IMPROVING MUSIC GENRE CLASSIFICATION BY SHORT TIME FEATURE INTEGRATION

Anders Meng, Peter Ahrendt and Jan Larsen
Informatics and Mathematical Modelling
Technical University of Denmark
am,pa,jl@imm.dtu.dk

IMM/ISP

Abstract

- Many different short-time features (derived from $10-30$ms of audio) have been proposed for music segmentation, retrieval and genre classification.
- Often the available time frame of the music to make a decision (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 (early information fusion) and late information fusion (assembling of probabilistic outputs or decisions from the classifier, e.g. majority voting) for music genre classification.
- **A new feature integration technique**, the **AR** model is proposed and seemingly outperforms the commonly used *mean-variance* features.

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 societies share the common need for features which effectively represent the music.
A lot of effort have been put in finding good short time features, however often the decision time horizon is in the order of seconds.
One problem using short time features is that they typically have no perceptual meaning, in contrast to longer time features such as beat and thus makes them difficult to evaluate. In this paper a classification task is applied to evaluate the performance of short time features at different decision time horizons using information fusion such as feature integration.

Features

In this paper features derived at three different time scales are investigated. The perceptual information at these time scales are illustrated in the table below

Timescale	Frame-size	Perceptual meaning
Short time	30ms	instant frequency (harmonics, pitch)
Medium time	740ms	timbre, modulation (instrumentation)
Long time	9.62s	beat, mood, vocal

The short time scale is selected from the stationarity of the audio signal. The other two time scales have been selected to catch the perceptual information shown in the table.

There exist a vast amount of both perceptual and non-perceptual features at the various time-scales. This section will explain the investigated features at the three time scales. Figure 1 illustrates the idea of feature integration.

Short Time Features (30ms)

Mel Frequency Cepstral Coefficients (**MFCC**) was originally derived for use in automatic speech recognition systems, but have been used with great success in audio mining tasks. Earlier results indicate superior performance of these features, see e.g. [1].

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.

