Temporal feature integration for music genre classification

Meng, Anders; Ahrendt, Peter; Larsen, Jan; Hansen, Lars Kai

Published in:
I E E E Transactions on Audio, Speech and Language Processing

Link to article, DOI:
10.1109/TASL.2007.899293

Publication date:
2007

Document Version
Publisher's PDF, also known as Version of record

Temporal Feature Integration for Music Genre Classification

Anders Meng, Peter Ahrendt, Jan Larsen, *Senior Member, IEEE*, and Lars Kai Hansen

***Abstract*—Temporal feature integration is the process of combining all the feature vectors in a time window into a single feature vector in order to capture the relevant temporal information in the window. The mean and variance along the temporal dimension are often used for temporal feature integration, but they capture neither the temporal dynamics nor dependencies among the individual feature dimensions. Here, a multivariate autoregressive feature model is proposed to solve this problem for music genre classification. This model gives two different feature sets, the diagonal autoregressive (DAR) and multivariate autoregressive (MAR) features which are compared against the baseline mean-variance as well as two other temporal feature integration techniques. Reproducibility in performance ranking of temporal feature integration methods were demonstrated using two data sets with five and eleven music genres, and by using four different classification schemes. The methods were further compared to human performance. The proposed MAR features perform better than the other features at the cost of increased computational complexity.**

***Index Terms*—Autoregressive (AR) model, music genre classification, temporal feature integration.**

I. Introduction

In RECENT years, there has been an increasing interest in the research area of music information retrieval (MIR). This is spawned by the new possibilities on the Internet such as online music stores like Apple's iTunes and the enhanced capabilities of ordinary computers. The related topic of music genre classification can be defined as computer-assigned genre labeling of sound clips. It has received much attention in its own right, but it is also often used as a good test-bench for music features in related areas where the labels are harder to obtain than the musical genres. An example of this is in [1], where rhythm features are assessed in a music genre classification task.

Music genre classification systems normally consist of feature extraction from the digitized music, followed by a classifier that uses features to estimate the genre. In this paper, we focus on identifying temporal feature integration methods which give consistent and good performance over different data sets and choices of classifier. A review of music genre classification systems is given in [2].

In several feature extraction models, perceptual characteristics such as the beat [3] or pitch [4] are modeled directly. This has the clear advantage of giving features which can be examined directly without the need of a classifier. However, most of the previous research has concentrated on short-time features, e.g., audio spectrum envelope and the zero-crossing rate [5] which are extracted from 20–40 ms frames of the sound clip. Such features are thought to represent perceptually relevant characteristics such as, e.g., music roughness or timbre. They have to be evaluated as part of a full classification system. A sound clip is thus represented by a multivariate time series of these features and different methods exist to combine this information into a single genre label for the whole sound clip. An example is in [6], based on a hidden Markov model of the time series of the cepstral coefficient features.

Temporal feature integration is another approach to combine information. It uses a sequence of short-time feature vectors to create a single new feature vector at a larger time scale. It assumes a minimal loss of the important temporal information for music genre classification in the short-time feature extraction stage. Temporal feature integration is a very common technique. Often basic statistic estimates like the mean and variance of the short-time features have been used [4], [7], [8].

Here, a new multivariate autoregressive temporal feature integration model is proposed as an alternative to the mean-variance feature set. The main advantage of the autoregressive model is its ability to model temporal dynamics as well as dependencies among the short-time feature dimensions. In fact, the model is a natural generalization of the mean-variance temporal feature integration model.

This paper provides an extension of our work in [9] at several levels. In [9], each short-time feature dimension was modeled independently; hence, dependencies among the feature dimensions were not modeled. In this paper, these are modeled in terms of correlation by applying the multivariate autoregressive model (MAR). Furthermore, in this paper, we give a more detailed explanation of the autoregressive model and its relation to the other investigated temporal feature integration methods. Computational complexity has been included for the different methods and in the experimental section, more classifiers and a data set with larger complexity have been added.

