ISMIR 2005

An Investigation of feature models for music genre classification using the support vector classifier

Anders Meng(am@imm.dtu.dk)[1], John Shawe-Taylor(jst@ecs.soton.ac.uk)[2]
[1]Technical University of Denmark, Informatics and Matematical Modelling.
[2]University of Southampton

Introduction

The purpose of this work is two-fold:

1. To investigate the *multivariate Gaussian model* and *multivariate autoregressive model* for modelling short time features (e.g. mel frequency cepstral coefficients) over a segment of audio.
2. Investigate how these two models can be formulated in a kernel framework.

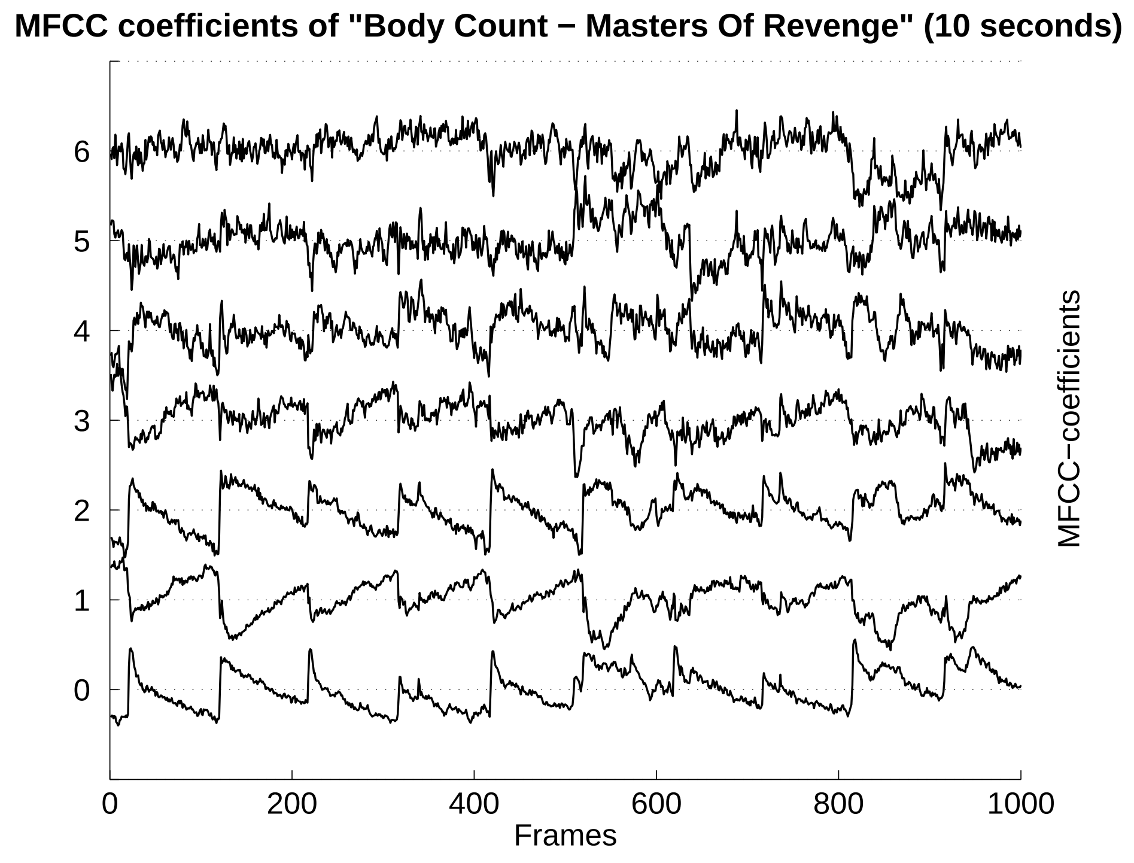

Figure 1. The first seven normalized MFCCs of a 10 second excerpt of the song "Masters of Revenge" by *Body Count*.

