IMPROVING MUSIC GENRE CLASSIFICATION BY SHORT-TIME FEATURE INTEGRATION

Anders Meng, Peter Ahrendt and Jan Larsen

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

ABSTRACT

Many different short-time features, using time windows in the size of 10-30 ms, have been proposed for music segmentation, retrieval and genre classification. However, often the available time frame of the music to make the actual decision or comparison (the decision time horizon) is in the range of seconds instead of milliseconds. The problem of making new features on the larger time scale from the short-time features (*feature integration*) has only received little attention. This paper investigates different methods for feature integration and late information fusion[1] for music genre classification. A new feature integration technique, the *AR* model, is proposed and seemingly outperforms the commonly used mean-variance features.

1. INTRODUCTION

Classification, segmentation and retrieval of music (and audio in general) are topics that have attracted quite some attention lately from both academic and commercial societies. These applications share the common need for features which effectively represent the music. The features ideally contain the information of the original signal, but compressed to such a degree that relatively low-dimensional classifiers or similarity metrics can be applied. Most efforts have been put in short-time features, which extract the information from a small sized window (often $10-30$ ms). However, often the decision time horizon is in the range of seconds and it is then necessary either to find features directly on this time scale or somehow integrate the information from the time series of short-time features over the larger time window. Additionally, it should be noted that in classification problems, the information fusion could also be placed after the actual classifications. Such late fusion could e.g. be majority voting between the classifications of each short-time feature.

In [1] and [2], features are calculated directly on the large time-scale (long-time features). They try to capture the perceptual beats in the music, which makes them intuitive and easy to test against a music corpora. In contrast, short-time features can only be tested indirectly through e.g. their performance in a classification task.

Feature integration is most often performed by taking the mean and variance of the short-time features over the decision time horizon (examples are [3], [4] and [5]). Computationally, the mean and variance features are cheap, but the question is how much of the relevant feature dynamics they are able to capture. As an attempt to capture the dynamics of the short-time features, [6] uses a spectral decomposition of the Mel-Frequency Cepstral Coefficients (*MFCCs*) into 4 different frequency bands. Another approach, by [7], takes the ratio of values above and below a constant times the mean as the long-time feature. Their short-time features are Zero-Crossing Rate and Short-Time Energy.

In a previous investigation [8], the authors examined feature integration by dynamic PCA where the idea is to stack short-time features over the decision time horizon and then use PCA to reduce the dimensionality (finding correlations both across time and features). Dynamic PCA was compared with late fusion in the form of majority voting, but the results did not strongly favor any of the methods.

Altogether, the idea of short-time feature integration seems scarcely investigated, although several researchers (necessarily) make use of it. This has been the main motivation for the current work, together with methods for late information fusion.

In Section 2, the investigated features and feature integration techniques are described. Section 3 concerns the employed classifiers and late information fusion schemes. In section 4, the results are analyzed and, finally, section 5 concludes on the results.

2. FEATURE MODEL

In this article the selected features exist either on a short, medium or long time scale. The timescales used can be seen from table 1. Short time only consider the immediate frequencies, and do not contain long structural temporal information. Medium time features can contain temporal information such as e.g. modulation (instrumentation) and long time features can contain structural information such as beat. Classification at short time only provide reasonable results using a computer, since human decision time horizons typically are 250ms or above for a moderate error [5].

Time scale	Frame size	Perceptual meaning
Short time	30ms	timbre (instant frequency)
Medium time	740ms	modulation (instrumentation)
Long time	9.62s	beat, mood vocal etc.

Table 1. The different time levels with corresponding perceptual interpretation.

[1] Late information fusion assemble the probabilistic output or decisions from a classifier over the short-time features (an example is majority voting). In early information fusion (which includes feature integration) the information is integrated before or in the classifier.

Depending on the decision time horizon, the performance at short time might not be adequate, in which more time is needed. There are several possibilities to increase the decision time horizon, either using the classifier in an early/late information fusion setting, which will be elaborated in section 3, or to use features derived at these time horizons. Figure 1 show the investigated features for the music genre setup and their relationships.