In Section IV, we will compare three temporal feature integration methods typically applied in the literature; the mean-variance (MeanVar), mean-covariance (MeanCov), filter bank

Manuscript received August 1, 2006; revised February 14, 2007. This work was supported in part by the European Commission through the sixth framework IST Network of Excellence: Pattern Analysis, Statistical Modeling, and Computational Learning (PASCAL) under Contract 506778 and in part by the Danish Technical Research Council under Project 26-04-0092 Intelligent Sound (www.intelligentsound.org). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Malcolm Slaney.

A. Meng, J. Larsen, and L. K. Hansen are with Informatics and Mathematical Modeling (IMM), Technical University of Denmark, Lyngby DK-2800, Denmark (e-mail: am@imm.dtu.dk; pea@widex.com; jl@imm.dtu.dk; lkh@imm.dtu.dk).

P. Ahrendt is with Widex, 3500 Vaerloese, Denmark.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2007.899293

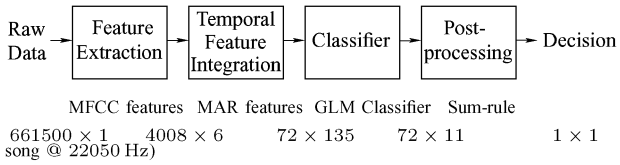

Fig. 1. Full music genre classification system. The flow-chart illustrates the different parts of the system, whereas the names just below the chart are the specific choices that gives the best performing system. The numbers in the bottom part of the figure illustrate the (large) dimensionality reduction that takes place in the system (the number of genres is 11).

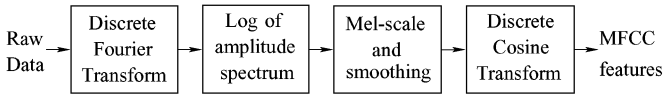

Fig. 2. Mel frequency cepstral coefficients feature extraction as described in [14].

coefficients (FC) with the proposed autoregressive models: the diagonal autoregressive (DAR) and MAR model. To generalize the result two different data sets consisting of five and 11 different genres have been used in the experiments. Both data sets have been evaluated by a group of persons to relate the obtained accuracies by the different automated methods. Furthermore, to ensure a fair comparison of the different temporal feature integration methods, four different classifiers have been applied; a linear model (LM), a generalized linear model (GLM), a Gaussian classifier (GC) and a Gaussian mixture model (GMM) classifer.

Fig. 1 illustrates the full music genre classification system that was used for evaluating the temporal feature integration methods.

Section II describes common feature extraction and integration methods, while Section III gives a detailed explanation of the proposed multivariate autoregressive feature model. Section IV reports and discusses the results of experiments that compare the newly proposed features with the best of the existing temporal feature integration methods. Finally, Section V concludes on the results.

II. Feature Extraction and Integration

Several different features have been suggested in music genre classification. The general idea is to process fixed-size time windows of the digitized audio signal with an algorithm which can extract relevant information in the sound clip. The size of the windows gives the time scale of the feature. The features are often thought to represent aspects of the music such as the pitch, instrumentation, harmonicity, or rhythm.

The following subsections explain popular feature extraction methods. They are listed on the basis of their time scale. The process of temporal feature integration is explained in detail in the end of the section.

A. Short-Time Features

Most of the features that have been proposed in the literature are short-time features, which usually employ frame sizes of 20–40 ms. They are often based on a transformation to the spectral domain using techniques such as the short-time Fourier transform. The assumption in these spectral representations is (short-time) stationarity of the signal which means that the frame size has to be small.

In [5], we found the so-called *Mel-frequency cepstral coefficient* (MFCC) to be very successful. Similar findings were observed in [10] and [11]. They were originally developed for speech processing [12]. The details of the MFCC feature extraction are shown in Fig. 2. It should be mentioned, however, that other slightly different MFCC feature extraction schemes exist, see, e.g., [13].

In the experiments, we used the *VOICEBOX*, written by Brookes, to extract the MFCC features.[1]

According to [15], short-time representations of the full time-frequency domain, such as the MFCC features, can be seen as models of the music timbre.