An 11 music genre classification problem was investigated using a support vector classifier and a linear neural network.

Two interesting kernels the convolutive kernel and the product probability kernel are promising candidates.

Feature Extraction / Integration

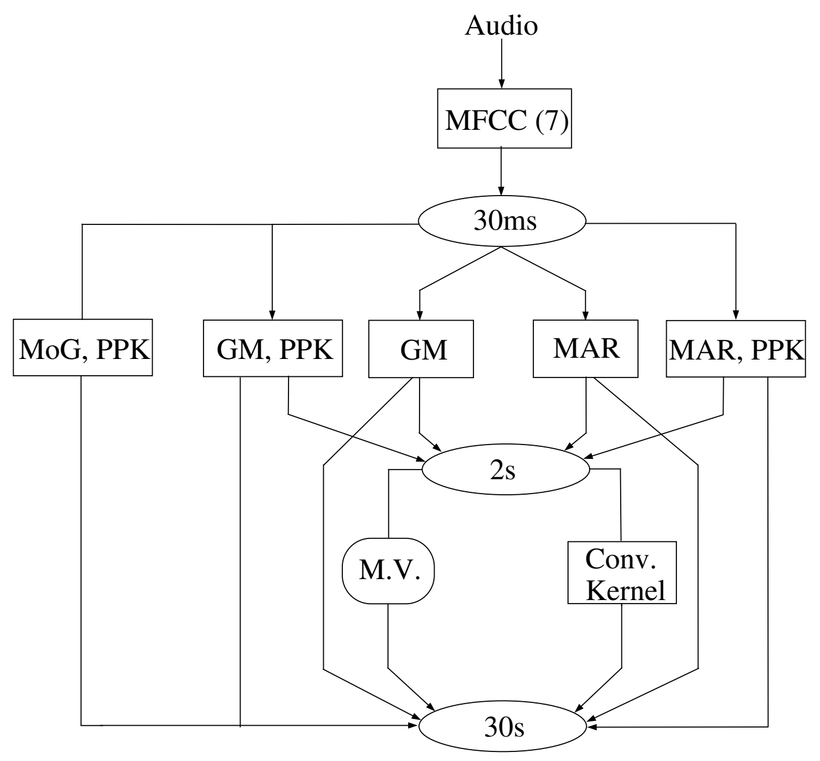

Figure 2. Combinations investigated in the music genre setup. *MAR:* Multivariate AR, *GM:* Multivariate Gaussian, *MoG:* Mixture of Gaussian, *M.V.:* Majority voting, *Conv. Kernel:* Convolution kernel.

Short time features ($\sim 10 - 50$ms)

There exist a wide range of short time features which typically are derived on a $10-50$ms basis. The MFCCs are thought to be good models of the "local" timbre and have with success been applied in various fields of MIR.

Feature integration (> 30ms)

Feature integration is a method for capturing the temporal structure over a segment of short time features.

Multivariate Gaussian model (GM): The mean and covariance are estimated over a segment of MFCC features and used as "new" features.

Multivariate autoregressive model (MAR): To include temporal correlations in the model of short time features, a multivariate AR model is used.

The ordinary multivariate AR model for a sequence of D dimensional short time features $\mathbf{x}_n$, where $n \in \text{segment}$ is given by

$$\mathbf{x}_n = \sum_{p=1}^{K} \mathbf{A}_p \mathbf{x}_{n-p} + \mathbf{u}_n, \qquad (1)$$

where the noise term $(\mathbf{u}_n)$ is assumed Gaussian distributed with mean $\mathbf{v}$ and covariance $\mathbf{C}$, and K is the model order.

Kernels and Classifiers

Linear neural network - linear model (LNN)

The LNN has c outputs (no. of genres) and is trained using a squared loss function. This classifier is fast and robust due to a discriminative training.

