DECISION TIME HORIZON FOR MUSIC GENRE CLASSIFICATION USING SHORT TIME FEATURES

Peter Ahrendt, Anders Meng and Jan Larsen

Informatics and Mathematical Modelling, Technical University of Denmark
Richard Petersens Plads, Building 321, DK-2800 Kongens Lyngby, Denmark
phone: (+45) 4525 3888,3891,3923, fax: (+45) 4587 2599, email: pa,am,jl@imm.dtu.dk, web: http://isp.imm.dtu.dk

ABSTRACT

In this paper music genre classification has been explored with special emphasis on the decision time horizon and ranking of tapped-delay-line short-time features. Late information fusion as e.g. majority voting is compared with techniques of early information fusion[1] such as dynamic PCA (DPCA). The most frequently suggested features in the literature were employed including mel-frequency cepstral coefficients (MFCC), linear prediction coefficients (LPC), zero-crossing rate (ZCR), and MPEG-7 features. To rank the importance of the short time features *consensus sensitivity analysis* is applied. A Gaussian classifier (GC) with full covariance structure and a linear neural network (NN) classifier are used.

1. INTRODUCTION

In the recent years, the demand for computational methods to organize and search in digital music has grown with the increasing availability of large music databases as well as the growing access through the Internet. Current applications are limited, but this seems very likely to change in the near future as media integration is a high focus area for consumer electronics [6]. Moreover, radio and TV broadcasting are now entering the digital age and the big record companies are starting to sell music on-line on the web. An example is the popular product iTunes by Apple Computer, which currently has access to a library of more than 500,000 song tracks. The user can then directly search and download individual songs through a website for use with a portable or stationary computer.

A few researchers have attended the specific problem of music genre classification, whereas related areas have received more attention. An example is the early work of Scheirer and Slaney [17] which focused on speech/music discrimination. Thirteen different features including *zero-crossing rate* (ZCR), *spectral centroid* and *spectral roll-off point* were examined together using both Gaussian, GMM and KNN classifiers. Interestingly, choosing a subset of only three of the features resulted in just as good a classification as with the whole range of features. In another early work Wold *et al.* [22] suggested a scheme for audio retrieval and classification. Perceptually inspired features such as pitch, loudness, brightness and timbre were used to describe the audio. This work is one of the first in the area of content-based audio analysis, which is often a supplement to the classification and retrieval of multimodal data such as video. In [12], Li *et al.* approached segment classification of audio streams from TV into seven general audio classes. They find that *mel-frequency cepstral coefficients* (MFCCs) and *linear prediction coefficients* (LPCs) perform better than features such as ZCR and *short-time energy* (STE).

The genre is probably the most important descriptor of music in everyday life. It is, however, not an intrinsic property of music such as e.g. tempo and makes it somewhat more difficult to grasp with computational methods. Aucouturier *et al.* [2] examined the inherent problems of music genre classification and gave an overview of some previous attempts. An example of a recent computational method is Xu *et al.* [23], where support vector machines were used in a multi-layer classifier with features such as MFCCs, ZCR and LPC-derived cepstral coefficients. In [13], Li *et al.* introduced DWCHs (Daubechies wavelet coefficient histograms) as novel features and compared these to previous features using four different classifiers. Lambrou *et al.* [11] examined different wavelet transforms for classification with a minimum distance classifier and a least-squares minimum distance classifier to classify into rock, jazz and piano. The state-of-art percentage correct performance is around 60% considering 10 genres, and 90% considering 3 genres.

In the MPEG-7 standard [8] audio has several *descriptors* and are meant for general sound, but in particular speech and music. Casey [5] introduced some of these descriptors, such as the *audio spectrum envelope* (ASE) to successfully classify eight musical genres with a hidden markov model classifier.

