy speech, telephone speech, music, silence and noise. The performance is roughly between 80 and 94% correct [Foote, 1997; Naphade & Huang, 2000a; Nahphade & Huang, 2000b; Li et al., 2001].

For the second stage, a number of classification schemes of varying complexity have been proposed. These schemes include Multivariate Gaussian models, Gaussian mixture models, self-organizing maps, neural networks, k-nearest neighbor schemes and hidden Markov models. Some authors have found that the classification scheme does not influence the classification accuracy [Scheirer & Slaney, 1997; Golub, 2000], suggesting that the topology of the feature space is relatively simple. An important implication of these results

is that, given the current state of audio classifiers, perhaps further advances could be made by developing more powerful features or at least understanding the feature space, rather than building new classification schemes.

Thus, our focus here is on features for classifying audio. We compare the two existing feature sets most commonly used, low-level signal properties and the MFCC, with two new feature sets and evaluate their performance in a general audio classification task with five classes of audio. The two new feature sets, described in detail below, are based on perceptual models of auditory processing. Additionally, a more advanced method of describing temporal feature behavior will be discussed which extends the traditional way of including mean and variance of feature trajectories.

6.2 Method

Our audio classification framework consists of two stages: feature extraction followed by classification. We compare four distinct feature extraction stages to evaluate their relative performance while in each case using the same classifier stage. The feature sets (described below) are: (1) standard low-level (SLL) signal properties; (2) MFCC; (3) psychoacoustic (PA) features including roughness, loudness and sharpness; and (4) an auditory filter representation of temporal envelope (AFTE) fluctuations. The audio database consists of a subset of a larger audio database (approximately 1000 items). Two volunteers listened to all tracks and classified the audio according to 21 pre-defined audio categories. Furthermore, a rating was given of how well each audio item matched the audio category. From these ratings, a 'quintessential' audio database was made of all tracks that had the same label for both volunteers and had a sufficiently high rating. Finally, five general audio classes were obtained by combining the 21 audio classes: classical music, popular music (including jazz, country, folk, electronica, latin, rock, rap, etc), speech (male, female, English, Dutch, German and French) crowd noise (applauding and cheering) and background noise (including traffic, fan, restaurant, nature, etc). The number of files in each general audio class is given in Table 6.1.

The classification process begins with the extraction of a set of features, i.e., feature vectors, from each sound file. Feature vectors are calculated on consecutive 32768-sample frames (743 ms at 44.1 kHz sampling rate) with a hop-size of 24576. The choice of a 743-ms frame length was based on a finding of Spina & Zue [1996]. Using a Gaussian-based classification mechanism to classify audio into several categories, they operated on MFCC and found that performance increased as the analysis frame size increased to about 700 msec and then saturated (and decreased a little) with further increases. The resulting feature vectors are grouped into classes based on the type of audio.

Table 6.1. Audio database by class.

class name	*files*
popular music	175
classical music	35
speech	31
noise	25
crowd noise	31

The resulting feature vectors from each class are divided into two groups, a training group and a test group: a randomly chosen 90% of the vectors are assigned to the training group and the remaining 10% are assigned to the test group.[1] An L-dimensional (where L is the length of each feature vector) Gaussian mixture model is then parameterized based on the *training* group. Subsequently, the predicted audio classes of the *test* group are compared to the actual audio classes to derive the classification performance. The controlled random division between training and test groups was performed 10 times. The average classification performance of these 10 divisions is used as the overall classification performance of the current feature set and model.

When classifying a feature vector from the test group, a feature vector $\mathbf{x}$ of length L with $\mathbf{x} \in \Re^L$ falls into one of J classes. The solution entails a rule for predicting the class membership based on the L features. This solution is based on the statistical properties of the training data which comprise a set of feature vectors $\mathbf{x}_{n,j}$ with their corresponding class membership j. It is assumed that the features in each class follow a multivariate Gaussian distribution with each class having its own mean vector and its own covariance matrix. Then the probability density in class j, $p(\mathbf{x}_j)$, is given by

$$p(\mathbf{x}_j) = p(\mathbf{x}|j) = p(\mathbf{x},\mu_j,\mathbf{S}_j) = \frac{1}{\sqrt{(2\pi)^L|\mathbf{S_j}|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu_\mathbf{j})^T\mathbf{S}_j^{-1}(\mathbf{x}-\mu_\mathbf{j})\right),$$

with $\mathbf{x}$ the feature vector, μ_j the average value for class j, which can be estimated from the training data:

$$\hat{\mu}_j = \frac{1}{N_j}\sum_{n=1}^{N_j}\mathbf{x}_{n,j},$$

[1]This method and the 90%-10% split of training and test data was used in an earlier study on music genre classification [Tzanetakis et al., 2001]. It is not clear that this is the optimal division of training and test data but we have not yet evaluated the effect of using different split sizes.

