# Features for Audio Classification

**Chapter** · January 2004

DOI: 10.1007/978-94-017-0703-9



**2 authors:**

Jeroen Breebaart
Dolby Laboratories, Inc.
**101** PUBLICATIONS **1,829** CITATIONS
SEE PROFILE

Martin Franciscus McKinney
Starkey Hearing Technologies
**59** PUBLICATIONS **920** CITATIONS
SEE PROFILE

# FEATURES FOR AUDIO CLASSIFICATION

Jeroen Breebaart and Martin McKinney
Philips Research Laboratories, Prof. Holstlaan 4 (WY82), 5656 AA Eindhoven, The Netherlands

email: {jeroen.breebaart, martin.mckinney}@philips.com

**Abstract**

Four audio feature sets are evaluated in their ability to differentiate five audio classes: popular music, classical music, speech, 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.

## 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. 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 [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11]. 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. These different stages will be discussed in more detail below.

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 [1, 2, 4, 6, 8, 9]. 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 [12] and music [13]. 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

[14, 4, 15]. The performance of classification schemes usually decreases if more audio classes are present [7, 10]. 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 [16, 17, 18, 11].

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 neighbour schemes and hidden Markov models. Some authors have found that the classification scheme does not influence the classification accuracy [4, 19], 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 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.

## 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, a Gaussian-based quadratic discriminant analysis (QDA) [20]. The feature sets (described below) are: (1) low-level signal properties; (2) MFCC; (3) psychoacoustic features including roughness, loudness and sharpness; and (4) an auditory model representation of temporal envelope fluctuations. The audio database consists of five general classes of audio: classical music, popular music (all styles but classical), speech (male and female, English, Dutch, German and French), crowd noise (applauding and cheering), and noise (including traffic, fan, restaurant, nature, etc. noises). The number of files in each class is given in Table 1.

| **Class Name** | Popular Music | Classical Music | Speech | Noise | Crowd Noise |
|---|---|---|---|---|---|
| **Number of Files** | 175 | 35 | 31 | 25 | 31 |

Table 1: Audio database by class

The classification process begins with the extraction of a set of features, i.e., feature vectors, from each sound file. Features are calculated on 10 consecutive 32768-sample frames (44.1 kHz sampling rate) with a hop-size of 24576. Thus, each audio file is represented by 10 feature vectors and these vectors are grouped into classes based on the type of audio. The 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 $N$-dimensional (where $N$ is the length of each feature vector) Gaussian mixture model is then parameterized based on the *training* group, assuming that each audio class comprises a

[1]This method and the 90%-10% split of training and test data was used in an earlier study on music genre classification [21]. 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.

single centroid with its own mean and variance. Each file in the *test* group is then classified based on its feature vectors using Bayesian theory to find which class centroid it most probably belongs. Details of this classification method, QDA, can be found in [20]. This 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.

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 is a symmetric normalized distance measure between two centroids based on the centroid means and (co)variances [22]. A high Bhattacharyya distance for a particular feature means that the centroids are well separable along that feature (dimension).

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 (by calculating the Chernoff bound) from the Bhattacharyya distances between the classes [22]. The feature which gave the lowest estimate 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.

## 2.1 Features

Features are calculated on 32768-sample frames of audio. 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 [4]. 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 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) 20-150 Hz (in the range of modulations contributing to perceptual roughness).

The two new feature sets introduced in Secs. 2.1.3 and 2.1.4 are based on models of human auditory processing. Each begins with a bank of bandpass filters which represent the frequency resolution of the peripheral human auditory system. These filters, termed critical band filters, reflect the channeling property of the auditory system, i.e., signals that are passed through different critical bands are, to a large extent, processed independently [23].

### 2.1.1 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 Dongge Li, et al., at Philips Research Briarcliff [11]. See the paper for mathematical details.

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**: 20-150 Hz modulation energy of the SLL feature set

### 2.1.2 MFCC

The second feature set is based on the first 13 MFCCs [24]. The final 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**: 20-150 modulation energy of the MFCC coefficients

### 2.1.3 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 and Fastl [25] and Daniel and Weber [26]. First we filter each frame of audio by a bank of gammatone filters [27], 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 [25], 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 and Fastl [28]. Here we assume that an RMS value of 1 in the real value 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 [29]. 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 [29, 30].

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**: 20-150 Hz loudness modulation energy
**10**: 20-150 Hz sharpness modulation energy

### 2.1.4 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

# 3 Results

## 3.1 SLL feature set

The results for the standard low-level feature set are shown in Fig. 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 (20-150 Hz modulation energy of the RMS) show some discriminative power.

