Features for Audio and Music Classification

Martin F. McKinney
Philips Research Laboratories
Prof. Holstlaan 4 (WY82)
5656 AA Eindhoven, The Netherlands
martin.mckinney@philips.com

Jeroen Breebaart
Philips Research Laboratories
Prof. Holstlaan 4 (WY82)
5656 AA Eindhoven, The Netherlands
jeroen.breebaart@philips.com

Abstract

Four audio feature sets are evaluated in their ability to classify five general audio classes and seven popular music genres. 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 both music and 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 provide information about the content. However, due to the fact that metadata are 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 and music (Wold et al., 1996; Spina & Zue, 1997; Scheirer & Slaney, 1997; Scheirer, 1998; Wang et al., 2000; Zhang & Kuo, 2001; Li et al., 2001; Tzanetakis & Cook, 2002).

Most audio classification systems combine two processing stages: feature extraction followed by classification. A variety of signal features have been used for this purpose, including low-level parameters such as the zero-crossing rate, signal bandwidth, spectral centroid, and signal energy. Another set of features used, inherited from automatic speech recognizers, is the set mel-frequency cepstral coefficients (MFCC). Typical performance of these feature sets in speech/music discrimination tasks is around 95% (Toonen Dekkers & Aarts, 1995; Scheirer & Slaney, 1997; Lu & Hankinson, 1998) but decreases as the number of audio classes increases (Zhang & Kuo, 1998, 2001).

There has also been some recent work on automatic music genre detection. Tzanetakis & Cook (2002) combine standard features with representations of rhythm and pitch content and show classification performance in the range of 60%.

Several different classification strategies have been employed in these studies, including multivariate Gaussian models, Gaussian mixture models, self-organizing maps, neural networks, k-nearest neighbor schemes and hidden Markov models. In some cases, the 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 findings is that, 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 and music. 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 the classification of a set of general audio classes and a set of popular music genres. We also examine how the characterization of features' temporal behavior can influence classification. The two new feature sets, described in detail below, are based on perceptual models of auditory processing.

2 Method

We compare four distinct feature extraction stages to evaluate their relative performance while using the same classifier stage, a Gaussian-based quadratic discriminant analysis (QDA) (Duda & Hart, 1973). 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 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 dif-

General Audio Class	Classical Music	Popular Music	Speech	Noise	Crowd Noise
Number of Files	35	188	31	25	31

Popular Music Class	Jazz	Folk	Electronica	R&B	Rock	Reggae	Vocal
Number of Files	38	23	27	43	37	11	9

Table 1: Audio database by class: number of audio files in each class.

ferent critical bands are, to a large extent, processed independently (Glasberg & Moore, 1990).

Previous studies have shown that, for speech-music discrimination, 2nd-order statistics of features (over time) provide a better basis for classification than the features themselves (Scheirer & Slaney, 1997). In addition, Peeters et al. (2002) showed that "dynamic features" provide a good basis for music summarization purposes. Here we apply this technique to audio and music genre classification and include parameterized analyses of features' temporal fluctuations as additional features.

Two types of classification were performed, one on a set of five general audio classes and a second on a set of popular music genres. The general audio classes were: classical music, popular music (non-classical genres), speech (male and female, English, Dutch, German and French), crowd noise (applauding and cheering), and noise (background noises including traffic, fan, restaurant, nature, etc. noises). The popular music class contained music from seven genres: Jazz, Folk, Electronica, R&B, Rock, Reggae, and Vocal. The database used in the current study is a "quintessential" subset of a larger database. Two subjects assigned each song a genre and rated the song as to how well it typifies the genre. Songs were selected for the quintessential database based on the overall rating. The number of files in each class is given in Table 1.

2.1 Features

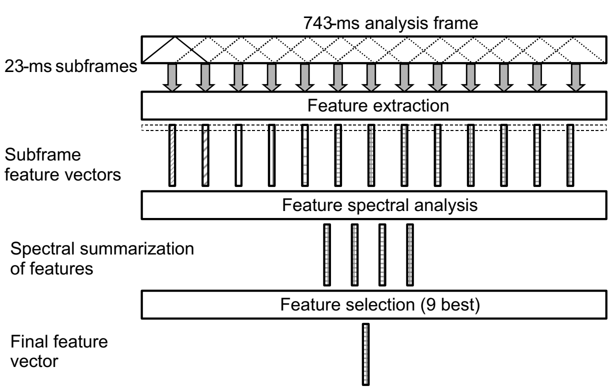

Figure 1: Feature extraction method

The feature extraction process, illustrated in Fig. 1, includes a summarized temporal analysis of features. Individual features are calculated from 23-msec half-overlapping subframes of audio. A power spectrum is then calculated for each feature, across 64 consecutive subframe values, resulting in an overall analysis frame of 743 msec. The power spectrum is normalized by the DC value and summarized by calculating the energy in four 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 lower range of modulations contributing to perceptual roughness). The top nine values (determined by a separate ranking procedure for each feature set - see below) of this spectral summarization were then selected and used in the classification process. This entire process was performed separately for each feature set.