where N_j is the number of training vectors belonging to class j and $\mathbf{x}_{n,j}$ the n-th learning vector for class j. The within-class covariance matrix $\mathbf{S}_j$ can be estimated by

$$\hat{\mathbf{S}}_j = \frac{1}{N_j - 1} \sum_{n=1}^{N_j} (\mathbf{x}_{n,j} - \hat{\mu}_j)(\mathbf{x}_{n,j} - \hat{\mu}_j)^T.$$

Classification of a new feature vector $\mathbf{x}$ is based on maximizing the discriminant function δ_j across classes J. The discriminant function for class j is given by

$$\delta_j(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \mu_j)^T \mathbf{S}_j^{-1}(\mathbf{x} - \mu_j) - \frac{L}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{S}_j| + \ln p(j),$$

which can be rewritten as

$$\delta_j(\mathbf{x}) = -\frac{1}{2}D(\mathbf{x}, \mu_j, \mathbf{S}_j) - \frac{1}{2}\ln(|\mathbf{S}_j|) + \ln(p(j)) - \frac{L}{2}\ln(2\pi),$$

where $D(..)$ denotes the *Mahalanobis distance* from $\mathbf{x}$ to μ_j, and is given by

$$D(\mathbf{x}, \mu_j, \mathbf{S}_j) = (\mathbf{x} - \mu_j)^T \mathbf{S}_j^{-1}(\mathbf{x} - \mu_j).$$

In addition to evaluating each feature set by its classification performance we also look at the discriminating power of individual features. To do this, we calculate the *Bhattacharyya distance* between classes based on single features. The Bhattacharyya distance $M(i, j)$ is a symmetric normalized distance measure between two centroids based on the centroid means and (co)variances [Papoulis, 1991]:

$$M(i,j) = M(j,i) = \frac{1}{8}(\mu_i - \mu_j)^T \left(\frac{\mathbf{S}_i + \mathbf{S}_j}{2}\right)^{-1} (\mu_i - \mu_j) + \frac{1}{2}\ln\left(\frac{|\frac{1}{2}(\mathbf{S}_i + \mathbf{S}_j)|}{\sqrt{|\mathbf{S}_i|}\sqrt{|\mathbf{S}_j|}}\right).$$

A high Bhattacharyya distance for a particular feature means that the centroids are well separable along that feature (dimension). An interesting property of the Bhattacharyya distance is *additivity*: the distances between two joint distributions of statistically independent random variables equals the sum of the marginal distances. Thus, *as long as features are statistically independent*, the distance of a certain subset can directly be calculated by summing marginal Bhattacharyya distances.

Although the size of the feature sets differ, we performed classification using the same number of features from each set. We chose the best 9 features from each set following an iterative ranking procedure. First the feature space was reduced to one feature and for each feature, the overall misclassification rate was estimated based on Gaussian data assumptions. The expected misclassification rate between classes i and j, ε_{ij} is given by

$$\varepsilon_{ij} = p(i)p(error|i) + p(j)p(error|j).$$

The likelihood ratio $K(\mathbf{x})$ at position $\mathbf{x}$ is given by

$$K(\mathbf{x}) = \frac{p(\mathbf{x}|i)}{p(\mathbf{x}|j)}.$$

The decision boundary for classification is set at $K(\mathbf{x}) = 1$. If the class centroids μ_j are assumed to be known, and if the shape of the class clusters is normal and can be characterized by their covariance matrices $\mathbf{S}_j$, the expected error rates can be calculated. Following [Fukunaga, 1972], the upper bound (Chernoff bound) of a misclassification error between classes i and j, ε_{ij}, is then given by

$$\varepsilon_{ij} \leq \sqrt{p(i)p(j)}\exp(-M_{ij}),$$

with M_{ij} the Bhattacharyya distance between class i and j. The upper bound for the total misclassification rate ε is given by

$$\varepsilon \leq \sum_{i}^{c}\sum_{j>i}^{c}\varepsilon_{ij}.$$

The feature which gave the lowest estimated misclassification rate was ranked as the top feature. Next the same process was performed using a two-feature space that included the top-ranked feature and one of the remaining features. The feature that gave, along with the top-ranked feature the lowest estimate of misclassification was ranked second. This process was repeated until all features were ranked (note that this method does not guarantee that the optimal combination is found since the search method may result in order effects). The top nine features of each set were chosen and used for the classification results described below.

