Music Genre Classification using the multivariate AR feature integration model

Peter Ahrendt
Technical University of Denmark (IMM)
Building 321, office 120, 2800 Kgs. Lyngby
Denmark
pa@imm.dtu.dk

Anders Meng
Technical University of Denmark (IMM)
Building 321, office 105, 2800 Kgs. Lyngby
Denmark
am@imm.dtu.dk

Keywords: Feature Integration, Multivariate AR, Generalized Linear Classifier

1 INTRODUCTION

Music genre classification systems are normally build as a feature extraction module followed by a classifier. The features are often short-time features with time frames of 10-30ms, although several characteristics of music require larger time scales. Thus, larger time frames are needed to take informative decisions about musical genre. For the MIREX music genre contest several authors derive long time features based either on statistical moments and/or temporal structure in the short time features. In our contribution we model a segment (1.2 s) of short time features (texture) using a multivariate autoregressive model. Other authors have applied simpler statistical models such as the mean-variance model, which also has been included in several of this years MIREX submissions, see e.g. Tzanetakis (2005); Burred (2005); Bergstra et al. (2005); Lidy and Rauber (2005).

2 FEATURES & FEATURE INTEGRATION

The system is designed to handle 22.5kHz mono signals, but could easily be extended to arbitrary sample-rate of the audio signal. Each song is represented by a 30s music snippet taken from the middle of the song. From the raw audio signal the first 6 Mel Frequency Cepstral Coefficients (MFCC) are extracted (including the 0th order coefficient) using a hop- and framesize of 7.5ms and 15ms, respectively. Thus, each song is now represented by a 6 dimensional multivariate time-series. The time series typically display dependency among feature dimensions as well as temporal correlations. Simple statistical moments can be used to characterize important information of the short time features or more elaborate models can be applied. Statistical models which include correlations among feature dimensions as well as time correlations is e.g. the multivariate autoregressive model. Assume that $\mathbf{x}_n$ for $n = 1, \ldots, N$ is the time series of short time features then the multivariate AR model (MAR) can be written as

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

where the noise term $\mathbf{u}_n$ is assumed i.i.d. with zero mean and finite covariance matrix $\mathbf{C}$. The 6 dimensional parameter vector $\mathbf{v}$ is a vector of intercept terms related to the mean of the time series. The $\mathbf{A}_p$'s are the autoregressive coefficient matrices and P denotes the model order. The parameters of the model are estimated using ordinary least squares method and the new feature now consists of elements of $\mathbf{v}$, $\mathbf{C}$ (diagonal + upper triangular part) and $\mathbf{A}_p$ for $p = 1, \ldots, P$. In the actual setup a hopsize of 400ms, framesize of 1200ms and a model order of $P = 3$ results in 72 medium time feature vectors each of dimension 135 ($\mathbf{v} \sim 6, \mathbf{C} \sim 15$ and $A_{1,2,3} \sim 36 * 3 = 108$) for each music snippet. The hopsize, framesize as well as the model order of $P = 3$ have been selected from earlier experiments on other data sets (a-priori information). Thus, not tuned specifically to the unknown data sets in contest. To avoid numerical problems in the classifier each feature dimension of the MAR features is normalized to unit variance and zero mean. The normalization constants for each dimension are calculated from the training set.

3 CLASSIFIER

A generalized linear model (GLM), Bishop (1995), with softmax activation function is trained on all the MAR-feature vectors from all the songs. This classifier is simply an extension of a logistic regression classifier to more than two classes. It has the advantage of being discriminative, which makes it more robust to non-equal classes. Furthermore, since it is a linear model it is less prone to overfitting (as compared to a generative model). Each frame of size 1200ms is classified as belonging to one of c classes, where c is the total number of music genres. In the actual implementation the *Netlab* package was used, see `http://www.ncrg.aston.ac.uk/netlab/` for more details.

3.1 Late information fusion

To reach a final decision for a 30s music clip the sum-rule, Kittler et al. (1998), is used over all the frames in the music clip. The sum-rule assigns a class as

$$\hat{c} = \arg\max_{c} \sum_{r=1}^{n_f} P(c|\mathbf{x}_r) \tag{2}$$

where r and n_f is the frame index and number of frames of the music clip, respectively, and $P(c|\mathbf{x}_r)$ is the estimated posterior probability of class c given the MAR feature vector $\mathbf{x}_r$. As mentioned earlier $n_f = 72$ frames for each music clip.

Figure 1 shows the full system setup of the music genre classification task from the raw audio to a decision on genre of each music snippet.