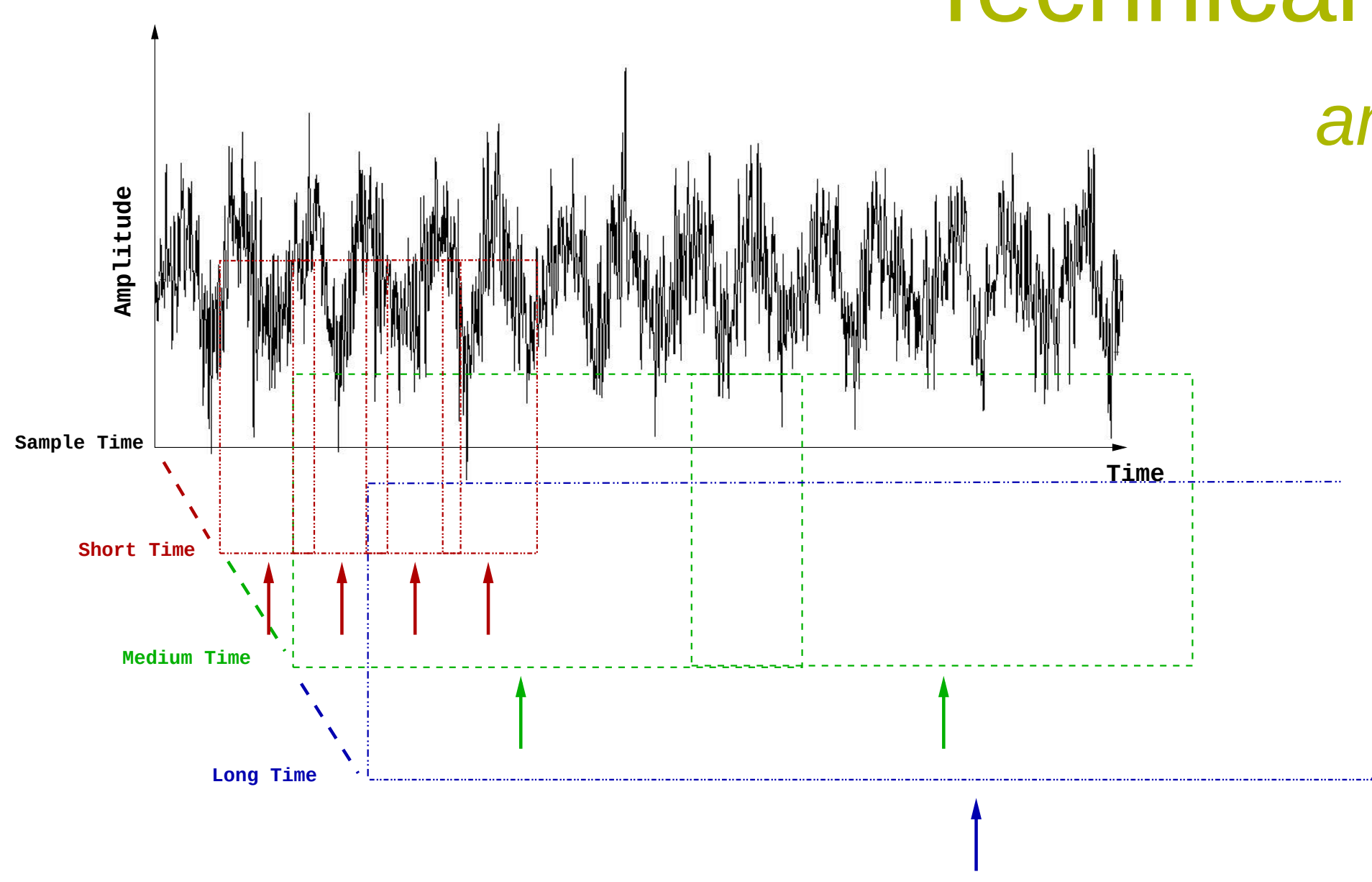

Figure 1: Feature integration. The arrows indicate the new feature at the corresponding time scale. Red, green and blue indicate the three time-scales.

Medium Time Features (740ms)

Mean and Variance (**MV**) are derived directly from the short time features.

Filterbank Coefficients (**FC**) [2] aggregates the power in four frequency bands which are 0Hz) DC, $1-2$Hz) modulation energy, $3-15$Hz) and $20-50$Hz) of each short time feature dimension separately.

High Zero-Crossing Rate Ratio (**HZCRR**) [3] defined as the ratio of number of frames who's time zero crossing rates (ZCR) are above 1.5 times the average ZCR.

Low Short-Time Energy Ratio (**LSTER**) [3] Ratio of the number of frames whose short time energy is less than 0.5 times the average.

Long Time Features (9.62seconds)

Features at this time scale can be derived directly from the audio (perceptual features such as beat) or from short- or medium time features (non-perceptual) by the above mentioned feature integration techniques.

Beat Spectrum (**BS**) [5] The frame similarity have been calculated using the cosine measure between frames. From the similarity matrix the beat-spectrum is derived. The power spectrum of the beat spectrum is aggregated in 6 discriminating bins.

Beat Histogram (**BH**) [4] is another method to determine main beat as well as sub-beats. Instead of utilizing the wavelet transform, octave spaced frequency bins have been used. The resulting beat-histogram is aggregated in 6 discriminating bins.

Autoregressive Model (**AR**)

The autoregressive model which have been used with great success on time series prediction problems is based on the formulation

$$x_n = \sum_{k=1}^{P} a_k x_{n-k} + \mu + u_n, \tag{1}$$

where each dimension of the features on which feature integration is performed upon is modelled separately. Furthermore in this setup $u_n \sim \mathcal{N}(u_n; 0, \sigma^2)$ and μ represents the mean of the time series. The new feature generated consist of the AR coefficients $[a_1, a_2, \ldots, a_P]$, mean value and variance estimate. The **AR** method have some resemblance with the filterbank approach since the power spectrum of each *MFCC* feature is approximated using the **AR** model [6]. Furthermore the number of AR-coefficients controls the number of peaks in the power spectrum.

An overview of the features derived at the different time scales is shown in figure 2.