McKinney *et al.* [15] approached audio and music genre classification with emphasis on the features. Two new feature sets based on perceptual models were introduced and compared to previously proposed features with the use of Gaussian-based quadratic discriminant analysis. It was found that the perceptually based features performed better than the traditional features. To include temporal behavior of the short-time features (23 ms frames), four summarized values of the power spectrum of each feature is found over a longer time frame (743 ms). In this manner, it is argued that temporal descriptors such as beat is included.

Tzanetakis and Cook [20] examined several features such as spectral centroid, MFCCs as well as a novel beat-histogram. Gaussian, GMM and KNN classifiers were used to classify music on different hierarchical levels such as e.g. classical music into choir, orchestra, piano and string quartet.

In the last two mentioned works, some effort was put into the examination of the time-scales of features and the decision time-horizon for classification. However, this generally seems to be a neglected area and has been the motivation for the current paper. How much time is, for instance, needed to make a sufficiently accurate decision about the musical genre? This might be important in e.g. hearing aids and streaming media. Often, some kind of early information fusion of the short-time features is achieved by e.g. taking the mean or another statistics over a larger window. Are the best features then the same on all time-scales or does it depend on the decision time horizon? Is there an advantage of early information fusion as compared to late information fusion such as e.g. majority voting among short-time classifications, see further e.g., [9]. These are the main questions to be addressed in the following.

In section 2 the examined features will be described. Section 3 deals with the methods for extracting information about the time scale behavior of the features, and in section 4 the results are presented. Finally, section 5 state the main conclusions.

2. FEATURE EXTRACTION

Feature extraction is the process of capturing the complex structure in a signal using as few features as possible. In the case of timbral textual features a frame size, in which the signal statistics are assumed stationary is analyzed and features are extracted. All

[1] This term refers to the decision making, i.e., early information fusion is an operation on the features *before* classification (and decision making). This is opposed to late information fusion (decision fusion) that assembles the information on the basis of the decisions.

features described below are derived from short-time 30 ms audio signal frames with a hop-size of 10 ms.

One of the main challenges when designing music information retrieval systems is to find the most descriptive features of the system. If good features are selected one can relax on the classification methodology for fixed performance criteria.

2.1 Spectral signal features

The spectral features have all been calculated using a Hamming window for the *short time Fourier transform* (STFT) to minimize the side-lobes of the spectrum.

MFCC and LPC. The MFCC and LPC both originate from the field of automatic speech recognition, which has been a major research area through several decades. They are carefully described in this context in the textbook by Rabiner and Juang [16]. Additionally, the usability of MFCCs in music modeling has been examined in the work of Logan [14]. The idea of MFCCs is to capture the short-time spectrum in accordance with human perception. The coefficients are found by first taking the logarithm of the STFT and then performing a mel-scaling which is supposed to group and smooth the coefficients according to perception. At last, the coefficients are decorrelated with the discrete cosine transform which can be seen as a computationally cheap PCA. LPCs are a short-time measure where the coefficients are found from modeling the sound signal with an all-pole filter. The coefficients minimizes a least-square measure and the LPC gain is the residual of this minimization. In this project, the autocorrelation method was used. The delta MFCC (DMFCC $\equiv$ MFCC$_n$ - MFCC$_{n-1}$) and delta LPC (DLPC $\equiv$ LPC$_n$ - LPC$_{n-1}$) coefficients are further included in the investigations.

MPEG-7 audio spectrum envelope (ASE). The *audio spectrum envelope* is a description of the power contents in log-spaced frequency bands of the audio signal. The log-spacing is done as to resemble the human auditorial system. The ASE have been used in e.g. audio thumbnailing and classification, see [21] and [5]. The frequency bands are determined using an $1/4$-octave between a lower frequency of 125 Hz, which is the "low edge" and a high frequency of 9514 Hz.

MPEG-7 audio spectrum centroid (ASC). The *audio spectrum centroid* describes the center of gravity of the log-frequency power spectrum. The descriptor indicates whether the power spectrum is dominated by low or high frequencies. The centroid is correlated with the perceptual dimension of timbre named *sharpness*.