6.2.1 Features

It has been reported, for speech-music discrimination, that the 2nd-order statistics of features (over time) are better features for classification than the features themselves [Scheirer & Slaney, 1997]. Here we carry the temporal analysis one step further and include a parameterized analysis of the features' temporal fluctuations. To do this we subdivide the audio frame into 1024-sample (23-ms) subframes with a 512-sample overlap, calculate feature values for each subframe and take the fast Fourier transform (FFT) on the array of subsequent feature calculations. Next the power spectrum is calculated and normalized by the DC value to reduce correlations. Finally the frequency axis is summarized by summing the energy in four frequency bands: 1) 0 Hz (average across observations), 2) 1-2 Hz (on the order of musical beat rates), 3) 3-15 Hz (on the order of speech syllabic rates), and 4) 20-43 Hz (in the range of modulations contributing to perceptual roughness).

Low-level signal parameters. This feature set, based on standard low-level (SLL) signal parameters, includes: (1) root-mean-square (RMS) level, (2) spectral centroid, (3) bandwidth, (4) zero-crossing rate, (5) spectral roll-off frequency, (6) band energy ratio, (7) delta spectrum magnitude, (8) pitch, and (9) pitch strength. This set of features is based on a recent paper by Li et al. [2001].

The final SLL feature vector consists of 36 features:

1-9: DC values of the SLL feature set
10-18: 1-2 Hz modulation energy of the SLL feature set
19-27: 3-15 Hz modulation energy of the SLL feature set
28-36: 20-43 Hz modulation energy of the SLL feature set

MFCC. The second feature set is based on the MFCCs [Slaney, 1998]. Mel-frequency cepstrum coefficients represent a parameterized description of the (frequency-warped) power spectrum. They are often used in automatic speech recognizers due to the following properties:

- Compactness: they are able to represent the spectral envelope with only a few parameters.
- Independence: the cepstral coefficients are approximately uncorrelated for speech signals [Hermansky & Malayath, 1998] and music [Logan, 2000]. The discrete cosine transform (DCT) used in the MFCC algorithm is a good approximation of the optimal diagonalizing principle-component analysis (PCA) transform.
- Gain independence: except for the zeroth MFCC coefficient, which is a function of the overall signal level, the remaining MFCC coefficients do not depend on the input level.

The full MFCC feature vector consists of 52 features:

1-13: DC values of the MFCC coefficients
14-26: 1-2 Hz modulation energy of the MFCC coefficients
27-39: 3-15 Hz modulation energy of the MFCC coefficients
40-52: 20-43 Hz modulation energy of the MFCC coefficients

Psychoacoustic features. The third feature set is based on estimates of the percepts roughness, loudness and sharpness. Roughness is the perception of temporal envelope modulations in the range of about 20-150 Hz and is maximal for modulations near 70 Hz. Loudness is the sensation of intensity and sharpness is a perception related to the spectral density and the relative strength of high-frequency energy. For loudness and sharpness, we characterize the temporal behavior in the same manner as for the SLL and MFCC feature sets. The estimate of roughness, however, is not treated the same way. Because roughness is based on mid-frequency temporal envelope modulations, an accurate

estimate can only be obtained for relatively long audio frames ($>\sim$ 180 msec). Thus, the temporal variation of roughness within an audio frame is represented by its mean and standard deviation over subframes of length $N_s = 8192$ (186 msec) with a hopsize of 4096.

Roughness. Our model for roughness is based on those of Zwicker & Fastl [1999a] and Daniel & Weber [1997]. First we filter each frame of audio by a bank of gammatone filters [Patterson et al., 1995], bandpass filters based on the effective frequency analysis of the ear, which are spaced logarithmically between 125 and 10 kHz. Next, the temporal (Hilbert) envelope of each filter output is calculated by taking the FFT, setting the negative frequency components to zero, multiplying the positive frequency components by 2, taking the inverse FFT and finally the absolute value. A correlation factor is then calculated for each filter based on the correlation of its output with that from two filters above and below it in the filter bank. This measure was introduced to decrease the estimated roughness of bandpass noise. The roughness estimate is then calculated by filtering the power in each filter output with a set of bandpass filters (centered near 70 Hz) that pass only those modulation frequencies relevant to the perception of roughness [Zwicker & Fastl, 1999a], multiplying by the correlation factor and then summing across frequency and across the filter bank.