Support Vector Classifier (SVC)

The SVM have been applied in various fields of machine learning. It is known for its good performance in high-dimensional spaces unaffected by the curse of dimensionality [1]. The SVM efficiently handles non-linear kernels such as the *convolution kernel* and *product probability kernel*.

The following two kernel functions measure similarity between audio segments. An audio segment is defined as $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_L]$ where L is the segment size.

Convolution Kernel [2] The convolution kernel between to audio segments ($\mathbf{X}$ and $\mathbf{X}'$) is defined as

$$\kappa(\mathbf{X}, \mathbf{X}') = \frac{1}{L^2} \sum_{v=1}^{L} \sum_{v'=1}^{L} \kappa_I \left(\mathbf{x}_v, \mathbf{x}'_{v'} \right), \qquad (2)$$

where $\kappa_I(\mathbf{x}, \mathbf{z})$ must be a valid kernel function.

The Product Probability Kernel (PPK) [3] between two probability densities is defined as

$$\kappa(\theta, \theta') = \int p(\mathbf{x}|\theta)^\rho p(\mathbf{x}|\theta')^\rho d\mathbf{x}, \qquad (3)$$

where $\theta(\theta')$ are the parameters from modelling $\mathbf{X}(\mathbf{X}')$, $\rho > 0$ and $p(\mathbf{x}|\theta)$ is the probabilistic model of the short time features over an audio segment. Typically, $\rho = 1/2$ is used, which normalizes the kernel. Note that latent models (such as the Mixture of Gaussian (MoG) or Hidden Markov Model (HMM)) can be applied in this framework.

Late fusion

The problem of combining the outputs of a classifier. Several fusion techniques exist [4], however, due to the nature of the SVM classifier, *majority voting* was applied.

- Majority voting: the votes received from the classifier are counted and the class with largest amount of votes is selected.

Results

The goal of the music genre classification was to investigate the accuracy of the various methods on a 30s time scale (duration of music snippets).

The Data set consists of 11 music genres distributed evenly among the genres: *Alternative, Country, Easy Listening, Electronica, Jazz, Latin, Pop&Dance, Rap&Hiphop, R&B and Soul, Reggae and Rock*. The data set consists of a training set of 1098 music snippets (30s each), 100 from each genre except for latin and a separate test set of 220 music snippets also distributed evenly among genres.

Parameter tuning The first seven MFCC coefficients were used with a hop- and framesize of 10 and 30ms, respectively. A 3rd order multivariate AR model was found adequate for modelling segments of 2s (intermediate time-scale) and segments of 30s (music snippets). The parameters of the model as well parameters of the SVC were optimized using re-sampling methods on the training set. The LIBSVM package was used for the SVC.

Human accuracy To access the integrity of the data set 9 persons evaluated the data set. The 95% binomial confidence interval was $[46.0 \quad 51.8 \quad 57.7\%]$.

Combinations Quite a few combinations exists to reach a decision at 30s using the two feature integration models (MAR and GM), kernels and majority voting. Figure 2 illustrates some of the combinations investigated.

The two best performing combinations were

- **MARMV** : A MAR model (order 3) is fitted to each 2s of data (using 50% overlap, $\sim$ hopsize 1s) and classification is performed using the LNN. Majority voting is applied to reach a decision at 30s.
- **MARPKK** : A MAR model (order 3) is estimated for each music snippet (30s) and a product probability kernel is generated ($\rho = 1/2$). The kernel is evaluated with a SVC.