MPEG-7 audio spectrum spread (ASS) . The *audio spectrum spread* describes the second moment of the log-frequency power spectrum. It indicates if the power is concentrated near the centroid, or if it is spread out in the spectrum. It is able to differentiate between tone-like and noise-like sounds [8].

MPEG-7 spectral flatness measure (SFM). The *audio spectrum flatness measure* describes the flatness properties of the spectrum of an audio signal within a number of frequency bands. The SFM feature expresses the deviation of a signal's power spectrum over frequency from a flat shape (noise-like or impulse-like signals). A high deviation from a flat shape might indicate the presence of tonal components. The spectral flatness analysis is calculated for the same number of frequency bands as for the ASE, except that the low-edge frequency is 250 Hz. The SFM seem to be very robust towards distortions in the audio signal, such as MPEG-$1/2$ layer 3 compression, cropping and dynamic range compression [1]. In [4] the centroid, spread and SFM have been evaluated in a classification setup.

All MPEG-7 features have been extracted in accordance with the MPEG-7 audio standard [8].

2.2 Temporal signal features

The temporal features have been calculated on the same frame basis as the spectral features.

Zero crossing rate (ZCR). ZCR measures the number of time domain zero-crossings in the frame. It can be seen as a descriptor of the dominant frequency of music and to find silent frames.

Short time energy (STE). This is simply the mean square power in the frame.

3. FEATURE RANKING - SENSITIVITY MAPS

3.1 Time stacking and dynamic PCA

To investigate the importance of the features at different time scales a tapped-delay line of time stacking features is used. Define an extended feature vector as

$$\mathbf{z}_n = [\mathbf{x}_n, \mathbf{x}_{n-1}, \mathbf{x}_{n-2}, \ldots, \mathbf{x}_{n-L}]^T ,$$

where L is the lag-parameter and $\mathbf{x}_n$ is the row feature vector at frame n. Since the extended vector increases in size as a function of L, the data is projected into a lower dimension using PCA. The above procedure is also known as dynamic PCA (DPCA) [10] and reveals if there is any linear relationship between e.g. $\mathbf{x}_n$ and $\mathbf{x}_{n-1}$; thus not only correlations but also cross-correlations between features. The decorrelation performed by the PCA will also include a decorrelation of the time information, e.g. is MFFC-1 at time n correlated with LPC-1 at time $n-5$?

At $L = 100$ the number of features will be 10403 which makes the PCA computational intractable due to memory and speed. A "simple" PCA have been used where only 1500 of the total of 10403 largest eigenvectors is calculated by random selection of training data, see e.g. [19]. To investigate the validity of the method 200 eigenvectors was used at $L = 50$ and the number of random selected data points was varied between $200 - 1500$. The variation in classification error was less than a percent, thus indicating that this is a robust method. Due to memory problems originating from the time stacking, the largest used lag time is $L = 100$, which corresponds to one second of the signal.

3.2 Feature ranking

One of the goals of this project is to investigate which features are relevant to the classification of music genres at different time scales. Selection of single best method for feature ranking is not possible, since several methods exists each with their advantages and disadvantages. An introduction to feature selection can be found in [7], which also explains some of the problems using different ranking schemes. Due to the nature of our problem a method known as the *sensitivity map* is used, see e.g. [18]. The influence of each feature on the classification bounds is found by computing the gradient of the posterior class probability $P(C_k|\mathbf{x})$ w.r.t. all the features. Here C_k denotes the k'th genre. One way of computing a sensitivity map for a given system is the *absolute value average sensitivities* [18]

$$\mathbf{s} = \frac{1}{NK} \sum_{k=1}^{K} \sum_{n=1}^{N} \left| \frac{\partial P(C_k|\tilde{\mathbf{x}}_n)}{\partial \mathbf{x}_n} \right| , \qquad (1)$$

where $\mathbf{x}_n$ is the n'th time frame of a test-set and $\tilde{\mathbf{x}}_n$ is the n'th time frame of the same test-set projected into the M largest eigenvectors of the training-set. Both $\mathbf{s}$ and $\mathbf{x}_n$ are vectors of length D - the number of features. N is the total number of test frames and K is the number of genres. Averaging is performed over the different classes as to achieve an overall ranking independent of the class. It should be noted that the sensitivity map expresses the importance of each feature individually - correlations are thus neglected.