Loudness. The loudness model is loosely based on the work of Zwicker & Fastl [1999b]. Here we assume that the maximum allowed sample value in the digital representation of the audio file corresponds to 96 dB SPL and we estimate the loudness level in sones. First, the power spectrum of the input frame is calculated and then normalized by subtracting (in dB) an approximation of the absolute threshold of hearing. This normalized power spectrum is then filtered by a bank of gammatone filters and summed across frequency to yield the power in each auditory filter, which corresponds to the internal excitation as a function of frequency. These excitations are then compressed, scaled and summed across filters to arrive at the loudness estimate.

Sharpness. The psychoacoustic percept of sharpness is based primarily on the relative strength of high-frequency components [von Bismarck, 1974]. It is estimated here using an algorithm almost identical to that of loudness with the only differences being a weight applied to each filter before the final summation and an additional normalization factor. The weights are larger for filters at higher center frequencies and were optimized to fit the psychoacoustic data on sharpness [von Bismarck, 1974; Zwicker & Fastl, 1999c].

The final psychoacoustic (PA) feature vector consists of 10 features:

1: average roughness
2: standard deviation of roughness
3: average loudness
4: average sharpness
5: 1-2 Hz loudness modulation energy
6: 1-2 Hz sharpness modulation energy
7: 3-15 Hz loudness modulation energy
8: 3-15 Hz sharpness modulation energy
9: 20-43 Hz loudness modulation energy
10: 20-43 Hz sharpness modulation energy

Auditory filterbank temporal envelopes. The fourth feature set is based on a model representation of temporal envelope processing by the human auditory system. Each audio frame is processed in two stages: (1) it is passed through a bank of gammatone filters, as in the PA feature set, which represent the spectral resolution of the peripheral auditory system and (2) a temporal analysis is performed by computing the modulation spectrum of the envelope (computed as in the roughness feature) of each filter output. In this implementation the filterbank includes every other critical band filter from 260-9795 Hz. Because the temporal analysis is performed directly on the entire 32768-sample frame we do not need to subdivide it into sub-frames as with the other features. The other features consist of only one value per audio frame and thus in order to evaluate their temporal behavior within a single frame, their values must be computed on a subframe basis. An advantage of being able to perform the temporal analysis directly at the level of the audio frame is that higher frequencies (up to the Nyquist frequency of the sampling rate) can be represented. After computing the envelope modulation spectrum for each auditory filter it is normalized by the average value (DC) and, parameterized by summing the energy in four frequency bands and taking the log: 0 Hz (DC), 3-15 Hz, 20-150 Hz, and 150-1000 Hz. The parameterized summary of high-frequency modulations is not calculated for some low-frequency critical band filters: a frequency band summary value is only computed for a critical band filter if the filter's center frequency is greater than the maximum frequency of the band. This process yields 62 features describing the auditory filterbank temporal envelopes (AFTE):

1-18: DC envelope values of filters 1-18
19-36: 3-15 Hz envelope modulation energy of filters 1-18
37-52: 20-150 Hz envelope modulation energy of filters 3-18
53-62: 150-1000 Hz envelope modulation energy of filters 9-18

6.3 Results

6.3.1 Standard low-level features (SLL)

The results for the standard low-level feature set are shown in Figure 6.1. The left panel shows the confusion matrix using the best 9 features of the SLL feature set. Classification performance is best for crowd noise with 99% correct classification and second best for classical music with 96% correct classification. Popular music is correctly classified in 80% of the cases, while in 20% of the cases it is classified as speech. Detection of background noise is not good (46% correct). It is often misclassified as classical music (28%) or crowd noise (21%). The overall classification accuracy is 82%.

The right panel shows the Bhattacharyya distance between all classes based on single features. Features 5 (spectral rolloff frequency) and 6 (band-energy ratio), and their second-order statistics (features 14, 15, 23, 24, 32, 33) show discriminative power between classical music and other classes. Furthermore, the 2-3 Hz and 3-15 Hz modulation energies of most features (feature numbers 19-27) contribute to discrimination between speech and background noise and between speech and crowd noise. Consistent with the confusion between speech and popular music in the classification results, no features show strong discrimination between popular music and speech. Only features 19 (3-15 Hz modulation energy of the signal RMS) and 28 (20-43 Hz modulation energy of the RMS) show some discriminative power. The small numbers above the x-axis at the right panel indicate the rank of the best 9 features.