ICASSP, Philidelphia, March, 2005. Paper #2698

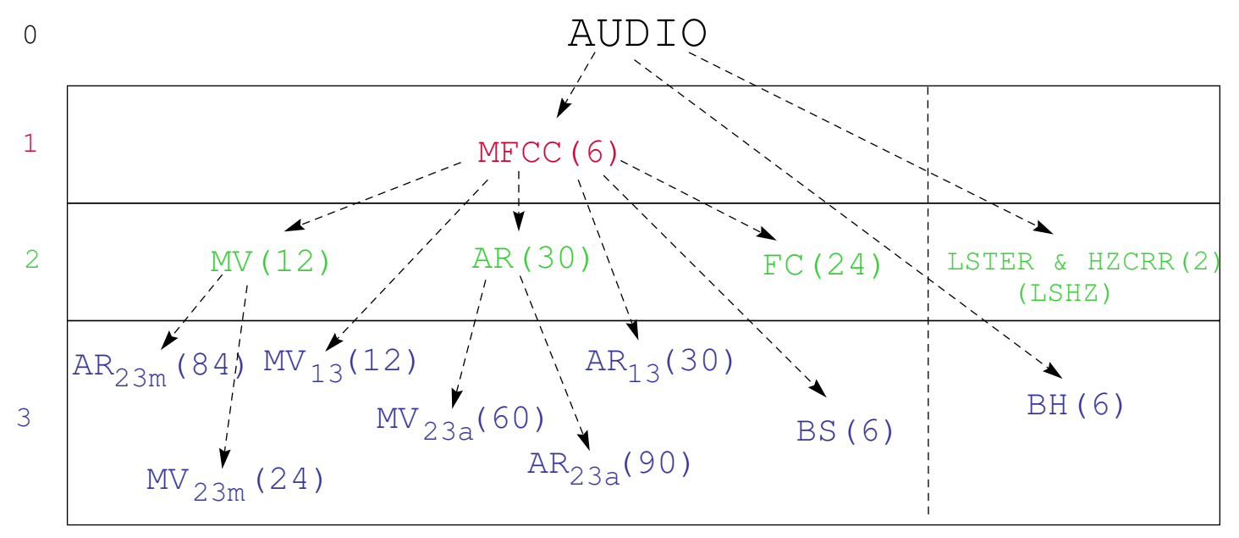

Figure 2: A summary of the applied features in this paper. The arrows indicate which time level the features are derived from and to. The number in parenthesis indicates the dimension of the new feature. E.g. $MV_{23a}(60)$ indicate feature integration from medium to long time scale using *mean-variance* feature integration technique of the **AR** derived features at medium time scale. The resulting dimension is 60.

Experiments

To test the integrity of the proposed **AR** feature integration technique, the various methods were tested in a music genre classification setup, using two different data sets.

Classifiers and information fusion

Two different classifiers were used in this investigation, 1) a linear neural network classifier (LNN) trained with sum of squares loss function and 2) a gaussian classifier (GC) with full covariance matrix. The prior of the classes were known a-priori.

Early information fusion refers to modelling the complex interactions among samples prior to the classifier : *feature integration*.

Late information fusion is the process of combining results provided from the classifier. Three methods was investigated: *majority voting*, *sum-rule* and *median-rule*. Initial studies showed that the *sum-rule* outperformed the other two voting schemes. The sum rule for a decision $D_k \in \{1,..,C\}$ at time k is given as

$$D_k = \max_{i=\{1,..,C\}} \sum_{n=0}^{L} P(c=i|x_n), \tag{2}$$

where C is the number of classes, k the index-variable at the new time-scale, $P(c|x)$ is the posterior probability of class c and L is the number frames which is fused upon.

Data set 1

Consists of 100 songs distributed evenly among *classical, (hard) rock, jazz, pop and techno*. The test set is fixed with 25 music snippets each of 30 seconds. Training set was generated from 15 songs in each genre using 3 music snippets of 30 seconds from each song resulting in a total of 225 music snippets. In each genre 35 of the 45 snippets were randomly selected 10 times to generate test-error bounds. This data set was further subject to a human evaluation by 22 persons which were asked to classify audio pieces at the medium and long time scale into the five genres mentioned above. Their performance are indicated in figure 3a (*human* feature).

Data set 2

This data set consists of 354 songs of each 30 seconds in 6 genres. The data was downloaded from the "Amazon.com Free-downloads database". The 354 songs where evenly distributed among *classical, country, jazz, rap, rock and techno*. In each genre the samples were split into 49 samples for the training- and 10 for a separate test set. Same procedure was applied as above to generate test-error bounds.

Results & Discussion

Figure 3 illustrates the test error on data set 1 and 2 respectively. Each part contains test errors from both the long decision time horizon (10s) and the medium decision time horizon (740ms). The results from both classifiers on the same features have been placed in the same block (GC : Gaussian Classifier) and (LNN : Linear Neural Network). The 95% confidence intervals have been shown for all features. Using the Mcnemar test on a 1% significance level it was found that the **AR** feature integration model differed from the *mean-variance (MV)* and *Filterbank (FC)* approach.