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[1], and (9) pitch strength. This set of features is based on a recent paper by Li et al. (2001). [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: 20-43 Hz modulation energy of the SLL feature set

2.1.2 *MFCC*

The second feature set is based on the first 13 MFCCs (Slaney, 1998). 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: 20-43 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, maximal at 70 Hz, and is generally thought to be a primary component of musical dissonance (Plomp & Levelt, 1965; Terhardt, 1974). Loudness is the sensation of signal strength and sharpness is a perception related to the spectral density and the relative strength of high-frequency energy. Estimates of these percepts were calculated based on current models (Zwicker & Fastl, 1990; Daniel & Weber, 1997; Bismarck, 1974). Temporal analyses of loudness and sharpness were calculated using the subframe process described above. However, because roughness is based on mid-rate 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 186-msec subframes with 93-msec overlap.

[1]The term *pitch* is used here to describe an estimate of the pitch percept derived using an autocorrelation-based method (see Li et al., 2001, for details).

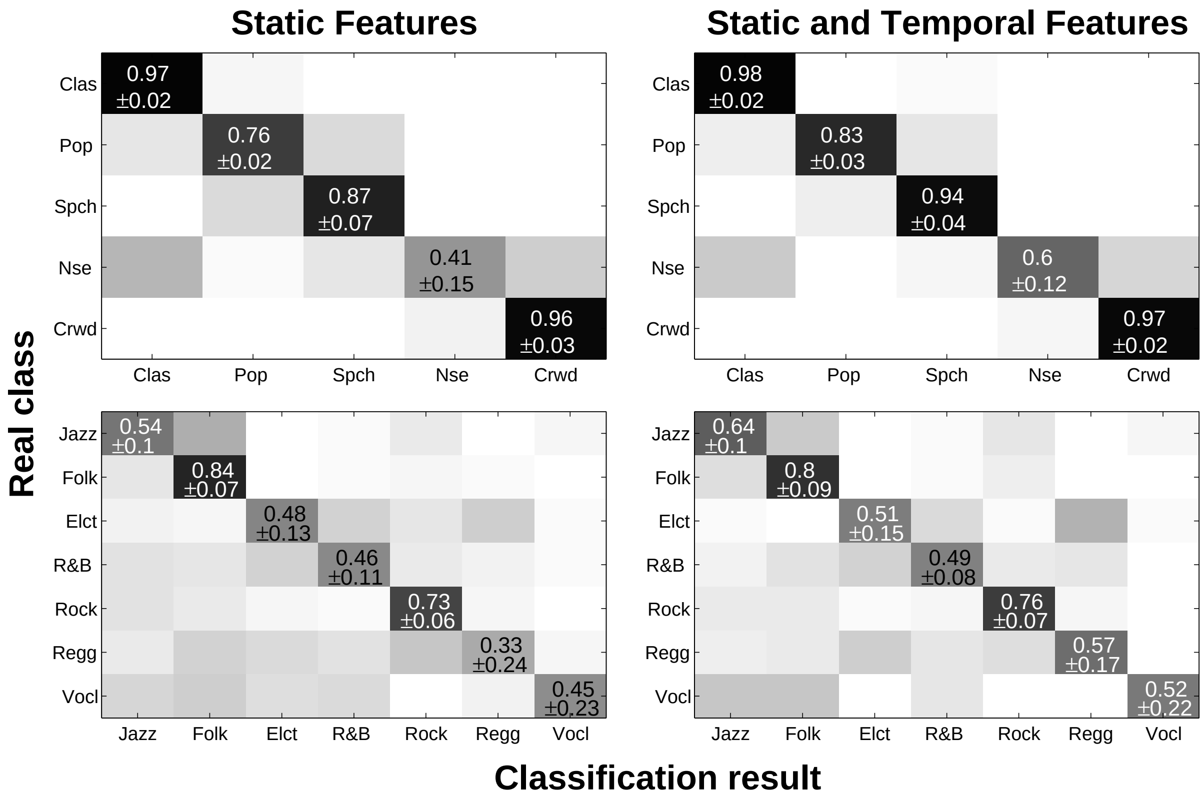

Figure 2: Classification performance using standard low-level features. Confusion matrices for classification based on the best 9 static features (left) and best 9 overall features, including static and temporal features (right). The top two panels show results for the general audio classes, the bottom panels for the music genre classes. The numbers in the boxes indicate the probability (± standard error) that the class on the left axis is classified as the class on the bottom axis. Where numbers are not present, the shade indicates the probability: white = 0, black = 1.