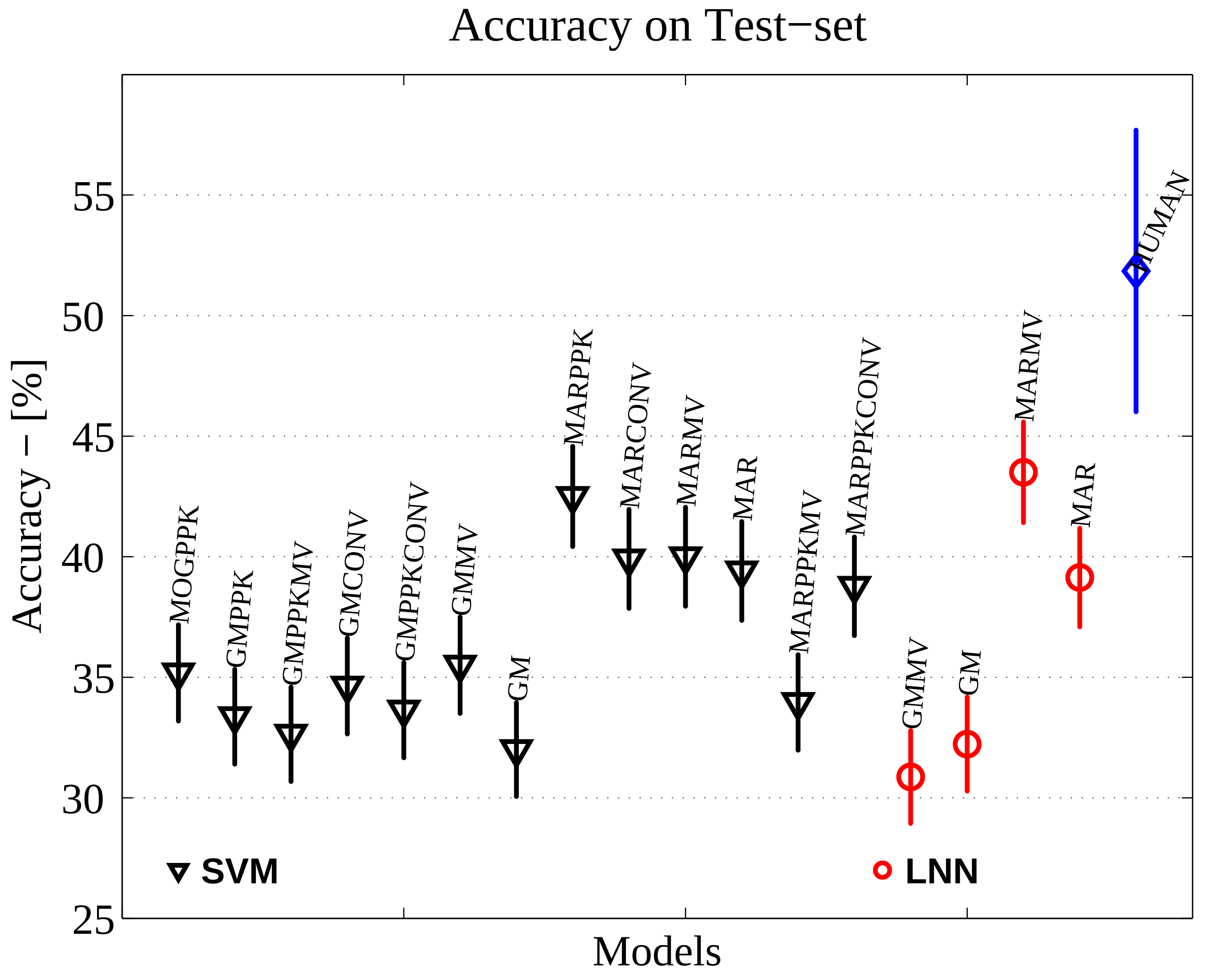

Figure 3. Accuracy on test-set. 95% binomial confidence intervals are also shown.

Performing a Mcnemar test on the MARMV and MARPPK it cannot be rejected that the two models are similar. Thus, instead of classification on a 2s time scale, it is adequate to classify once for each 30s.

Conclusion

- In the music genre classification setup an accuracy of 43% was achieved with the MARPPK method compared with 44% accuracy of the MARMV. The human accuracy of the test set across 9 people was $\sim$ 52%.
- The advantage of the product probability kernel is the possibility of handling complex models of the short time features.
- Future work consists of investigating other probability models which efficiently encode high-dimensional feature spaces.