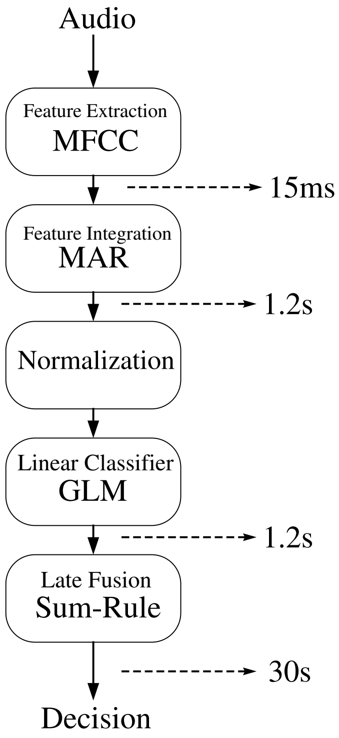


Figure 1: Overview of system from audio to a genre decision at 30s. The time scale at each step is indicated to the right.

4 CONTEST RESULTS

This years *Audio Genre Classification* contest consisted of two audio databases

- *USPop* (single level genre), `http://www.ee.columbia.edu/~dpwe/research/musicsim/uspop2002.html`
- *Magnatune* (hierarchical genre taxonomy) `www.magnatune.com`

from which two independent data sets were compiled. Originally, a third database, *Epitonic* (`http://www.epitonic.com`), was proposed, but due to lack of time only the first two databases were investigated.

The first data set was generated from the USPop database and consisted of a training set of 940 music files distributed un-evenly among 6 genres (Country, Electronica/Dance, Newage, Rap/Hiphop, Reggae and Rock) and a test set of 474 music files. The second data set was generated from the Magnatune database with a training/test set of $1005/510$ music files distributed un-evenly among the 10 genres: Ambient, Blues, Classical, Electronic, Ethnic, Folk, Jazz, Newage, Punk and Rock.

4.1 Parameter optimization

The various parameters of both the feature extraction and integration step as well as nuisance parameters for the GLM classifier were preselected, and therefore not tuned to the specific data sets. Cross-validation or an approximative approach could have been utilized in order to optimize the values of the classifier and feature extraction/integration step.

4.2 Results & Discussion

Figure 2 shows the raw mean classification accuracy of both data sets of the methods, which completed within the 24 hour time limit (8th of September). A 95% binomial confidence interval was applied on each method to illustrate the possible variation in mean value. Our algorithm, denoted as *Ahrendt&Meng*, shows a mean accuracy of 60.98% for uncorrected classes on the Magnatune data set and a mean accuracy of 78.48% on the USpop data set. Our method showed a mean accuracy of 71.55% when averaging across data sets compared with the best performing method of 78.8% by *Mandel&Ellis*. There is several observations, which can be made from this years contest. Our model is solely based on the first 6 MFCCs, which subsequently are modelled by a multivariate autoregressive model, hence, the temporal structure is modelled. The best performing method in this years contest is by Mandel and Ellis (2005) (8th of September), see

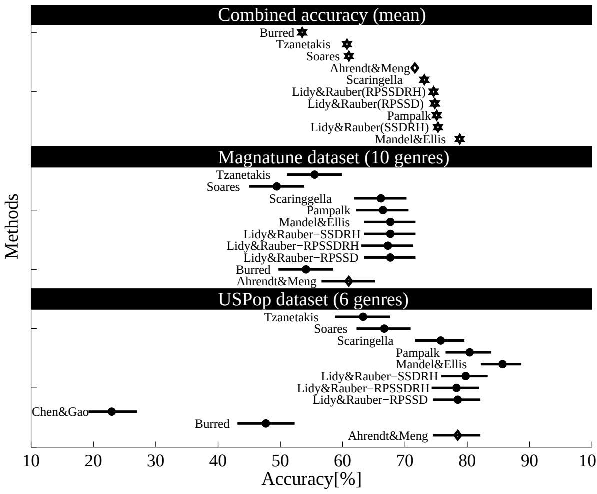
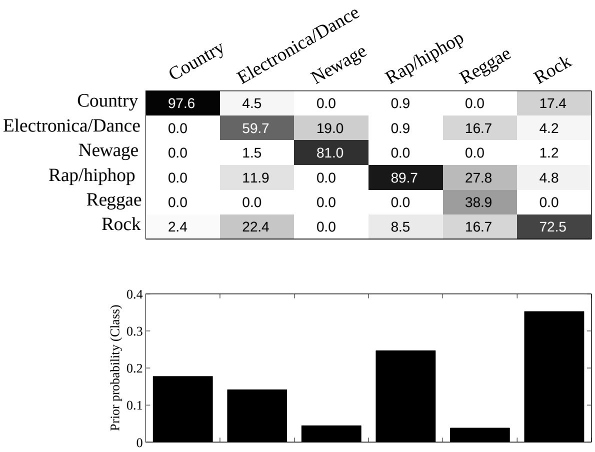


Figure 2: Mean accuracy on both USPop and Magnatune data sets illustrated with a 95% binomial confidence interval. The ”Combined accuracy” is the mean accuracy on the two data sets.

figure 2). Their approach consist of extracting the first 20 MFCCs and then model the MFCCs of the entire song by a multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. This model is then used in a modified KL-divergence kernel, from which a support vector classifier can be applied. Since the mean and covariance are static components no temporal information is modelled in this approach, however, good results were observed. Even better results might have been achieved by using models, which include temporal information.