For the linear neural network an estimate of the posterior distribution is needed to use the sensitivity measure. This is achieved using the softmax-function, see e.g. [18].

4. RESULTS

Two different classifiers were used in the experiments: a Gaussian classifier with full covariance matrix and a simple single-layer neural network which was trained with sum-of-squares error function to facilitate the training procedure. These classifiers are quite similar, but they differ in the discriminant functions which are quadratic and linear, respectively. Furthermore the NN is inherently trained discriminatively. They are also quite simple, but after experimentation with more advanced methods, like the Gaussian mixture models and HMMs, this became a necessity in order to carry out the vast amount of training operations needed. Further, the purpose of this study is not to obtain optimal performance rather to investigate the relevance of relevant short-time features.

The data set was split into training, validation and test sets. The validation set was used only to select the number of DPCA-components. The best classification was found with 50 components at both $L = 50$ and $L = 100$. The data was split with 50, 25 and 25 sound files in each set, respectively, and each of these were distributed evenly into five music genres: Pop, Classical, Rock, Jazz, Techno. All sound files have a duration of 10 s and with a hop-size of 10 ms. This resulted in 1000 30 ms frames per sound file. The used sampling frequency is 22050 Hz. The size of the training set as well as duration of the sound files was determined from learning curves[2] (results not shown). After the feature extraction, the features were normalized to zero mean and unit variance to make them comparable.

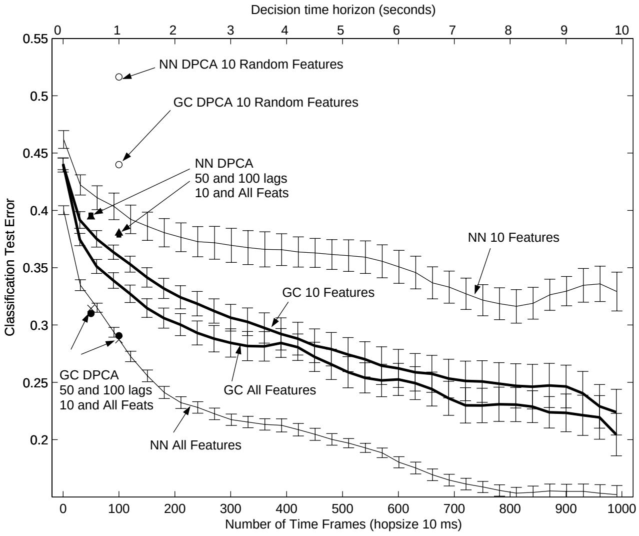

Figure 1: Classification error as a function of the lag of the GC and NN using DPCA and majority voting, respectively.

Figure 1 summarizes the examination of the decision time horizon as well as the comparison between early and late information fusion using DPCA and majority voting, respectively. It is seen from the figure that there is not an obvious advantage of using the DPCA transform instead of the computationally much cheaper majority voting. However, it can be seen from table 1 and 2 that the methods' performance depends on the genre. The tables show test classification error for each genre with error-bars obtained by repeating the experiment 50 times on randomly selected training data. The number in parenthesis shows the percentage relative to lag $L = 0$ of the classifier. For instance, it is seen that the DPCA gives remarkably better classification of jazz than majority voting. This might be used constructively to create a better classifier.