There exists many other short-time feature extraction methods, see, e.g., [16] and [17].

B. Medium-Time Features

Medium-time features are here defined as features, which are extracted on time scales around 1000–2000 ms. [4] uses the term "texture window" for this time scale where important aspects of the music "lives" such as note changes and tremolo [18]. Examples of features for this time scale are the low short-time energy ratio (LSTER) and high zero-crossing rate ratio (HZCRR) [19].

C. Long-Time Features

Long-time features describe important statistics of, e.g., a full song or a larger sound clip. An example is the beat histogram feature [20] which summarize the beat content in a sound clip.

D. Temporal Feature Integration

Temporal feature integration is the process of combining all the feature vectors in a time window into a single feature vector that captures the temporal information of this window. The new features generated do not necessarily capture any explicit perceptual meaning such as perceptual beat or mood, but captures information which are useful for the subsequent classifier. In [3], the "beat-spectrum" is used for music retrieval by rhythmic similarity. The beat-spectrum can be derived from short-time features such as the STFT or MFCCs as noted in [3]. This clearly indicates that the evolution of the short-time features contain important temporal information. Fig. 3 shows the first six MFCCs of a 10-s excerpt of the music piece "Masters of Revenge" by "Body Count." This example shows a clear repetitive structure in the short-time features. Another important property of temporal feature integration is data reduction. Consider a 4-min piece of music represented as short-time features (using the first six MFCCs). With a hop- and frame-size of 10 and 20 ms, respectively, this results in approximately 288 kB of data using a 16-bit representation of the features. The hop-size is defined as the frame-size minus the amount of overlap between frames and specifies the "effective sampling rate" of the features. This is a rather good compression compared to the original size of the music (3.84 MB, MPEG1-layer 3 at 128 kb/s). However, if

[1][Online]. Available: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html.

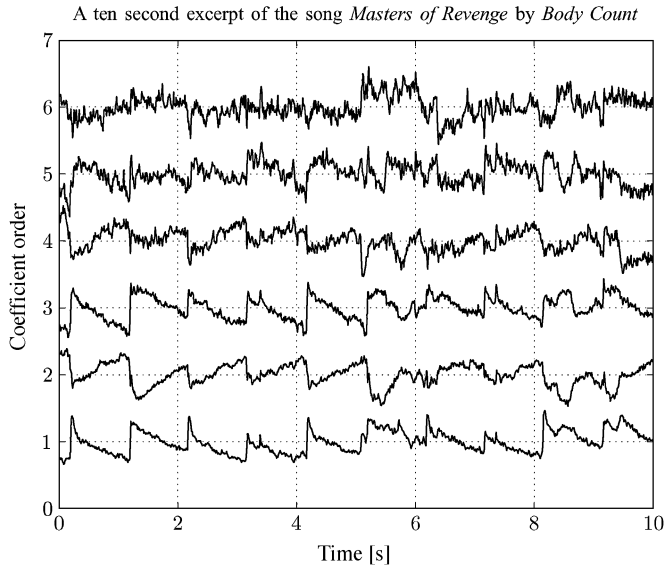

Fig. 3. First six normalized MFCCs of a 10-S snippet of "Body Count—Masters of Revenge." The temporal correlations are very clear from this piece of music as well as the cross correlations among the feature dimensions. This suggests that relevant information is present and could be extracted by selecting a proper temporal feature integration model.

the relevant information can be summarized more efficiently in less space, this must be preferred.

The idea of temporal feature integration can be expressed more rigorously by observing a sequence of consecutive short-time features, $\mathbf{x}_i$ of dimension D, where i represents the ith short-time feature. These are integrated into a new feature $\mathbf{z}_k$ of dimension M

$$\mathbf{z}_k = \mathbf{f}\left(\mathbf{x}_{k \cdot H_s}, \ldots, \mathbf{x}_{k \cdot H_s + W_s - 1}\right) \tag{1}$$

where H_s is the hop-size and W_s window-size (both defined in number of samples) and $k = 0, 1, \ldots,$ is the discrete time index of the larger time scale. There exists a lot of different models, here denoted by $\mathbf{f}(\cdot)$, which map a sequence of short-time features into a new feature vector.