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

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 18 4th-order bandpass GammaTone filters (Glasberg & Moore, 1990; Hartmann, 1997, chap. 10) spaced logarithmically from 26 to 9795 Hz; and (2) the modulation spectrum of the temporal envelope is calculated for each filter output. The spectrum of each filter is then summarized by summing the energy in four bands: 0 Hz (DC), 3-15 Hz, 20-150 Hz, and 150-1000 Hz. The parameterized summary of high modulation rates is not calculated for some low-frequency filters: a modulation rate summary value is only computed for a critical band filter if the filter's center frequency is greater than the maximum rate 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

2.2 Classification

Classification of audio files was performed using quadratic discriminate analysis (see Duda & Hart, 1973), which provided better preliminary results than linear discriminate analysis. Features were calculated from each file on 10 consecutive 743-msec frames with a 558-msec hop-size. The feature vectors were grouped into classes based on the type of audio and were used to parameterize an N-dimensional Gaussian mixture model (one Gaussian with its own mean and variance for each class), where N is the length of the feature vector. Training and cross-validation were done using the *.632+ bootstrap* method, an improved version of the leave-one-out bootstrap (Efron & Tibshirani, 1997, 1993). This method has been shown to provide estimates of prediction error with less variance than standard k-fold cross-validation techniques, especially for small databases. Bootstrap replications were performed 500 times for each class. Classification was done per audio file and was assigned based on the majority of 10 consecutive audio frame classifications.