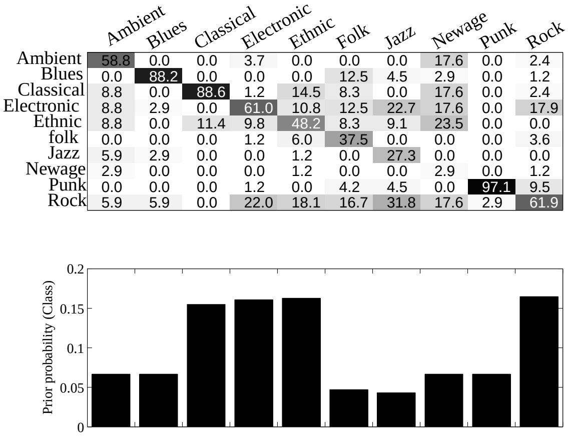
In order to make a proper statistical comparison of the different methods the raw classifications should have been known.



	Country	Electronica/Dance	Newage	Rap/hiphop	Reggae	Rock
Country	97.6	4.5	0.0	0.9	0.0	17.4
Electronica/Dance	0.0	59.7	19.0	0.9	16.7	4.2
Newage	0.0	1.5	81.0	0.0	0.0	1.2
Rap/hiphop	0.0	11.9	0.0	89.7	27.8	4.8
Reggae	0.0	0.0	0.0	0.0	38.9	0.0
Rock	2.4	22.4	0.0	8.5	16.7	72.5

Figure 3: *Upper:* Confusion matrix (accuracy) of proposed method on the USPop data set. *Lower:* The prior probabilities of the genres.

The upper figure of figure 3 and 4 shows the confusion matrix of our method on the USPop and Magnatune data set, respectively. The lower figures shows the prior probability on the genres calculated from the test sets. The true genre is shown along the horizontal axis. The confusion matrix on the Magnatune data set illustrates that our method provides reasonable predictive power of *Punk, Classical and Blues*, whereas *Newage* is actually below a random guessing of 2.9%.



	Ambient	Blues	Classical	Electronic	Ethnic	Folk	Jazz	Newage	Punk	Rock
Ambient	58.8	0.0	0.0	3.7	0.0	0.0	0.0	17.6	0.0	2.4
Blues	0.0	88.2	0.0	0.0	0.0	12.5	4.5	2.9	0.0	1.2
Classical	8.8	0.0	88.6	1.2	14.5	8.3	0.0	17.6	0.0	2.4
Electronic	8.8	2.9	0.0	61.0	10.8	12.5	22.7	17.6	0.0	17.9
Ethnic	8.8	0.0	11.4	9.8	48.2	8.3	9.1	23.5	0.0	0.0
folk	0.0	0.0	0.0	1.2	6.0	37.5	0.0	0.0	0.0	3.6
Jazz	5.9	2.9	0.0	0.0	1.2	0.0	27.3	0.0	0.0	0.0
Newage	2.9	0.0	0.0	0.0	1.2	0.0	0.0	2.9	0.0	1.2
Punk	0.0	0.0	0.0	1.2	0.0	4.2	4.5	0.0	97.1	9.5
Rock	5.9	5.9	0.0	22.0	18.1	16.7	31.8	17.6	2.9	61.9

Figure 4: *Upper:* Confusion matrix (accuracy) of proposed method on the Magnatune data set. *Lower:* The prior probabilities of the genres.

5 CONCLUSION & DISCUSSION

A mean accuracy over the two data sets of 71.6% was achieved using only the first 6 MFCCs as compared to a mean accuracy of 78.8% by Mandel and Ellis (2005) (8th of September) using the first 20 MFCCs. A further performance increase could have been achieved by optimizing nuisance parameters of the classifier and by correcting for uneven classes. Furthermore, the model order of the multivariate autoregressive model could have been optimized using cross-validation on the training set. Future perspectives would be to use a support vector classifier, which would alleviate problems of overfitting. The approach presented in this extended abstract could easily have been applied in the *Audio Artist Identification* contest as well.

References

J. Bergstra, N. Casagrande, and D. Eck. Music genre classification, mirex contests, 2005.

C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

Juan Jose Burred. Music genre classification, mirex contests, 2005.

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.

T. Lidy and A. Rauber. Music genre classification, mirex contests, 2005.

Michael Mandel and Daniel Ellis. Music genre classification, mirex contests, 2005.

George Tzanetakis. Music genre classification, mirex contests, 2005.