2.1. Short time features (1)

The short time features have been derived using a hop- and frame size of 10 and 30ms, respectively. Typically the frame size is selected such that the in-frame signal is approximately stationary.

Mel Frequency Cepstral Coefficients were originally developed for automatic speech recognition systems [9, 10], but have lately been used with success in various audio information retrieval tasks. Recent studies [8, 11] indicate that they outperform other features existing at a similar time level. From the previous investigations [8], good performance was achieved, hence, these are the only features considered at this decision time horizon. It was found that the first 6 *MFCCs* were adequate for the music genre classification task, in line with [5].

2.2. Medium time features (2)

The medium time features are based on a frame size of 740ms similar to [6] and a hop size of 370ms.

Mean and variance (*MV*) of the *MFCCs*. Mean and variance is a simple way to perform feature integration and the most commonly used, see e.g. [1, 3, 4].

Filterbank Coefficients (*FC*) is another method of feature integration. This method was proposed in [6] and suggests to calculate the power spectrum for each *MFCC* on a frame size of 740ms. The power is summarized in four frequency bands: 1) 0 Hz average of *MFCCs*, 2) $1-2$ Hz modulation energy of the *MFCCs*, 3) 3-15Hz and 4) 20-50 Hz (50Hz is half the sampling rate of the *MFCCs*). Experiments suggested that better performance could be achieved using more than 4 bins, which seems reasonable since these features was originally developed for general sound recognition.

Autoregressive model (*AR*) is a well-known technique for time series regression. Due to its simplicity and good performance in time-series modelling, see e.g. [12], this model is suggested for feature integration of the *MFCCs*. The *AR* method and *FC* approach resembles each other since the integrated ratio of the signal spectrum to the estimated spectrum is minimized in the *AR* method [13]. This suggests that the power spectrum of each *MFCC* is modelled. The *AR* parameters have been calculated using the windowed autocorrelation method, using a rectangular window. To the authors knowledge an *AR*-model has not previously been used for music feature integration. In all of the AR-related features, the mean and gain are always included along with a number of AR-coefficients. This number is given by the model order, which is found by minimizing validation classification error on the data set.

High Zero-Crossing Rate Ratio (*HZCRR*) is defined as the ratio of the number of frames whose time zero crossing rates (No. of times the audio signal crosses 0) are above 1.5 times the average.

Low Short-Time energy ratio (*LSTER*) is defined as the ratio of the number of frames whose short time energy is less than 0.5 times the average.

Both the *LSTER* and *HZCRR* features are explained further in [7]. They are derived directly from the audio signal, which makes them computationally cheap. It should be mentioned that the *HZCRR* and *LSTER* were originally meant for speech/music segmentation. In the experiments, they were combined into the feature *LSHZ* to improve their performance.

2.3. Long time features (3)

All the long time features have a hop- and frame size of 4.81 and 9.62 seconds, respectively. Many of the features at this decision time have been derived from features at an earlier timescale (feature integration), e.g. AR_{23a} is integrated from medium time to long time using an *AR* model on each of the *AR* medium time features. The different combinations applied can be seen from figure 1, where the arrows indicate which features are integrated to a longer time scale. Additionally, all the long-time features have been combined into the feature, *All*, and PCA was used for dimensionality reduction.

Beat spectrum (*BS*) has been proposed by [2] as a method to determine the perceptual beat. The *MFCCs* are used in the beat spectrum calculation. To calculate the frame similarity matrix, the cosine measure has been applied. The beat spectrum displays peaks when the audio has repetitions. In the implementation the discrete fourier transform is applied to the beat spectrum in order to extract the main beat and sub beats. The power spectrum is then aggregated in 6 discriminating bins wrt. music genre.