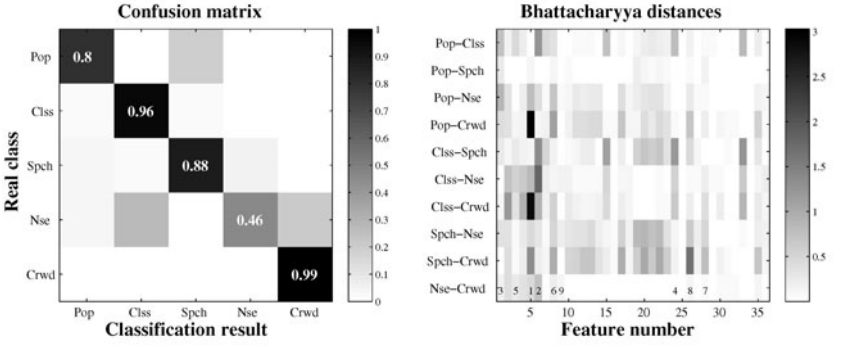

Figure 6.1. *Standard low-level features*: Classification performance (left) and feature discrimination power, i.e., distance between classes as a function of feature (right). The numbers above the x-axis indicate the rank of the best 9 features.

6.3.2 Mel-frequency cepstral coefficients (MFCC)

Figure 6.2 shows the results for the MFCC feature set. The format of the figure is the same as Figure 6.1: the confusion matrix of classification using the best 9 features is shown in the left panel and the Bhattacharyya distances between classes based on single features are shown in the right panel. The overall classification accuracy using the best 9 MFCC features is 85%, which is better than the SLL feature set. However, some of the individual audio classes show worse classification accuracy. For example, classical music is correctly identified in 90% of all cases, compared to 96% for the SLL feature set. Furthermore, crowd noise is correctly recognized in 86% of the cases, compared to 99% for the SLL feature set. Classification of background noise shows a large increase in performance, at 75% for the MFCC feature set compared to only 46% for the SLL feature set.

The Bhattacharyya distances in the right panel show that the second MFCC feature, which is the 2nd discrete cosine transform coefficient of the input spectrum, is a powerful feature, especially for discriminating crowd noise from other classes. This feature can be interpreted as the relative levels of low- and high-frequency energy in the signal. Features 6-13, which describe the input spectrum at a fine detail level, do not contribute to the classification process. On the other hand, second-order statistics of the first few MFCCs contribute to the discrimination between various classes. As with SLL features, discrimination between popular music and speech and between background noise and crowd noise is poor. This is consistent with the low Bhattacharyya distances between those classes.

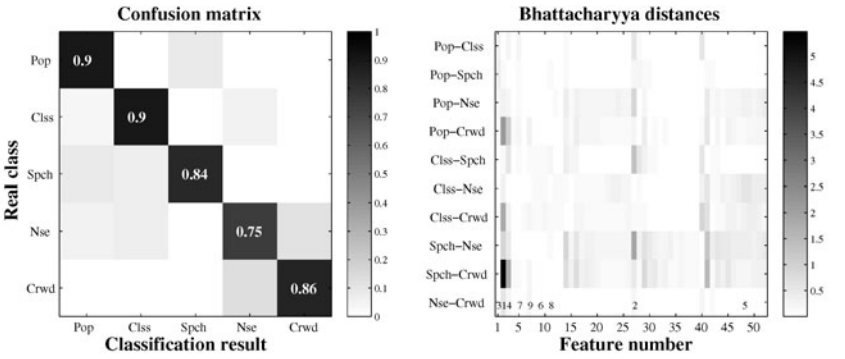

Figure 6.2. *Mel-frequency cepstral coefficients (MFCC)*: Classification performance (left) and feature discrimination power, i.e., distance between classes as a function of feature (right). The numbers above the x-axis indicate the rank of the best 9 features.