Acknowledgements

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

References

[1] J. Shawe-Taylor and N. Cristianini. On the generalisation of soft margin algorithms. *IEEE Transactions on Information Theory*, 48(10):2721–2735, 2002.

[2] David Haussler. Convolution kernels on discrete structures. Technical report, University of California at Santa Cruz, July 1999.

[3] T. Jebara, R. Kondor, and A. Howard. Probability product kernels. *Journal of Machine Learning Research*, pages 819–844, July 2004.

[4] J. Kittler, M. Hatef, Robert P.W. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998.

MIREX Contest 2005:

Audio Genre Classification

Peter Ahrendt (pa@imm.dtu.dk) &
Anders Meng (am@imm.dtu.dk)

Figure 4 illustrate the method applied for the *Audio Genre Classification* contest.

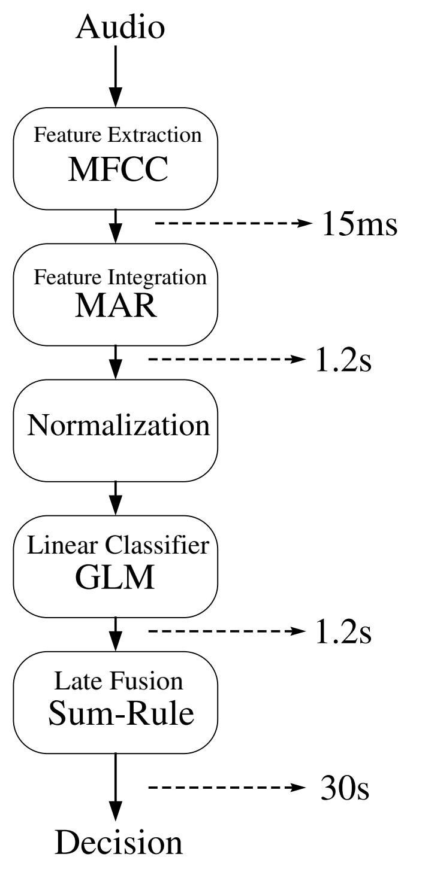

Figure 4. Overview of the MIREX contest setup from audio to decision at 30s.

Short method description

The first six MFCCs were extracted using a hopsize and framesize of 7.5ms and 15ms, respectively. Each sequence of MFCCs were integrated over segments of 1.2s using a 3rd order multivariate autoregressive model (MAR, see equation (1)) with a hopsize of 400ms. The estimated parameters of the MAR model from the 1.2s segment were stacked into a new feature vector. A music snippet of 30s was extracted from the middle of each song, yielding 72 MAR feature vectors per music snippet. Each of the 72 MAR feature vectors were classified as belonging to one of the c genres by a generalized linear model (GLM, the Netlab package was used). To reach a final decision at 30s the sum-rule [4] (related to majority voting) was applied to all the 1.2s decisions.

Nuisance parameters of the classifier and parameters for the MAR model have been preselected from earlier experiments on other databases, thus have not been tuned for the contest data sets. Furthermore, the GLM did not take uneven classes into account.

Results

Two data sets were constructed for the *Audio Genre Classification* contest. One data set was compiled from the *USPop2002* database, consisting of 6 genres and another data set from the *Magnatune* (`www.magnatune.com`) database consisting of 10 genres.