Beat histogram (*BH*) was proposed in [1] as a method for calculating the main beat as well as sub-beats. The implementation details can be found in [1]. In our implementation the discrete wavelet transform is not utilized, but instead an octave frequency spacing has been used. The resulting beat histogram is aggregated in 6 discriminating bins.

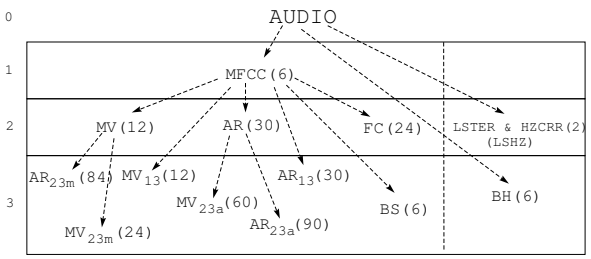


Fig. 1. Short(1), medium(2) and long(3) time features and their relationships. The arrow from e.g. medium time *MV* to the long time feature AR_{23m} indicate feature integration. Thus, for each of the 12 time-series of *MV* coefficients, 7 *AR* features have been found, resulting in a $7 \cdot 12 = 84$ dimensional feature vector AR_{23m}. The optimal feature dimension (shown in parenthesis) for the various features have been determined from a validation set, hence selecting the dimension which minimizes the validation error.

3. CLASSIFIERS AND COMBINATION SCHEMES

For classification purposes two classifiers were considered: 1) A simple single-layer neural network (LNN) trained with sum-of -squares error function to facilitate the training procedure and 2) A gaussian classifier (GC) with full covariance matrix. The two

classifiers differ in their discriminant functions which are linear and quadratic, respectively. Furthermore the LNN is inherently trained discriminatively. More sophisticated methods could have been used for classification, however, the main topic of this research was to investigate methods of information fusion in which the proposed classifiers will suffice.

The two fusion schemes considered were early and late information fusion. In early information fusion the complex interactions that exist between features in time is modelled in or before the statistical classification model. The feature integration techniques previously mentioned (such as the *AR*, *FC*, AR_{23a} and MV_{13} features) can be considered as early fusion. Late information fusion is the method of combining results provided from the classifier. There exists several combination schemes for late information fusion, see e.g. [14]. In the present work, the majority vote rule, sum rule and the median rule were investigated. In the majority vote rule, the votes received from the classifier are counted and the class with the largest amount of votes is selected, hereby performing consensus decision. In sum-rule the posterior probabilities calculated from each example are summed and a decision is based on this result. The median rule is like the sum rule except being the median instead of the sum. During the initial studies it was found that the sum rule outperformed the majority voting and median rule, consistent with [14], and therefore preferred for late information fusion in all of the experiments.

4. RESULTS AND DISCUSSION

Experiments were carried out on two different data sets. The purpose was not so much to find the actual test error on the data sets, but to compare the relative performances of the features.

For some of the features, dimensionality reduction by PCA was performed. Learning curves, which are plots of the test error as a function of the size of the training set, were made for all features. From these curves, it was found necessary to use PCA on AR_{23a}, AR_{23m}, MV_{23a} and the combined long-time features set (denoted All). It was found that approximately 20 principal components gave optimal results.

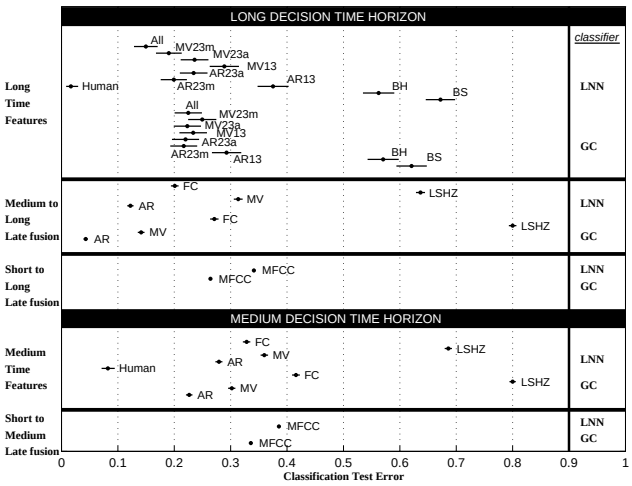
The classification test errors are shown in figure 2 for both of the data sets and both the medium time and long time classification problems.