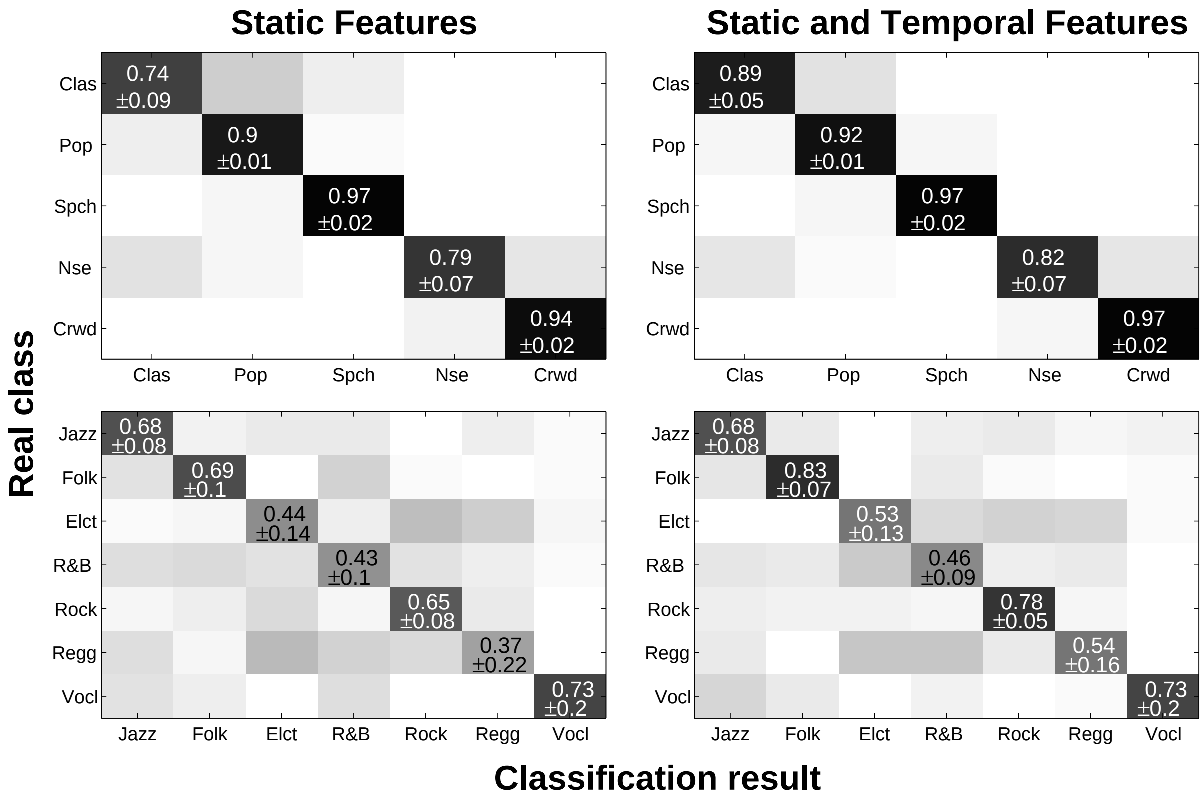

Figure 3: Classification performance using mel-frequency cepstral coefficients (MFCC): Same format as Fig. 2.

Although the size of the feature sets differ, we performed classification using the same number of features from each set. We chose the best nine[2] features from each set following an iterative ranking procedure: for each feature, we estimated classification error from the *Bhattacharyya distances* (see, e.g., Papoulis, 1991) between classes and designated the top ranked feature as that which gave the lowest error; we repeated this for 2, 3, ... 9 features.

Class Set	**Feature Type**	**Feature Rank**								
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
General Audio	**Static**	5	6	1	8	3	9	2	4	7
	Stat. & Temp.	5	6	1	24	8	3	28	26	9
Music Genre	**Static**	3	5	1	8	9	4	6	2	7
	Stat. & Temp.	24	5	1	4	8	6	3	19	28

Table 2: Feature ranking for the standard low level feature set. Feature numbers correspond to the features described in Sec. 2.1.1.

3 Results

3.1 SLL feature set

The ranking results for the SLL feature set are shown in Table 2. Different rankings are shown for each combination of audio-class set (general audio or music genre) and feature type (static or static-and-temporal). For general audio classification, features 5 (spectral roll-off frequency), 6 (band-energy ratio), and 1 (RMS level) rank the highest. When temporal features are included, features 24 (3-15 Hz modulations of band-energy ratio), 28 (20-43 Hz modulations of RMS level), and 26 (3-15 Hz modulations of pitch) are included in the top nine. For music genre classification, the top ranked features are slightly different: feature 3 (spectral bandwidth) is the top static feature and feature 19 (3-15 Hz modulations of RMS level) is included in the top nine. It is clear from these results that, when available, temporal modulations (over a range of rates) of features are important for classification.

Classification results for the SLL feature set are shown in Fig. 2.

Each panel shows a confusion matrix that indicates the probability ($\pm$ standard error) of each audio or music class (left axis) being classified as each class in the group (bottom axis). The top panels show results for general audio classification, the bottom panels for music genre classification; the left panels show results for classification based on static features; the right panels for classification based on both static and temporal features.

In general classification is better when temporal features are included. Although only one audio class (popular music) shows a significant improvement, there is only one class (folk music) for which performance decreased slightly. With temporal and static features, overall classification performance is $86 \pm 4\%$ for the general audio classes and $61 \pm 11\%$ for the music genres. For the general audio classes, classification is best for classical music ($98 \pm 2\%$) and worst for background noise ($60 \pm 12\%$), which is confused with crowd noise and classical music. For the music genres, classification is best for folk ($80 \pm 9\%$) and worst for R&B ($49 \pm 8\%$) which is confused with electronica and reggae.

[2] We limited all feature sets to their top nine features because the standard low-level feature set consists of only nine basic features.