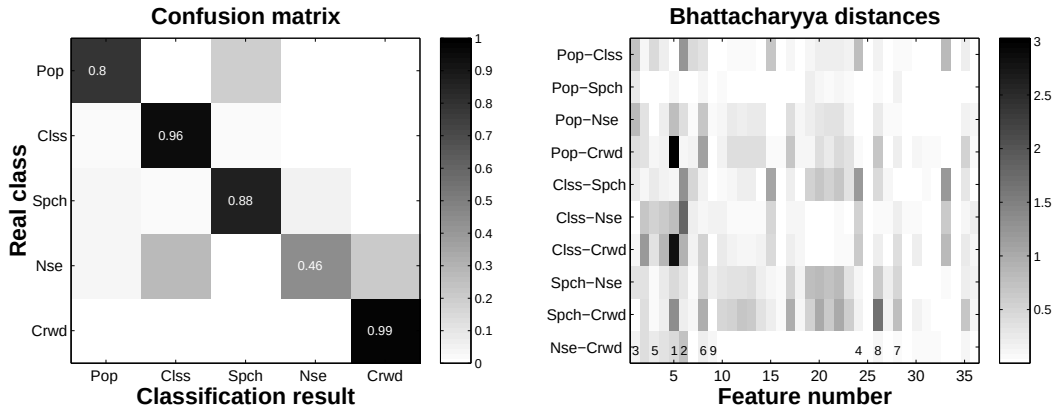


Figure 1: *Standard low-level features*: Classification performance (left) and feature discrimination power, i.e., distance between classes as a function of feature (right). The numbers above the x-axis indicate the rank of the best 9 features.

## 3.2 MFCC feature set

Figure 2 shows the results for the MFCC feature set. The format of the figure is the same as Fig. 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.

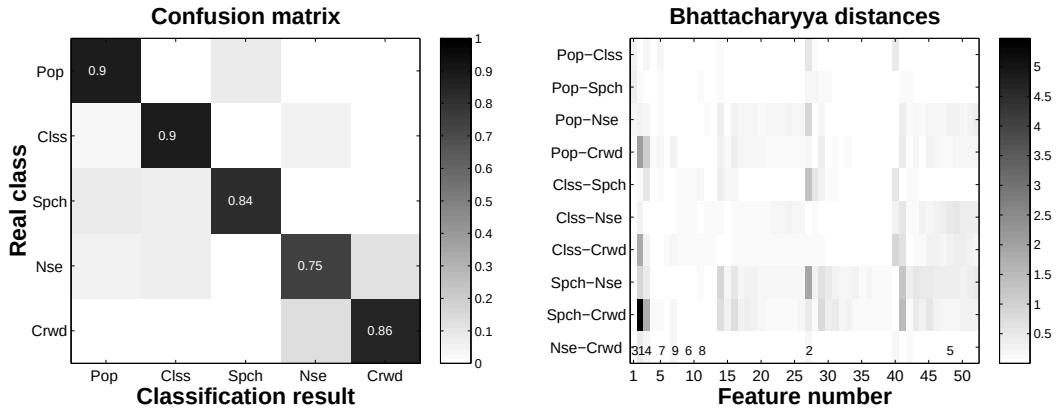


Figure 2: *Mel-frequency cepstral coefficients (MFCC)*: Classification performance (left) and feature discrimination power, i.e., distance between classes as a function of feature (right). The numbers above the x-axis indicate the rank of the best 9 features.

## 3.3 PA feature set

The results for the PA feature set are shown in Fig. 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 the best 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.

## 3.4 AFTE feature set

The feature analysis results for the auditory filter temporal envelope modulation feature vector are shown in Fig. 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

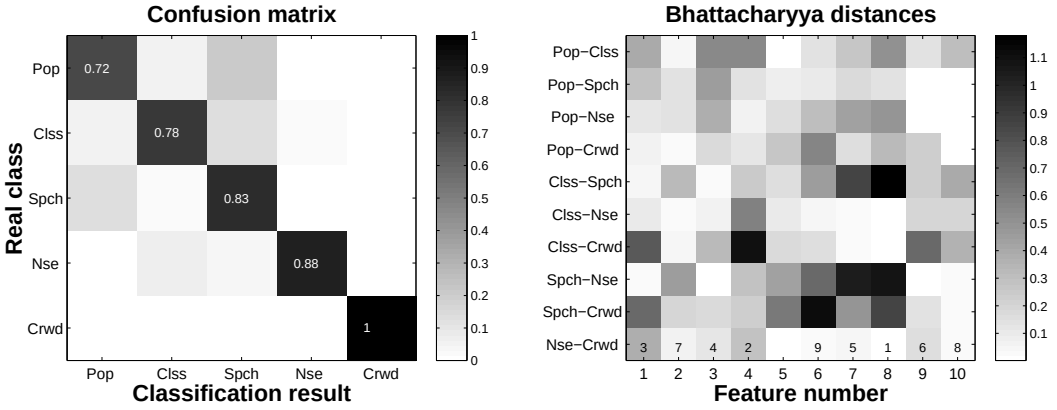


Figure 3: *Psychoacoustic features*: Classification performance (left) and feature discrimination power, i.e., distance between classes as a function of feature (right). The numbers above the x-axis indicate the rank of the best 9 features.