6.3.3 Psychoacoustic (PA) feature set

The results for the PA feature set are shown in Figure 6.3. The overall classification accuracy of this feature set is 84%. The confusion matrix shows that most classes were classified with an accuracy between 72 and 88% correct, with the exception of crowd noise which was classified correctly in 100% of the cases. The features that best discriminate between the classes are the 3-15 Hz modulation energy of the sharpness (feature 8), the average sharpness (feature 4), the average roughness (feature 1), the average loudness (feature 3) and the 3-15 Hz loudness modulation energy (feature 7). The panel of Bhattacharyya distances for individual class contrasts shows that feature 8 (3-15 Hz modulation energy of the sharpness) is key in the discrimination of speech from crowd noise, background noise and classical music. In addition, the average sharpness (feature 4) provides a relatively large distance between classical music and crowd noise.

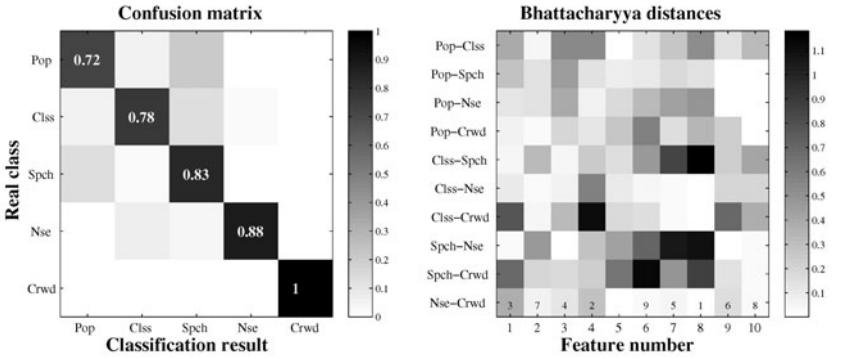

Figure 6.3. *Psychoacoustic features*: Classification performance (left) and feature discrimination power, i.e., distance between classes as a function of feature (right). The numbers above the x-axis indicate the rank of the best 9 features.

6.3.4 AFTE feature set

The feature analysis results for the auditory filter temporal envelope modulation feature vector are shown in Figure 6.4. The layout of the figure is the same as previous figures. The overall classification accuracy using the best 9 features is high (90%). Crowd noise is detected correctly in all cases and background noise and popular music are detected quite accurately (91%). Speech and classical music have lower scores (85% and 83%, respectively) and are both sometimes misclassified as popular music. Low and high Bhattacharyya distances (right panel) are somewhat scattered across features and audio classes, however there is a clear maximum for features 1-5 (steady-state

values of the auditory filters 1-5 centered at 260-376 Hz) for the discrimination between popular music and crowd noise. Other than that, no other individual feature sticks out as a powerful discriminator; the high performance of the AFTE feature set is due to a combination of features.

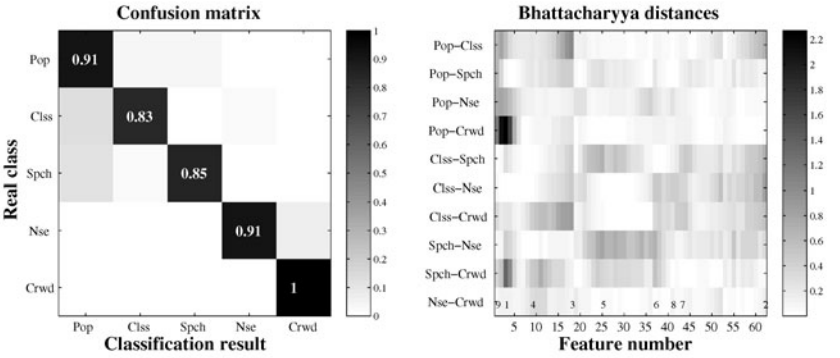

Figure 6.4. *Auditory filterbank temporal envelope*: Classification performance (left) and feature discrimination power, i.e., distance between classes as a function of feature (right). The numbers above the x-axis indicate the rank of the best 9 features.

6.3.5 Results summary

The results are summarized in Table 6.2. A comparison across all feature sets shows that, overall, the AFTE feature set are the most powerful for classification with our audio classes. For some individual classes, however, other feature sets perform slightly better: for classical music, the MFCC set performs the best with 90% classification; and for speech, the SLL set performs the best with 88% classification. Although it is not shown here, the performance of the AFTE feature set increases as more features are included. With the best 20 features, average classification performance increases to 95% with a 95% classification accuracy for classical music as well.