4.1. Data set 1

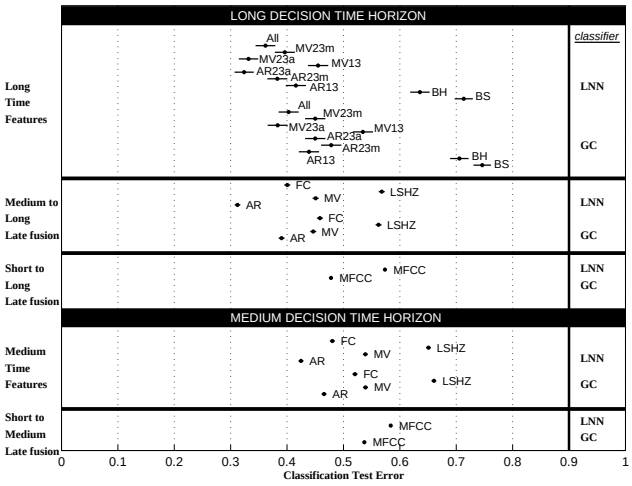
The data set consisted of the same 100 songs, that were also used in [8]. The songs were distributed evenly among classical, (hard) rock, jazz, pop and techno. The test set was fixed with 5 songs from each genre and using 30 seconds from the middle of the songs. The training set consisted of three pieces each of 30 seconds from each song, resulting in 45 pieces. For cross-validation, 35 of these pieces were picked randomly for each of the 10 training runs.

4.1.1. Human classification

To test the integrity of the music database, a human classification experiment was carried out on the data set. 22 persons were asked each to classify (by forced-choice) 100 of the 740 ms and 30 of 10 s samples from the test set. The average classification rate across people and across samples was 98% for the 10 s test and 92% for the 740 ms test. The lower/upper 95% confidence limits were



(a) Experiment on data set 1



(b) Experiment on data set 2

Fig. 2. The figure illustrates the classification test errors for data set 1 in the upper part and data set 2 in the lower. Each part contains test errors from both the long decision time horizon (10 s) and the medium decision time horizon (740 ms). Thus, the block "Medium to Long Late Fusion" under "Long Decision Time Horizon" include all the medium-time features, such as *AR* and *FC* features, where the sum rule has been used to fuse information from the medium to long time scale. The results for the same medium-time features without any late fusion, would then be placed in "Medium Time Features" under "Medium Decision Time Horizon". The results from both classifiers on the same features are placed in the same block (GC is Gaussian Classifier, LNN is Linear Neural Network). All the abbreviations of the features are explained in section 2. The 95%- confidence intervals have been shown for all features.

97/99% and 91/93%, respectively. This suggests that the genre labels, that the authors used, are in good agreement with the common genre definition.

4.2. Data set 2

The data set consisted of 354 music samples each of length 30 seconds from the "Amazon.com Free-Downloads" database [15]. The songs were classified evenly into the six genres classical, country, jazz, rap, rock and techno and the samples were split into 49 for training and 10 for testing. From the training samples, 45 were randomly chosen in each of the 10 cross-validation runs. The authors found it much harder to classify the samples in this data set than in the previous, but it is also considered as a much more realistic representation of an individuals personal music collection.

4.3. Discussion

Notably, as seen in figure 2, the feature *LSHZ*, *BS* and *BH* perform worse than the rest of the features on both data sets. This may not be surprising since they were developed for other problems than music classification and/or they were meant as only part of a larger set of features. The *FC* did not do as well as the *AR* features. A small investigation indicated that *FCs* have the potential to perform better by changing the number of frequency bins, though still not as good as *ARs*.