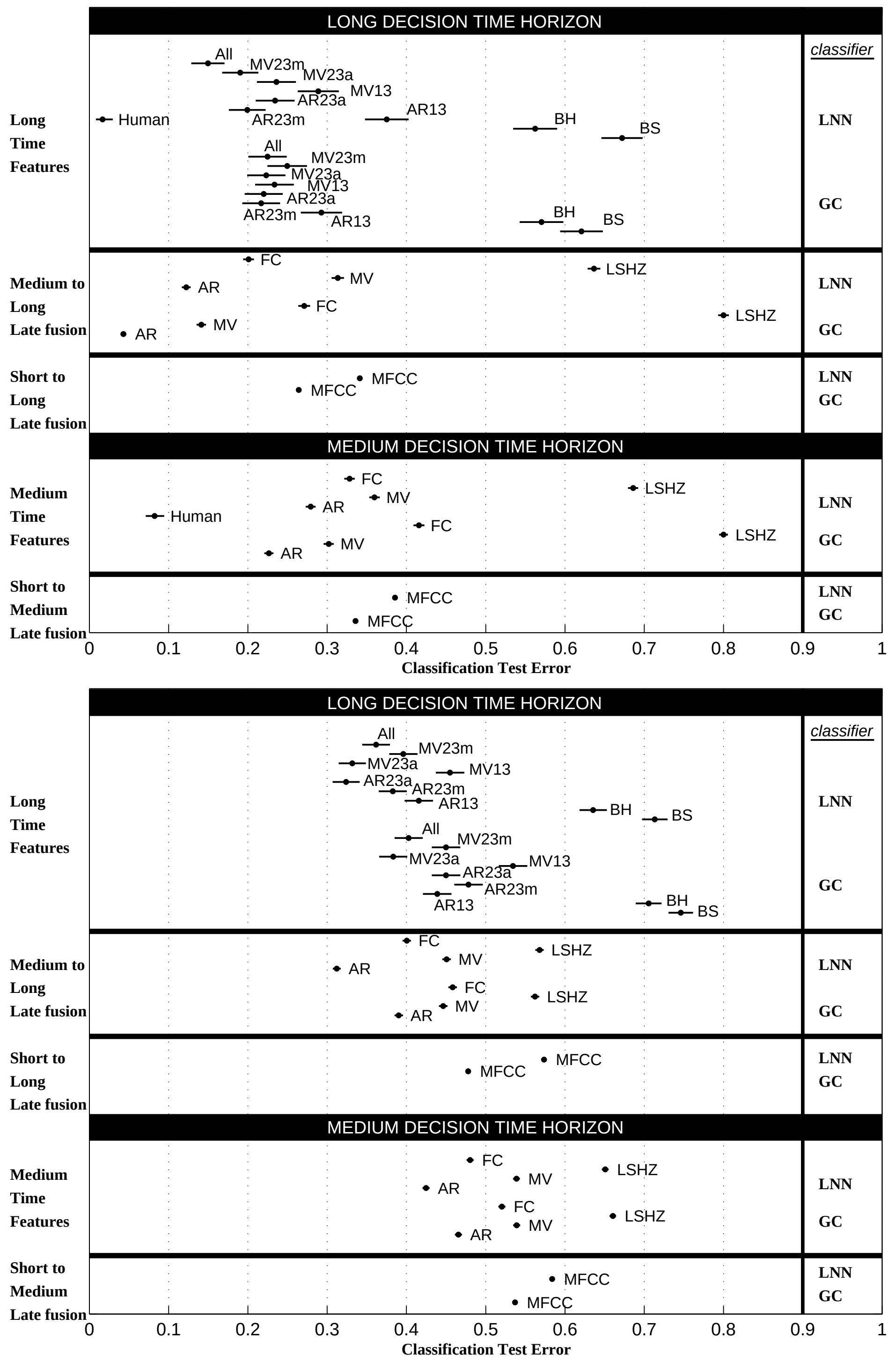

Figure 3: The upper figure shows the experiments on data set 1 and the lower figure the experiments on data set 2.

Conclusion

- This paper carefully investigated different methods with the purpose of music genre classification on longer time scales.
- A new feature integration technique, the **AR** model was suggested, and it performed significantly better on the two data sets.
- Particular good results was found by performing feature integration from short to medium time scale followed by late information fusion from medium to long time scale.

References

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

[2] McKinney, M.F. and Breebart, J.: *Features for audio and music classification,* ISMIR, 2003, pp. 151–158.

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

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

[5] Foote, J. and Uchihashi, S.:*The beat spectrum: A new approach to rhytm analysis,* ICME 2001, pp. 1088–1091.

[6] Makhoul, J. :*Linear Prediction: A Tutorial Review,"* Proceedings of the IEEE, vol. 63, no. 4, 1975, pp. 561–580.