Figure 1 also shows the results after choosing the 10 features with the best sensitivity consensus ranks (see below). There is a small deviation for the GC and a large deviation for the NN between the 10 best features and the full feature set when majority voting is used. This might be connected to the differences in the number of variables in the two classifiers which implies that the curve for the NN with 10 features is dominated by bias since the number of variables is only $5 \cdot 11 = 55$. Thus, 10 features is not really enough for this classifier. In contrast, the GC with 103 features has more than 25000 different variables and might be dominated by variance which increases the test error. However, the sensitivity ranking still seems reasonable when compared to the full feature sets and when comparisons are made with the classification error from a set of 10 random features (illustrated in the figure).

Another examination of early information fusion was also carried out by using the mean values of the short-time features over increasing time frames (from 1 to 1000 frames). The classification results are not illustrated, however, since approximately the same classification rate as without the time information (lag $L = 0$) was achieved at all time scales, though with a lot of fluctuations.

Full Feature Set	**Pop**	**Classic**	**Rock**	**Jazz**	**Techno**
NN (L=0)	36% ± 0.8%	27% ± 2%	29% ± 1.1%	67% ± 1.1%	41% ± 0.7%
Maj.Vote (L=100)	17%(−19)	19%(−8)	26%(−3)	63%(−4)	29%(−12)
Time Stacking (L=100)	21%(−15)	22%(−5)	21%(−8)	45%(−22)	34%(−7)
GC (L=0)	50% ± 0.2%	39% ± 0.5%	27% ± 0.2%	71% ± 0.5%	31% ± 0.3%
Maj.Vote (L=100)	32%(−18)	28%(−11)	22%(−5)	68%(−3)	17%(−14)
Time Stacking (L=100)	28%(−22)	29%(−10)	21%(−6)	39%(−32)	26%(−5)

Table 1: Test error classificstion rates of Gaussian Classifier (GC) and Neural Network (NN) using the full feature set.

Best 10 Feat.	**Pop**	**Classic**	**Rock**	**Jazz**	**Techno**
NN (L=0)	38% ± 1.4%	30% ± 2.5%	40% ± 2.1%	86% ± 1.4%	37% ± 0.96%
Maj.Vote (L=100)	27%(−11)	23%(−7)	38%(−2)	88%(+2)	25%(−12)
Time Stacking (L=100)	21%(−17)	23%(−7)	45%(+5)	65%(−21)	37%(0)
GC (L=0)	34% ± 0.6%	35% ± 1.5%	38% ± 1.4%	65% ± 1.2%	47% ± 0.8%
Maj.Vote (L=100)	22%(−12)	26%(−9)	32%(−6)	62%(−3)	39%(−8)
Time Stacking (L=100)	36%(+2)	32%(−3)	22%(−16)	43%(−22)	12%(−35)

Table 2: Test error classification rates of Gaussian Classifier (GC) and Neural Network (NN) using the 10 best features.

The training of the models has been repeated 50 times on different song clips, and the sensitivies have been calculated and ranked. It is now possible to obtain a consensus ranking from the cumulated sensitivity histograms of the 103 features, which is shown in figure 2. Each row shows the cumulated sensitivity histogram where dark color corresponds to large probability. For $L = 0$ the number of features is $D = 103$, but for $L = 100$ the amount of features is $D = 10403$ due to the time stacking. A similar plot could be generated at $L = 100$ but the histograms of each feature would not be easy to see and interpret. To rank the features, at e.g. $L = 100$, the mean value of the sensitivity over time of each feature is applied, which results in only 103 time-averaged features in figure 2. The mean value is applied since only low frequency variation in sensitivity over lag-parameters are present (below 5 Hz). To provide the consensus features, the feature which has the highest cumulated histogram frequency in each column is selected.