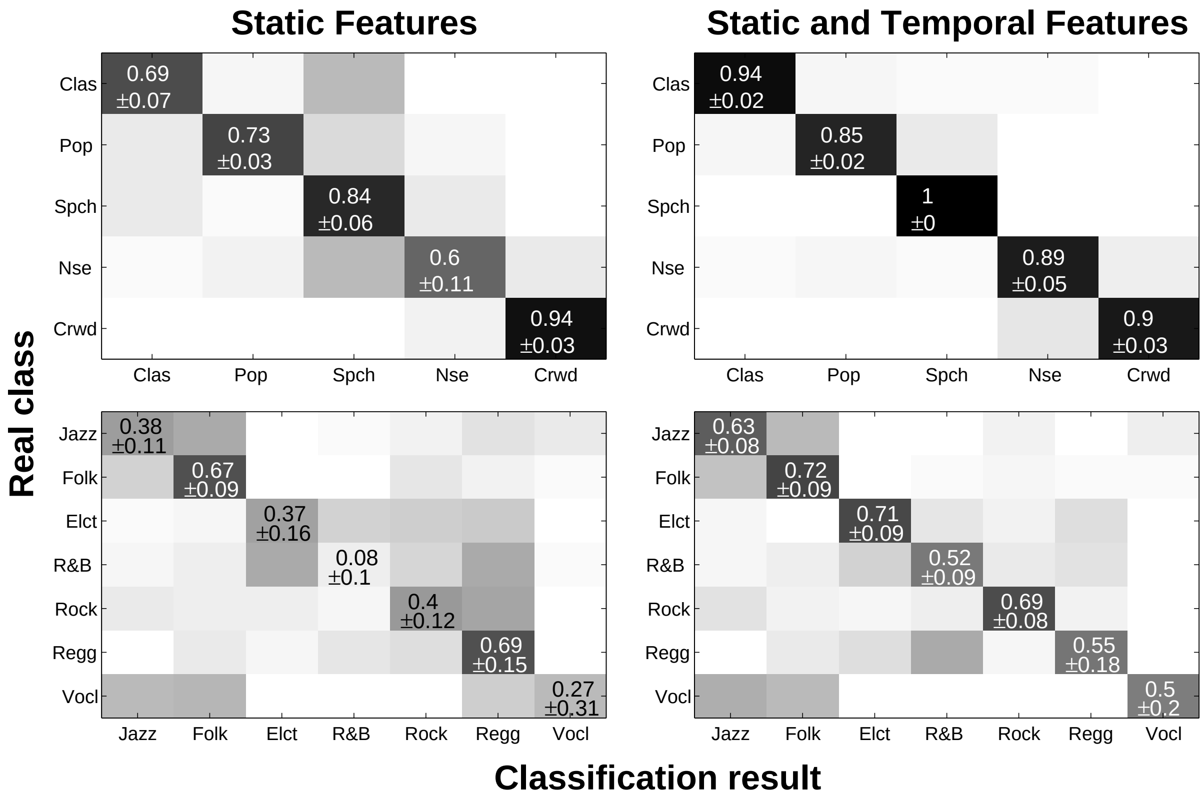

Figure 4: Classification performance using psychoacoustic features. Same format as previous figures. Results shown on the left panels were computed with only three static features: average loudness, roughness, and sharpness.

3.2 MFCC feature set

Table 3 shows the top ranked MFCC features for each audio-class set and feature-type combination. For general audio classification, the first three MFCCs are highly ranked. When temporal features are included, features 27 (3-15 Hz modulations of MFCC 1) and 49 (20-43 Hz modulations of MFCC 10) are included in the top nine. For classification of music genres, the rankings are slightly different, but the first few MFCCs also seem to be the most important. When temporal features are included, feature 40 (20-43 Hz modulations of MFCC 1) is ranked the highest. These results show that temporal modulations of MFCCs are also important for classification.

Class Set	**Feature Type**	**Feature Rank**								
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
General Audio	**Static**	2	3	1	9	11	5	12	10	7
	Stat. & Temp.	2	27	1	49	3	5	9	11	7
Music Genre	**Static**	1	4	2	6	3	7	11	9	13
	Stat. & Temp.	40	1	4	2	6	27	3	11	13

Table 3: Feature ranking for the MFCC feature set. Feature numbers correspond to the features described in Sec. 2.1.2.

Figure 3 shows the classification results for the MFCC feature set. As with the SLL feature set, classification is better when based on the top nine static-and-temporal features, rather than the top nine static features. Overall classification based on both feature types (right panels) is 92 ± 3% for the general audio classes and 65 ± 10% for the music genres. Overall performance is slightly better than the SLL feature set for general audio classification, largely helped by an increased ability to classify background noise and popular music. Classification of classical music is, however, worse for the MFCC feature set (89 ± 5% with MFCC vs. 98 ± 2% with SLL). Overall classification performance of music genres is not significantly better with MFCC features than SLL features (65 ± 10% with MFCC vs. 61 ± 11% with SLL), although classification performance of rock music is significantly increased with MFCC features and no music genres show a decrease in classification performance.