A very simple temporal feature integration method is that of stacking consequtive short-time features, see, e.g., [5] and [6] into a new feature vector and thereby maintaining information. This method requires a robust machine learning algorithm to cope with the often high-dimensional feature vectors created, and does not introduce any sort of compression. In the following, the MeanVar, MeanCov, and FCs will be discussed. These methods have been suggested for temporal feature integration in the literature. All of these methods introduce a good level of compression compared to that of stacking. Furthermore, as will become clear, these methods have closed form solutions.

1) Gaussian Model: A very simple model for temporal feature integration is the so-called MeanVar model which has been used in work related to music genre classification, see, e.g., [9] and [20]. This model implicitly assumes that consecutive samples of short-time features are independent and Gaussian distributed and, furthermore, that each feature dimension is independent. Using maximum-likelihood, the parameters for this model are estimated as

$$\begin{aligned} \mathbf{m}_k &= \frac{1}{W_s} \sum_{n=0}^{W_s - 1} \mathbf{x}_{k \cdot H_s + n} \\ c_{i,k} &= \frac{1}{W_s} \sum_{n=0}^{W_s - 1} \left(\mathbf{x}_{i, kH_s + n} - m_{i,k}\right)^2 \end{aligned} \tag{2}$$

for $i = 1, \ldots, D$, which results in the following feature at the new time scale

$$\mathbf{z}_k = \begin{bmatrix} \mathbf{m}_k \\ \mathbf{c}_k \end{bmatrix} \tag{3}$$

where $\mathbf{z}_k$ is of dimension $2D$ and $\mathbf{m}$, and $\mathbf{c}$ are the estimated mean and variance of the short-time features. As seen in Fig. 3, the assumption that each feature dimension is independent is not correct. A more reasonable temporal feature integration model is the multivariate Gaussian model, denoted in the experimental section as MeanCov, where correlations among features are modeled. This model of the short-time features can be formulated as $\mathbf{x} \sim N(\mathbf{m}, \mathbf{C})$, where the mean and covariance are calculated over the given window of short-time features. Thus, the diagonal of $\mathbf{C}$ contains the variance features from MeanVar. The mean vector and covariance matrix are stacked into a new feature vector $\mathbf{z}_k$ of dimension $(D/2)(3 + D)$

$$\mathbf{z}_k = \begin{bmatrix} \mathbf{m}_k \\ \mathrm{vech}(\mathbf{C}_k) \end{bmatrix} \tag{4}$$

where $\mathrm{vech}(\mathbf{C})$ refers to stacking the upper triangular part of the matrix including the diagonal.

One of the drawbacks of the Gaussian model, whether this is the simple (MeanVar) or the multivariate model (MeanCov), is that temporal dependencies in the data are not modeled.

2) Filter Bank Coefficients (FC): The filter bank approach considered in [21] aims at capturing some of the dynamics in the sequence of short-time features. They investigated the method in a general audio and music genre classification task. The idea is to extract a summarized power of each feature dimension independently in four specified frequency bands. The temporal feature integration function $\mathbf{f}(\cdot)$ for the filter bank approach can be written compactly as

$$\mathbf{z}_k = \mathrm{vec}(\mathbf{P}_k \mathbf{W}) \tag{5}$$

where $\mathbf{W}$ is a filter matrix of dimension $N \times 4$, and $\mathbf{P}_k$ contains the periodograms of each short-time feature and has dimension $D \times N$, where $N = W_s/2$ when W_s is even and $N = (W_s - 1)/2$ for odd values.

The four frequency bands in which the periodograms are summarized are specified in the matrix $\mathbf{W}$. In [21], the four filters applied to handle the short-time features are: 1) a dc-filter; 2) 1–2 Hz modulation energy; 3) 3–15 Hz modulation energy; 4) 20–43 Hz perceptual roughness [22].