Experiments with ranking of the features at $L = \{0, 50, 100\}$ clearly indicates that delta features generally ranks lower at higher lag time, see also area **B** and **D** in figure 2 for $L = 100$. The MFCC(**A**) and LPC(**C**) generally rank better than e.g. the ASE(**E**) and SFM(**F**) coefficients. However, the high frequency components of both the ASE and SFM also show relevance, which is an indicator of "noise-like" parts in the music. The 10 best consensus features for $L = \{0, 50, 100\}$ are shown in table 3. A sanity check of the sen-

[2] Classification error or log-likelihood as a function of the size of the training set.

sitivity map was performed using the Optimal Brain Damage [3] for $L = 0$ and showed similar results.

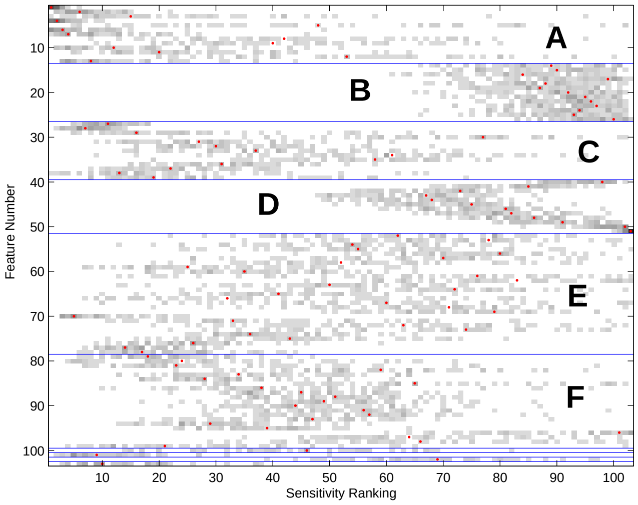

Figure 2: Consensus feature ranking of individual feature at $L = 100$. See text for interpretation. The features are MFCC(**A**), DMFCC(**B**), LPC(**C**), DLPC(**D**), ASE(**E**), SFM(**F**) and the single features ASC, ASS, STE and ZCR. The ten best features in decreasing order are: $\{1, 4, 6, 7, 70, 2, 28, 13, 101, 103\}$.

L=0 (1 to 5)	LPC2	LPC1	MFCC2	LPC3	MFCC4
L=50 (1 to 5)	MFCC1	MFCC4	MFCC6	MFCC2	LPC2
L=100 (1 to 5)	MFCC1	MFCC4	MFCC6	MFCC7	ASE19
L=0 (6 to 10)	LPC4	LPC5	GAIN	MFCC1	MFCC3
L=50 (6 to 10)	MFCC7	ASE19	LPC1	ASS	MFCC10
L=100 (6 to 10)	MFCC2	LPC2	MFCC13	ASS	ZCR

Table 3: The 10 best consensus features of the NN classifier as a function of the time stack lag, L. The DPCA transform was employed.

5. CONCLUSION

Music genre classification has been explored with special emphasis on the decision time horizon and ranking of tapped-delay line short-time features. A linear neural network and a Gaussian classifier were used for classification. Information fusion showed increasing performance with time horizon, thus state-of-art 80% correct classification rate is obtained within 5 s decision time horizon. Early and late information fusion showed similar results, thus we recommend the computational efficient majority decision voting. However, investigation of individual genres showed that e.g. jazz is better classified using DPCA. Consensus ranking of feature sensitivities enabled the selection and interpretation of the most salient features. MFCC, LPC and ZCR showed to be most relevant, whereas MPEG-7 features showed less consistent relevance. DMFCC and DLPC showed to be least important for the classification. With only the 10 best features, 70% classification accuracy was obtained using a 5 s decision time horizon.

REFERENCES

[1] E. Allamanche, J. Herre, O. Helmuth, B. Frba, T. Kasten, and M. Cremer, “Content-Based Identification of Audio Material Using MPEG-7 Low Level Description,” in *Proc. of the ISMIR*, Indiana University, USA, Oct. 2001.

[2] J.-J. Aucouturier and F. Pachet, “Representing musical genre: A state of the art,” *Journal of New Music Research*, vol. 32, pp. 83–93, Jan. 2003.