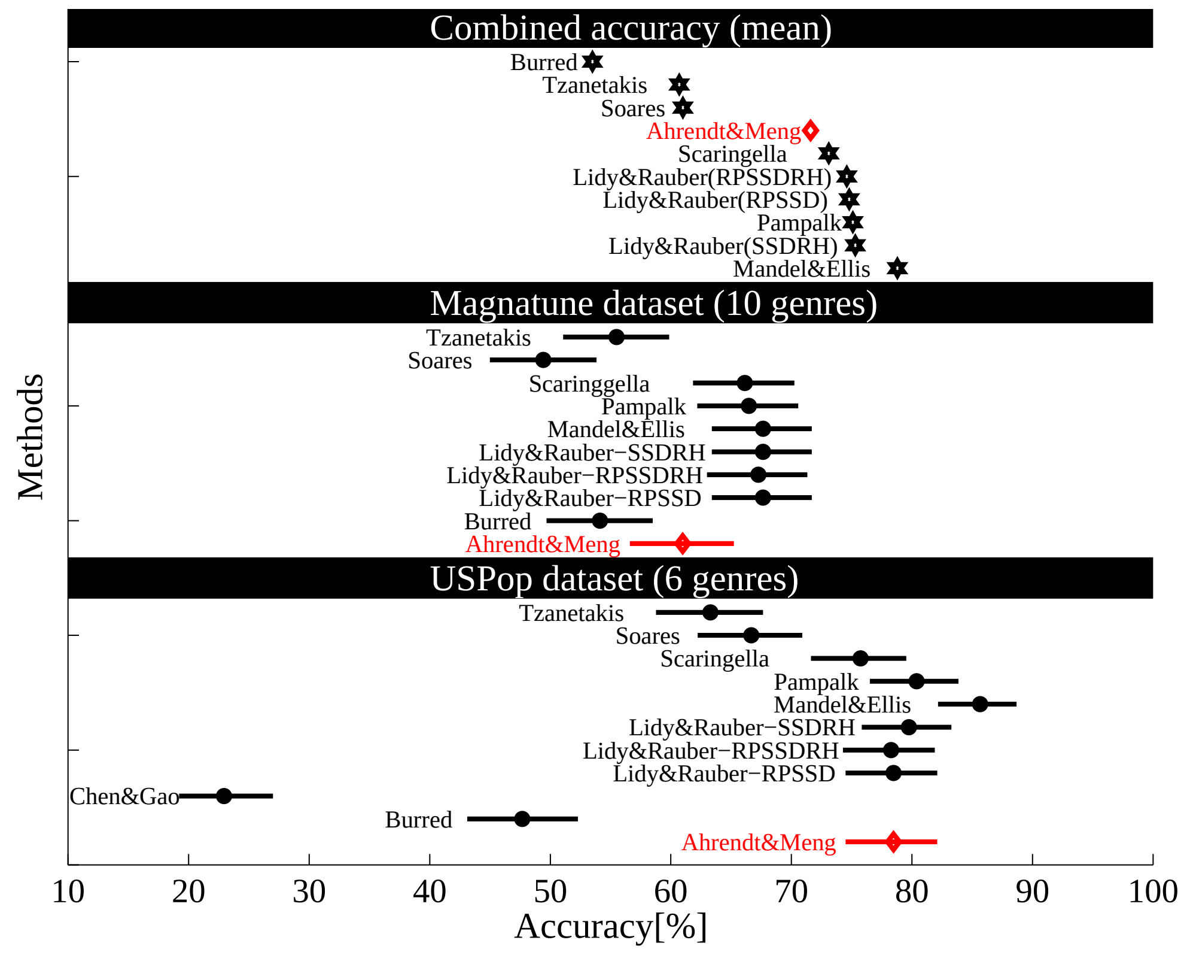

Figure 5. The accuracy of both data sets as well as the combined mean accuracy. The authors algorithm *Ahrendt&Meng* is colored with red (as of the 8th of September).

Conclusion

- The authors method had an accuracy (raw classification accuracy) of 71.6% as compared to the best(as of 8th of September) by *Mandel-Ellis* method which scored an average accuracy of 78.8%.
- It should be noticed though, that only the first 6 MFCCs were used by our method, which states the importance of temporal information in the short time features.

Comments: For a more detailed explanation of the method see the extended abstract, or send a mail to one of the authors.

Acknowledgements: Thanks to the MIREX team for their huge effort in making this contest possible.