3.3 PA feature set

Class Set	**Feature Type**	**Feature Rank**								
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
General Audio	**Static**	4	1	3	·	·	·	·	·	·
	Stat. & Temp.	8	4	1	3	9	7	2	10	6
Music Genre	**Static**	4	1	3	·	·	·	·	·	·
	Stat. & Temp.	4	1	8	9	3	7	6	5	10

Table 4: Feature ranking for the psychoacoustic feature set. Feature numbers correspond to the features described in Sec. 2.1.3. Rankings 4-9 for static features are not applicable because there are only 3 static psychoacoustic features.

Table 4 shows the top ranked psychoacoustic features for each audio-class set and feature-type combination. Of the three static features, feature 4 (average sharpness) is ranked highest, followed by features 1 (average roughness) and 3 (average loudness) for both general audio and music genre classification. When temporal features are included, the top ranked features include 8 (3-15 Hz modulations of sharpness) and 9 (20-43 Hz

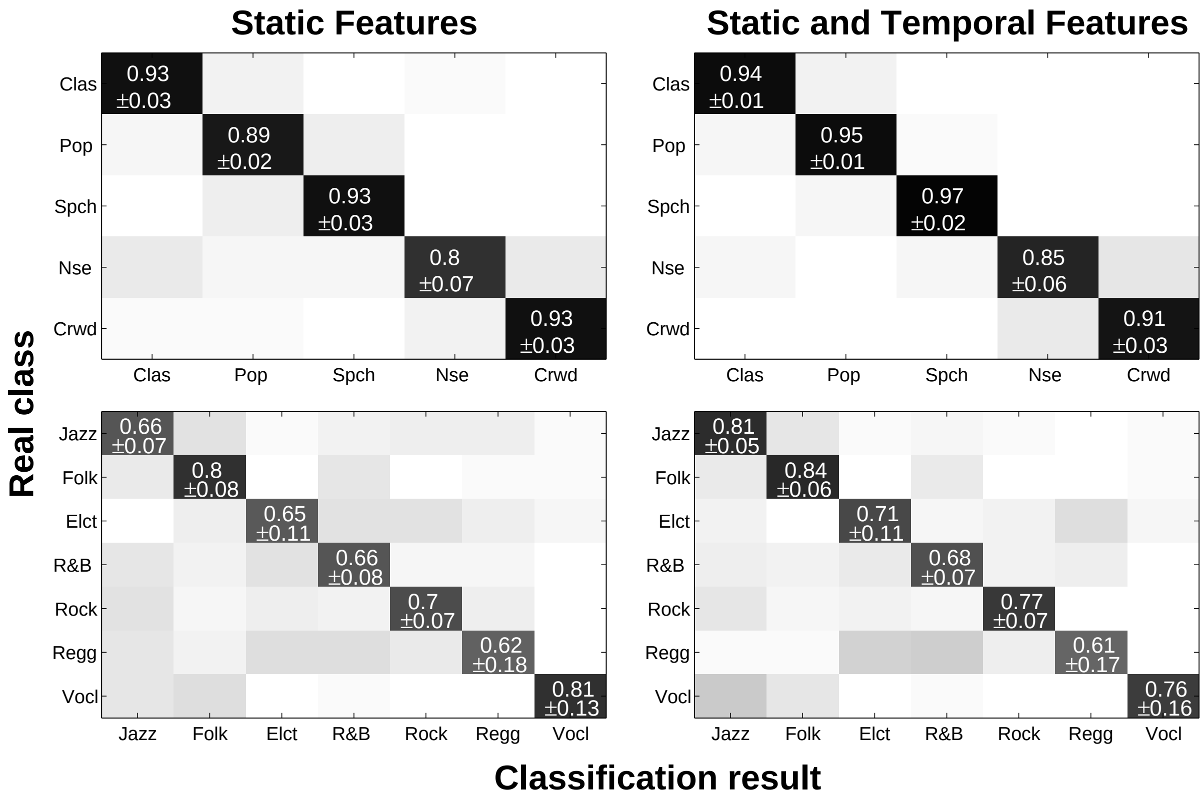

Figure 5: Classification performance using the auditory filterbank temporal envelope. Same format as previous figures.

modulations of loudness).

Classification results for the PA feature set are shown in Fig. 4. As with the previous feature sets, the inclusion of temporal features increases overall classification performance. (In this case, the number of features used for classification also increased.) The overall classification performance using both static and temporal features is $92 \pm 3\%$ for general audio classes and $62 \pm 10\%$ for music genres. These values are roughly the same as those for the MFCC feature set.