In comparing Bhattacharyya distances across features it is important to note that they are not normalized across feature sets. The MFCC set gives the single largest Bhattacharyya distance, 5.5 between the speech and crowd noise classes. Despite this large distance, the MFCC feature set is not the best basis for classifying speech or crowd noise: speech is often confused with popular and classical music and crowd noise is often confused with noise (see left panel of Figure 6.2). The PA and AFTE feature sets, on the other hand, give the lowest maximum Bhattacharyya distances at 1.2 and 2.2 respectively, but they are not the worst feature set overall. This combination of low Bhattacharyya distances and high classification performance may be due to a high correlation

Table 6.2. *Classification Results Summary.* Each entry gives the percent correct classification for the given audio class (top row) and feature set (left column). The right column shows, for each feature set, the average percent correct across all classes.

feature set	*popular music*	*classical music*	*speech*	*noise*	*crowd noise*	*average*
SLL	80%	96%	88%	46%	99%	82%
MFCC	90%	90%	84%	75%	86%	85%
PA	72%	78%	83%	88%	100%	84%
AFTE	91%	83%	85%	91%	100%	90%

between features and/or a better distribution of distances across features and audio classes.

6.4 Discussion

One can see from the ranking of the top nine features (see right panels of Figures 6.1–6.4) that temporal variations of the basic features are important for classification. In all cases, there are at least a few features in the top nine that incorporate temporal modulations. In addition, although we don't show the results here, performance of the SLL feature set is reduced to 71% overall if only the average values (DC) of the feature set are used for classification.

Our assumption of Gaussian-shaped clusters in the feature space may not be valid. Based on reasonably favorable results, it appears that it is not a bad assumption but we have not analyzed the feature space to the point where we can quantitatively evaluate this assumption. Classification performance could be further improved by such an analysis followed by the incorporation of perhaps more appropriate probability density functions.

Further improvements in classification performance could also come from changes to the classifier. For example, it is possible that sequential classification using fewer classes at each stage (i.e. grouping several classes initially) could result in improved performance. One could use different features, perhaps based on the Bhattacharyya distances between classes, for each sequential stage. In addition, as more powerful features for class discrimination are developed, different classification schemes (self-organizing maps, neural networks, k-nearest neighbor schemes and hidden Markov models) may begin to show differences in performance.

Finally, combinations of the best features from each set could also lead to improvements in classification performance. One could rank the features

across sets in the same manner that we rank features within each feature set, and then choose the combination that yields the best performance.

6.5 Conclusions

We have shown that audio classification can be improved by developing and working with improved audio features. Our comparison of current feature sets for this purpose shows that, overall, the AFTE feature set is the most powerful. However, for classifying particular audio classes, namely classical music and speech, the SLL feature set performs best.

From our ranking of features we have also shown that temporal variations in features are important for audio class discrimination. In all of our feature sets, the nine top-ranked features include at least two features representing temporal fluctuations.

Finally, we have seen that the Bhattacharyya distance can be a useful measure for determining the power of a particular feature. However a high Bhattacharyya distance between two clusters does not necessarily guarantee good classification performance for those cluster classes. In order to better relate Bhattacharyya distance and classification performance, one must look at correlations between features and at the entire feature vs. distance space (right panels of Figures 6.1–6.4).

Future work will involve the development of new features, further analysis of the feature space to test the Gaussian assumption, examination of alternative classification schemes, and the incorporation of more audio classes.

Acknowledgments

The authors would like to thank Armin Kohlrausch of Philips Research for helpful comments on this manuscript and Nick de Jong and Fabio Vignoli of Philips Research for their assistance in building the audio database.

References

Bismarck, G. von [1974]. Sharpness as an attribute of the timbre of steady sounds. *Acustica*, 30:159–172.

Daniel, P., and R. Weber [1997]. Psychoacoustical roughness: Implementation of an optimized model. *Acustica·Acta Acustica*, 83:113–123.

Davis, S.B., and P. Mermelstein [1980]. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-28:357–366.

Duda, R.O., and P.E. Hart [1973]. *Pattern classification and scene analysis*. Wiley, New York.

Foote, J. [1997]. A similarity measure for automatic audio classification. In *Proc. AAAI 1997 Spring Symposium on Intelligent Integration and Use of Text, Image, Video, and Audio Corpora.*

Fukunaga, K. [1972]. *Introduction to Statistical Pattern Recognition*. Academic press, New York, London.

Glasberg, B.R., and B.C.J. Moore [1990]. Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, 47:103–138.

Golub, S. [2000]. *Classifying Recorded Music*. Master's thesis, University of Edinburgh. http://www.aigeek.com/aimsc/.

Hermansky, H., and N. Malayath [1998]. Spectral basis functions from discriminant analysis. In *International Conference on Spoken Language Processing*.