The advantage of this method is that the temporal structure of the short-time features is taken into account; however, correlations among feature dimensions are not modeled. In order to model these, cross-correlation spectra would be required.

III. Multivariate Autoregressive Model for Temporal Feature Integration

The simple mean-variance model does not model temporal feature correlations; however, these features have shown to perform remarkably well in various areas of music information retrieval, see, e.g., [20] and [23]. The dependencies among features could be modeled using the MeanCov model, but still do not model the temporal correlations. The FC approach includes temporal information in the integrated features, but the correlations among features are neglected.

This section will focus on the MAR for temporal feature integration, since it has the potential of modeling both temporal correlations and dependencies among features.

For simplicity, we will first study the DAR. The DAR model assumes independence among feature dimensions similar to the MeanVar and FC feature integration approaches. The full multivariate autoregressive model (MAR) in considered in Section III-B.

A. DAR

The DAR model was investigated in [9], where different temporal feature integration methods were tested and showed improved performance compared to the MeanVar and FC approaches; however, the theory behind the model was not fully covered. For completeness, we will here present a more detailed description of the model.

Assuming independence among feature dimensions, the Pth order causal autoregressive model[2] for each feature dimension can be written as

$$x_n = \sum_{p=1}^{P} a_p x_{n-p} + G u_n \tag{6}$$

for $n = 0, \ldots, W_s - 1$, where a_p for $p = 1, \ldots, P$ is the autoregressive coefficients, u_n is the noise term, assumed independent and identically distributed (i.i.d.) with unit variance and mean value v. G sets the scale of the noise term. Note that the mean value of the noise process v is related to the mean m of the time series by $m = \left(1 - \sum_{p=1}^{P} a_p\right)^{-1} v$.

Equation (6) expresses the "output" x_n as a linear function of past outputs and present inputs u_n. There are several methods for estimating the parameters of the autoregressive model, either in the frequency domain [24] or directly in time-domain [25]. The most obvious and well-known method is the ordinary least squares method, where the mean squared error is minimized. Other methods suggested are the generalized (or weighted) least squares where the noise process is allowed to be colored. In our case, the noise process is assumed white; therefore, the least squares method is applied and described in the following. The prediction of a new sample based on estimated parameters a_p becomes

$$\tilde{x}_n = \sum_{p=1}^{P} a_p x_{n-p} \tag{7}$$

[2]In the speech community, this is known as a linear predictive coding (LPC) model, however, here applied to a sequence of short-time features instead of the raw sound signal.

and the error signal e_n measured between $\tilde{x}_n$ and x_n is

$$e_n = x_n - \tilde{x}_n = x_n - \sum_{p=1}^{P} a_p x_{n-p} \tag{8}$$

where e_n is known as the residual. Taking the z-transformation on both sides of (8), the error can now be written as

$$E(z) = \left(1 - \sum_{p=1}^{P} a_p z^{-p}\right) X(z) = A(z) X(z). \tag{9}$$

In the following, we will switch to frequency representation $z = e^{j\omega}$ and in functions use $X(\omega)$ for representing $X(e^{j\omega})$. Assuming a finite energy signal x_n, the total error to be minimized in the ordinary least squares method $\mathcal{E}_{\text{tot}}$ is then according to Parseval's theorem given by

$$\mathcal{E}_{\text{tot}} = \sum_{n=0}^{W_s - 1} e_n^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} |E(\omega)|^2 \, d\omega. \tag{10}$$

To understand why this model is worthwhile to consider, we will now explain the spectral matching capabilities of the model. First, we look at the model from (6) in the z-transformed domain, which can now be described as

$$X(z) = \sum_{p=1}^{P} a_p X(z) z^{-p} + G U(z) \tag{11}$$

where $v = 0$ is assumed without loss of generalizability. The system transfer function becomes

$$H(z) \equiv \frac{X(z)}{U(z)} = \frac{G}{1 - \sum_{p=1}^{P} a_p z^{-p}} \tag{12}$$

and its corresponding model power spectrum

$$\hat{P}(\omega) = |H(\omega) U(\omega)|^2 = |H(\omega)|^2 = \frac{G^2}{|A(\omega)|^2}. \tag{13}$$