[3] C. Bishop, *Neural Networks for Pattern Recognition*. Oxford: Clarendon Press, 1995.

[4] J.J. Burred and A. Lerch, “A Hierarchical Approach to automatic musical genre classification,” in *Proc. 6th Int. Conf. on Digital Audio Effects ’03*, London, Great Britain, Sept. 2003.

[5] M. Casey, “Sound Classification and Similarity Tools,” in B.S. Manjunath, P. Salembier and T. Sikora (eds), *Introduction to MPEG-7: Multimedia Content Description Language*, J.Wiley, 2001.

[6] 2004 International Consumer Electronics Show, Las Vegas, Nevada, Jan. 8–11, 2004, www.cesweb.org

[7] I. Guyon and A. Elisseeff, “An Introduction to Variable and Feature Selection,” *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, Mar. 2003.

[8] *Information technology Multimedia content description interface - Part 4: Audio*, ISO/IEC FDIS 15938-4:2002(E) Retrieval (ISMIR 2003), Baltimore, Oct. 2003, www.chiariglione.org/mpeg/.

[9] J. Kittler, M. Hatef, R.P.W. Duin and J. Matas, “On Combining Classifiers,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 226–239, Mar. 1998.

[10] W. Ku, R.H. Storer, C. Georgakis, ”Disturbance Detection and Isolation by Dynamic Principal Component Analysis”, *Chemometrics and Intell Lab Sys.*, pp. 179–196, Sept. 1995.

[11] T. Lambrou *et al.*, “Classification of Audio Signals using Statistical Features on Time and Wavelet Transform Domains,” in *Proc. ICASSP ’98* , Seattle, USA, May 1998, pp. 3621–3624.

[12] D. Li *et al.*, “Classification of General Audio Data for Content-Based Retrieval,” *Pattern Recognition Letters*, vol. 22, pp. 533–544, Apr. 2001.

[13] T. Li and M. Ogihara and Q. Li, “A comparative study on content-based music genre classification,” in *Proc. ACM SIGIR ’03*, Toronto, Canada, July 2003, pp. 282–289.

[14] B. Logan, “Mel Frequency Cepstral Coefficients for Music Modeling,” in *Proc. of the International Symposium on Music Information Retrieval 2000*, Plymouth, USA, Oct. 2000.

[15] M.F. McKinney and J. Breebaart, “Features for Audio and Music Classification,” in *4th International Conference on Music Information*, http://ismir2003.ismir.net/papers/McKinney.PDF

[16] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Prentice Hall, 1993.

[17] E. Scheirer and M. Slaney, “Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator,” in *Proc. ICASSP ’97*, Munich, Germany, 1997, pp. 1331–1334.

[18] S. Sigurdsson *et al.*, “Detection of Skin Cancer by Classification of Raman Spectra,” accepted for *IEEE Transactions on Biomedical Engineering, 2003*.

[19] H. Schweitzer, “A Distributed Algorithm for Content Based Indexing of Images by Projections on Ritz Primary Images,” *Data Mining and Knowledge Discovery 1*, pp. 375-390, 1997.

[20] G. Tzanetakis and P. Cook, “Musical Genre Classification of Audio Signals,” *IEEE Transactions on Speech and Audio Processing*, vol. 10, pp. 293–302, July 2002.

[21] J. Wellhausen and M. Höynck, “Audio Thumbnailing Using MPEG-7 Low Level Audio Descriptors,” in *Proc. ITCom ’03*, Orlando, USA , Sept. 2003.

[22] E. Wold, T. Blum, D. Keislar and J. Wheaton, “Content-based classification, search and retrieval of audio,” *IEEE Multimedia Mag.*, vol. 3, pp. 27–36, July 1996.

[23] C. Xu *et al.*, “Musical Genre Classification using Support Vector Machines,” in *Proc. ICASSP ’03*, Hong Kong, China, Apr. 2003, pp. 429–432.