3.4 AFTE feature set

Table 5 shows the top ranked features from the set describing the auditory filter temporal envelope. The static features are the DC outputs (no modulations) of the 18 bandpass filters, which cover a range of center frequencies (260-9795 Hz). The top nine static features for both general audio and music genre classification span the range of center frequencies. When temporal features are included, several appear in the top nine rankings: 62 (150-1000 Hz modulations of filter 18), 25 (3-15 Hz modulations of filter 7), 41 (20-150 Hz modulations of filter 5), 52 (20-150 Hz modulations of filter 18), and 20 (3-15 Hz modulations of filter 2). As with all the other feature sets, temporal variations of features are important for classification.

Classification results for the AFTE feature vector are shown in Fig. 5. Classification based on temporal and static features is slightly better than on static features alone. Overall classification performance (right panels, $93 \pm 2\%$ for general audio classes and $74 \pm 9\%$ for music genres) is better than the other feature sets, although the improvement is significant only for the SLL feature set. For general audio classification, the most difficult classes are background noise ($85 \pm 6\%$) and crowd noise ($91 \pm 3\%$), which are confused with each other. For music genre classification, the most difficult genre is Reggae ($61 \pm 17\%$), which is most often confused with electronica, R&B, and rock.

Class Set	**Feature Type**	**Feature Rank**								
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
General Audio	**Static**	3	18	11	1	4	8	7	6	5
	Stat. & Temp.	3	62	18	9	25	41	1	4	43
Music Genre	**Static**	2	18	3	16	1	12	17	15	6
	Stat. & Temp.	2	18	3	52	16	1	20	12	46

Table 5: Feature ranking for the AFTE feature set. Feature numbers correspond to the features described in Sec. 2.1.4.

Overall classification results are summarized in Table 6. A comparison across all feature sets shows that, given our choices of classes/genres, the AFTE features provide the highest mean classification rates for both general audio and music genre classification. The differences in overall classification performance, however, are only significant in a few cases: AFTE-SLL for general audio; and AFTE-PA if only static features are used. If we compare classification of specific classes, the AFTE is only significantly outperformed by other feature sets in a few cases: SLL features provide better classification of classical music and crowd noise; MFCC features provide better classification of crowd noise; and PA features provide better classification of speech. On the other hand, the AFTE feature set significantly outperforms all other feature sets in the classification of popular music (general audio class) and jazz and R&B (music genres).

Table 6 also shows that, for every feature set, the inclusion of

	Static Features				Static and Temporal Features			
	SLL	**MFCC**	**PA**	**AFTE**	**SLL**	**MFCC**	**PA**	**AFTE**
General Audio	$80 \pm 5\%$	$87 \pm 4\%$	$76 \pm 5\%$	$90 \pm 3\%$	$86 \pm 4\%$	$92 \pm 3\%$	$92 \pm 3\%$	$93 \pm 2\%$
Music Genre	$55 \pm 12\%$	$57 \pm 12\%$	$41 \pm 14\%$	$70 \pm 9\%$	$61 \pm 11\%$	$65 \pm 10\%$	$62 \pm 10\%$	$74 \pm 9\%$

Table 6: Classification Results Summary. Each entry gives the percent correct classification $\pm$ standard error for the given set of audio classes (left column) and feature set (top rows).

temporal features increases the mean overall classification performance. The differences are not significant in overall classification rates but they are for many individual classes.

4 Discussion

It is well known that, for audio signals, temporal envelope fluctuations at specific rates play an important role in perception. We have shown here that the explicit inclusion of parameters describing these modulations (not only in estimates of the temporal envelope but in other features as well) can increase the performance of audio and music classifiers. We have also shown that a feature set based on a model of auditory perception outperforms other current standard feature sets in the classification of general audio and music genre.

While the overall classification performance of our general audio classes is quite high ($93 \pm 2\%$), music genre classification is far from perfect ($74 \pm 9\%$). While this measure of performance may seem low, it should be pointed out that the classes of music genre do not always have distinct boundaries, which makes their classification a fuzzy problem. We have attempted, in our selection of audio files, to create an internally consistent database so that each music genre contains examples with similar audio qualities. In this manner we can evaluate which features or properties of features are important for characterizing audio qualities relevant to musical genre. Nevertheless there are a number of other (non-acoustic) properties that contribute to labeling a piece of music as a specific genre, including artist, album, and record label. These aspects of genre labeling will not likely be accounted for in features extracted from the audio. So, while we are using the classification of musical genre as a means to measure how relevant our features are, they may never be able to do the job perfectly. In comparison to results of other studies of music genre classification (Tzanetakis et al., 2001; Tzanetakis & Cook, 2002), our features looks quite promising.