Combining the information in (9), (10), and (13) and the fact that $P(\omega) = |X(\omega)|^2$, the total error to be minimized can be written as

$$\mathcal{E}_{\text{tot}} = \frac{G^2}{2\pi} \int_{-\pi}^{\pi} \frac{P(\omega)}{\hat{P}(\omega)} d\omega. \tag{14}$$

The first observation is that trying to minimize the total error $\mathcal{E}_{\text{tot}}$ is equivalent to minimization of the integrated ratio of the signal spectrum $P(\omega)$ and its estimated spectrum $\hat{P}(\omega)$. Furthermore, at minimum error $\mathcal{E}_{\text{tot}} = G^2$, the following relation holds

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{P(\omega)}{\hat{P}(\omega)} d\omega = 1. \tag{15}$$

Equations (14) and (15) result in two major properties, a "global" and "local" property [24].

- The global property states that since the contribution to the total error $\mathcal{E}_{\text{tot}}$ is determined as a ratio of the two spectra, the matching process should perform uniformly over the

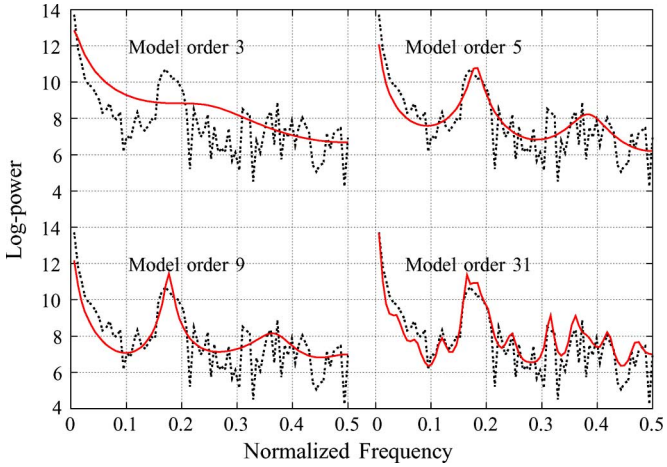

Fig. 4. Power density of the zero-order MFCC of a piano note $A5$ played for a duration of 1.2 s. The four figures show the periodogram as well as the AR-model power spectrum estimates of orders 3, 5, 9, and 31, respectively. The normalized frequency of 1/2 corresponds to 66.67 Hz.

whole frequency range, irrespective of the shaping of the spectrum. This means that the spectrum match at frequencies with small energy is just as good as frequencies with high energy.

- The local property deals with the matching of the spectrum in each small region of the spectrum. In [24], the author concludes that a better fit of $\hat{P}(\omega)$ to $P(\omega)$ will be obtained at frequencies where $P(\omega)$ is larger than $\hat{P}(\omega)$, than at frequencies where $P(\omega)$ is smaller. Thus, for harmonic signals, the peaks will be better approximated than the area in between the harmonics.

It is now seen that the autoregressive (AR) method and the FC approach are related, since in the latter method, the periodogram is summarized in four frequency bands where for the AR-model approach a selection of frequency bands is unnecessary since the power spectrum is modeled directly.

Fig. 4 shows the periodogram of the zero-order MFCC of the piano note $A5$ corresponding to the frequency 880 Hz recorded over a duration of 1.2 s as well as the AR-model approximation for four different model orders, 3, 5, 9, and 31. The hop-size of the MFCCs were 7.5 ms corresponding to a sample rate of 133.33 Hz. As expected, the model power spectrum becomes more detailed as the model order increases.

B. MAR

In order to include both temporal and among feature correlations, the multivariate AR model with full matrices is applied instead of only considering the diagonal of the matrices as in the DAR model. A full treatment of the MAR models are given in [25] and [26].