A careful analysis of the *MV* and *AR* features, and the feature integration combinations of these, has been made. By comparing the early fusion combinations of these, as seen in figure 2 (in the part "Long-time features"), it is quite unclear which of these perform the best. When the late fusion method is used (in the part "Medium to long late fusion"), the results are more clear and it seems that the *AR* feature performs better than the *MV* and *FC* features. This view is supported by the results in the "Medium-time features" part. Using the McNemar-test, it was additionally found that the results from the *AR* feature differ from the *MV* and *FC* features on a 1% significance level.

The late fusion of the *MFCC* features directly did not perform very well compared to the *MV* and *AR* features. This indicates the necessity of feature integration up to at least a certain time scale before applying a late fusion method.

5. CONCLUSION

The problem of music genre classification addresses many problems and one of these being the identification of useful features. Many short-time features have been proposed in the literature, but only few features have been proposed for longer time scales.

In the current paper, a careful analysis of feature integration and late information fusion has been made with the purpose of music genre classification on longer decision time horizons. Two different data sets were used in combinations with two different classifiers. Additionally, one of the data sets were manually classified in a listening test involving 22 test persons to test the integrity of the data set.

A new feature integration technique, the *AR* model, has been proposed as an alternative to the dominating mean-variance feature integration. Different combinations of the *AR* model and the mean-variance model have been tested, both based on the *MFCC* features. The *AR* model is slightly more computationally demanding, but performs significantly better on the tested data sets. A particularly good result was found with the three-step information fusion of first calculating *MFCC* features, then integrating with the *AR* model and finally using the late fusion technique *sum rule*. This combination gave a classification test error of only 5% on data set 1, as compared to the human classification error of 3%.

6. ACKNOWLEDGEMENTS

The work is supported by the European Commission through the sixth framework IST Network of Excellence: Pattern Analysis, Statistical Modelling and Computational Learning (PASCAL), contract no. 506778.

7. REFERENCES

[1] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, July 2002.

[2] J. Foote and S. Uchihashi, "The beat spectrum: A new approach to rhythm analysis," *Proc. International Conference on Multimedia and Expo (ICME)*, pp. 1088–1091, 2001.

[3] S. H. Srinivasan and M. Kankanhalli, "Harmonicity and dynamics-based features for audio," in *IEEE Proc. of ICASSP*, May 2004, vol. 4, pp. 321–324.

[4] Y. Zhang and J. Zhou, "Audio segmentation based on multi-scale audio classification," in *IEEE Proc. of ICASSP*, May 2004, pp. 349–352.

[5] G. Tzanetakis, *Manipulation, Analysis and Retrieval Systems for Audio Signals*, Ph.D. thesis, Faculty of Princeton University, Department of Computer Science, 2002.

[6] M. F. McKinney and J. Breebaart, "Features for audio and music classification," in *Proc. of ISMIR*, 2003, pp. 151–158.

[7] L. Lu, H-J. Zhang, and H. Jiang, "Content analysis for audio classification and segmentation," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 7, pp. 504–516, October 2002.

[8] P. Ahrendt, A. Meng, and J. Larsen, "Decision time horizon for music genre classification using short-time features," in *Proc. of EUSIPCO*, 2004, pp. 1293–1296.

[9] 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*, vol. ASSP, no. 28, pp. 357–366, August 1980.

[10] C. R. Jankowski, H.-D. Vo, and R. P. Lippmann, "A comparison of signal processing front ends for automatic word recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 3(4), pp. 286–293, 1995.

[11] Kim H.-Gook. and T. Sikora, "Audio spectrum projection based on several basis decomposition algorithms applied to general sound recognition and audio segmentation," in *Proc. of EUSIPCO*, 2004, pp. 1047–1050.

[12] A. C. Harvey, *Forecasting, structural time series models and the Kalman filter*, Cambridge University Press, 1994.

[13] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[14] J. Kittler, M. Hatef, Robert P.W. Duin, and J. Matas, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226–239, 1998.

[15] *www.amazon.com*, "Free-downloads section," 2004.