Several limitations of the current study should be mentioned. Our audio database is far from complete. We have shown clear advantages of particular feature sets operating on our database but these methods should be performed on larger data sets for confirmation. In addition, a larger database would likely reduce variance in our estimates of classification performance, and allow more conclusive comparisons between the different feature sets.

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.

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 temporal modulations of features are important for the classification of audio and music.

Overall, we saw that the AFTE feature set is the most powerful. However, for a few particular audio classes, classification was better with other feature sets (crowd noise: SLL and MFCC; classical music: SLL; speech: PA).

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., & Weber, R. (1997). Psychoacoustical roughness: Implementation of an optimized model. *Acustica·acta acustica*, *83*, 113-123.

Duda, R., & Hart, P. (1973). *Pattern classification and scene analysis.* New York: Wiley.

Efron, B., & Tibshirani, R. (1997). Improvements on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association*, *92*(438), 548-560.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap.* New York: Chapman & Hall.

Glasberg, B. R., & Moore, B. C. J. (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, *47*, 103–138.

Golub, S. (2000). *Classifying recorded music.* Unpublished master's thesis, University of Edinburgh. (Retrieved January 12, 2003, from `http://www.aigeek.com/aimsc/`)

Hartmann, W. M. (1997). *Signals, sound, and sensation.* Woodbury, New York: American Institute of Physics Press.

Li, D., Sethi, I. K., Dimitrova, N., & McGee, T. (2001). Classification of general audio data for content-based retrieval. *Pattern Recognition Letters*, *22*, 533-544.

Lu, G., & Hankinson, T. (1998). A technique towards automatic audio classification and retrieval. In *4th International Conference on Signal Processing.* Beijing. (Retrieved November 3, 2002, from `http://www.gscit.monash.edu.au/~guojunl/icsp98-1.pdf`)

Papoulis, A. (1991). *Probability, random variables and stochastic processes.* New York: McGraw-Hill.

Peeters, G., Burthe, A. L., & Rodet, X. (2002). Toward automatic music audio summary generation from signal analysis. In M. Fingerhut (Ed.), *Proceedings of the Third International Conference on Music Information Retrieval* (p. 94-100). Paris, France.

Plomp, R., & Levelt, W. J. M. (1965). Tonal consonance and critical bandwidth. *Journal of the Acoustical Society of America.*, *38*(2), 548-560.

Scheirer, E., & Slaney, M. (1997). Construction and evaluation of a robust multifeature speech/music discriminator. In *Proceedings of the ieee 22nd International Conference on Acoustics, Speech and Signal processing* (pp. 1331–1334). Munich, Germany. (Retrieved January 30, 2002, from `citeseer.nj.nec.com/scheirer97construction.html`)

Scheirer, E. D. (1998). Tempo and beat analysis of acoustical musical signals. *Journal of the Acoustical Society of America*, *103*, 588–601.

Slaney, M. (1998). *Auditory toolbox* (Tech. Rep. No. 1998-010). Interval Research Corporation. (Retrieved 30 January, 2002 from `http://rvl4.ecn.purdue.edu/~malcolm/interval/1998-010`)

Spina, M. S., & Zue, V. W. (1997). Automatic transcription of general audio data: Preliminary analysis. In *Proceedings of the 4th International Conference on Spoken Language Processing.* Philadelphia, PA.

Terhardt, E. (1974). On the perception of periodic sound fluctuations (roughness). *Acustica*, *30*, 201-213.

Toonen Dekkers, R. T. J., & Aarts, R. M. (1995). *On a very low-cost speech-music discriminator* (Tech. Rep. No. 124/95). Eindhoven, Netherlands: Philips Research Nat.Lab. Technical Note.

Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, *10*(5), 293-302.

Tzanetakis, G., Essl, G., & Cook, P. (2001). Automatic musical genre classification of audio signals. In *Proceedings of the 2nd Annual International Symposium for Music Information Retrieval.* Princeton, NJ.

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

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

Zhang, T., & Kuo, C. (1998). Content-based classification and retrieval of audio. In *SPIE's 43rd Annual Meeting - Conference on Advanced Signal Processing Algorithms, Architectures, and Implementations VIII.* San Diego.

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

Zwicker, E., & Fastl, H. (1990). *Psychoacoustics: Facts and models.* Berlin: Springer.