Li, D., I.K. Sethi, N. Dimitrova, and T. McGee [2001]. Classification of general audio data for content-based retrieval. *Pattern Recognition Letters*, 5:533–544.

Logan, B. [2000]. Mel frequency cepstral coefficients for music modeling. In *International Symposium on Music Information Retrieval.*

Lu, G., and T. Hankinson [1998]. A technique towards automatic audio classification and retrieval. In *4th Int. Conference on Signal Processing*, Beijing.

Naphade, M.R., and T.S. Huang [2000a]. A probabilistic framework for semantic indexing and retrieval in video. In *IEEE International Conference on Multimedia and Expo (I)*, pages 475–478.

Naphade, M.R., and T.S. Huang [2000b]. Stochastic modeling of soundtrack for efficient segmentation and indexing of video. In *Proc. SPIE, Storage and Retrieval for Media Databases*, San Jose, CA, pages 168–176.

Papoulis, A. [1991]. *Probability, Random Variables and Stochastic Processes*. McGraw-Hill series in electrical engineering. McGraw-Hill, New York.

Patterson, R.D., M.H. Allerhand, and C. Giguere [1995]. Time domain modeling of peripheral auditory processing: A modular architecture and software platform. *J. Acoust. Soc. Am.*, 98:1890–1894.

Scheirer, E., and M. Slaney [1997]. Construction and evaluation of a robust multifeature speech/music discriminator. In *Proc. ICASSP*, Munich, Germany, pages 1331–1334.

Scheirer, E.D. [1998]. Tempo and beat analysis of acoustical musical signals. *J. Acoust. Soc. Am.*, 103:588–601.

Slaney, M. [1998]. *Auditory Toolbox*. Technical Report 1998-010, Interval Research Corporation. http://rvl4.ecn.purdue.edu/ñalcolm/interval/1998-010/.

Spina, M.S., and V.W. Zue [1996]. Automatic transcription of general audio data: Preliminary analysis. In *Proc. 4th Int. Conf. on Spoken Language Processing*, Philadelphia, PA.

Spina, M.S., and V.W. Zue [1997]. Automatic transcription of general audio data: Effect of environment segmentation on phonetic recognition. In *Proceedings of Eurospeech*, Rhodes, Greece.

Toonen Dekkers, R.T.J., and R.M. Aarts [1995]. *On a Very Low-Cost Speech-Music Discriminator*. Technical Report 124/95, Nat.Lab. Technical Note.

Tzanetakis, G., G. Essl, and P. Cook [2001]. Automatic musical genre classification of audio signals. In *Proceedings International Symposium for Audio Information Retrieval (ISMIR)*, Princeton, NJ.

Wang, H., A. Divakaran, A. Vetro, S.F. Chang, and H. Sun, [2000a]. *Survey on Compressed-Domain Features used in Video/Audio Indexing and Analysis*. Technical report, Department of electrical engineering, Columbia University, New York.

Wang, Y., Z. Liu, and J.C. Huang [2000b]. Multimedia content analysis using both audio and visual cues. *IEEE Signal Processing Magazine*, 17:12–36.

Wold, E., T. Blum, D. Keislar, and J. Wheaton [1996]. Content-based classification, search, and retrieval of audio. *IEEE Multimedia*, Fall:27–36.

Zhang, M., K. Tan, and M.H. Er [1998]. Three-dimensional sound synthesis based on head-related transfer functions. *J. Audio. Eng. Soc.*, 146:836–844.

Zhang, T., and C.C.J. Kuo [2001]. Audio content analysis for online audiovisual data segmentation and classification. *IEEE Transactions on Speech and Audio Processing*, 9:441–457.

Zwicker, E., and H. Fastl [1999a]. *Psychoacoustics: Facts and Models*, volume 22 of *Springer series on information sciences*, chapter Roughness, pages 257–264. Springer-Verlag, Berlin, 2nd edition.

Zwicker, E., and H. Fastl [1999b]. *Psychoacoustics: Facts and models*, volume 22 of *Springer series on information sciences*, chapter Loudness, pages 203–238. Springer-Verlag, Berlin, 2nd edition.

Zwicker, E., and H. Fastl [1999c]. *Psychoacoustics: Facts and models*, volume 22 of *Springer series on information sciences*, chapter Sharpness and Sensory Pleasantness, pages 239–246. Springer-Verlag, Berlin, 2nd edition.