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 260-3760 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.

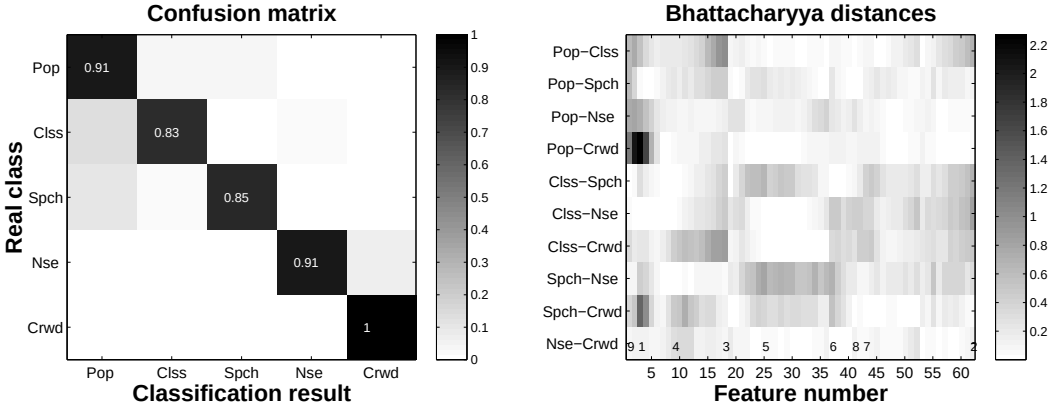


Figure 4: *Auditory filterbank temporal envelope*: Classification performance (left) and feature discrimination power, i.e., distance between classes as a function of feature (right). The numbers above the x-axis indicate the rank of the best 9 features.

The results are summarized in Table 2. A comparison across all feature sets shows that, overall, the AFTE features 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.

| 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% |

Table 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.

In comparing Bhattacharyya distances across feature sets it is important to note that they are not normalized across feature set. 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 Fig. 2). The PA and AFTE feature sets, on the other hand, gives 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 between features and/or a better distribution of distances across features and audio classes.

## 4 Discussion

One can see from the ranking of the top nine features (see right panels of Figs. 1-4) 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 features are used for classification.

The choice of a 32768-sample (743 msec) frame length was based on a finding of Spina and Zue [3]. 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 500 msec and then saturated (and decreased a little) with further increases. Thus, we chose the next power of 2 (for FFT performance reasons) larger than 500 msec. Whether this is the optimum frame size for all feature sets is not known but should be examined in future studies. Nonetheless, from a perceptual point of view, this is a remarkably short section of audio on which classification is performed. Improvements in classification performance could surely be made if classification were based on more than one audio frame.

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 neighbour 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.

# 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 SSL 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 Figs. 1-4).

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.

# 6 Acknowledgements

This work was supported by the CASSANDRA project and the "Learning Features for Classifying Audio" project. 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

[1] S. B. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on acoustics, speech and signal processing*, ASSP-28:357–366, 1980.

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

[3] M. S. Spina and V. W. Zue. Automatic transcription of general audio data: Preliminary analysis. In *Proc. 4th Int. Conf. on spoken language processing*, Philadelphia, PA, 1997.

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

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

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

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

[8] H. Wang, A. Divakaran, A. Vetro, S. F. Chang, and H. Sun. Survey on compressed-domain features used in video/audio indexing and analysis. Technical report, Department of electrical engineering, Columbia University, New York, 2000.

[9] Y. Wang, Z. Liu, and J. C. Huang. Multimedia content analysis using both audio and visual cues. *IEEE signal processing magazine*, 17:12–36, 2000.

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

[11] D. Li, I. K. Sethi, N. Dimitrova, and T. McGee. Classification of general audio data for content-based retrival. *Pattern recognition letters*, submitted:000–000, 2002.

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

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

[14] R. T. J. Toonen Dekkers and R. M. Aarts. On a very low-cost speech-music discriminator. Technical Report 124/95, Nat.Lab. Technical Note, 1995.

[15] G. Lu and T. Hankinson. A technique towards automatic audio classification and retrieval. In *4th int. conference on signal processing*, Beijing, 1998.

[16] J. Foote. 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*, 1997.

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

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

[19] Seth Golub. Classifying recorded music. Master's thesis, University of Edinburgh, Sep 2000. http://www.aigeek.com/aimsc/.

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

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

[22] A. Papoulis. *Probability, random variables and stochastic processes*. McGraw-Hill series in electrical engineering. McGraw-Hill, New York, 1991.

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

[24] Malcolm Slaney. Auditory toolbox. Technical Report 1998-010, Interval Research Corporation, 1998. http://rvl4.ecn.purdue.edu/m̃alcolm/interval/1998-010/.

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

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

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

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

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

[30] Eberhard Zwicker and Hugo Fastl. *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, 1999.