For a stationary time series of state vectors $\mathbf{x}_n$, the general multivariate AR model is defined by

$$\mathbf{x}_n = \sum_{p=1}^{P} \mathbf{A}_p \mathbf{x}_{n-I(p)} + \mathbf{u}_n \tag{16}$$

where the noise term $\mathbf{u}_n$ is assumed i.i.d. with mean $\mathbf{v}$ and finite covariance matrix $\mathbf{C}$. The above formulation is quite general since I refers to a general set; e.g., for a model order of 3, the set could be selected as $I = \{1, 2, 3\}$ or as $I = \{2, 4, 8\}$ indicating that $\mathbf{x}_n$ is predicted from these previous state vectors. Note that the mean value of the noise process $\mathbf{v}$ is related to the mean $\mathbf{m}$ of the time series by $\mathbf{m} = \left(\mathbf{I} - \sum_{p=1}^{P} \mathbf{A}_p\right)^{-1} \mathbf{v}$.

The matrices $\mathbf{A}_p$ for $p = 1, \ldots, P$ are the coefficient matrices of the Pth order multivariate autoregressive model. They encode how much of the previous information in $\{\mathbf{x}_{n-I(1)}, \mathbf{x}_{n-I(2)}, \ldots, \mathbf{x}_{n-I(P)}\}$ is present in $\mathbf{x}_n$. In this paper, the usual form of the multivariate AR model have been used, hence, $I = \{1, 2, \ldots, P\}$.

A frequency interpretation of the multivariate autoregressive model can, as for the univariate case, be established for the multivariate case. The main difference is that all cross spectra are modeled by the MAR model. In, e.g., [27], a frequency domain approach is used for explaining the multivariate autoregressive model by introducing the "autocovariance function," which contains all cross covariances for the multivariate case. The power spectral matrix can be defined from the autocovariance function as

$$\mathbf{f}(\omega) = \sum_{h=-W_s+1}^{W_s-1} \mathbf{\Gamma}(h) e^{-ih\omega} \tag{17}$$

where the autocovariance function $\mathbf{\Gamma}(h)$ is a positive function and fulfills $\sum_{h=-\infty}^{\infty} ||\mathbf{\Gamma}(h)||_2 < \infty$, under stationarity.

As with the DAR model, the ordinary least squares approach has been used in estimating the parameters of the MAR model; see, e.g., [25] for detailed explanation of parameter estimation.

The parameters which are extracted from the least squares approach for both the DAR and MAR models are the AR-matrices: $\{\mathbf{A}_1, \ldots, \mathbf{A}_P\}$, the intercept term $\mathbf{v}$ and the noise covariance $\mathbf{C}$. The temporal feature integrated vector of window k then becomes

$$\mathbf{z}_k = \left[\mathrm{vec}(\mathbf{B}_k)^T, \mathbf{v}_k^T \mathrm{vech}(\mathbf{C}_k)^T\right]^T \tag{18}$$

where $\mathbf{B} = [\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_P]$ is of dimension $D \times PD$ and $\mathbf{z}_k$ of dimension $(P + 1/2)D^2 + (3/2)D$. Note that for the DAR model, only the diagonals of the $\mathbf{A}_p$ and $\mathbf{C}$ matrices are used.

C. Issues on Stability

Until now, we have assumed that the time-series under investigation is stationary over the given window. The window-size, however, is optimized to the given learning problem which means that we are not guaranteed that the time-series is stationary within each window. This could, e.g., be in transitions from silence to audio, where the time-series might locally look nonstationary. In some applications, this is not a problem, since reasonable parameter estimates are obtained anyhow. In the considered music genre setup, the classifier seems to handle the nonstationary estimates reasonably. In other areas of MIR, the power-spectrum estimate provided through the AR-model might be more critical, hence, in such cases it would be relevant to investigate the influence of nonstationary windows.

D. Selection of Optimal Length

There exists multiple order selection criteria. Examples are Bayesian information criterion (BIC) and Akaike information

TABLE I
COMPUTATIONAL COMPLEXITY OF ALGORITHMS OF A WINDOW OF SHORT-TIME FEATURES

METHOD	MULTIPLICATIONS & ADDITIONS
MeanVar	$4DW_s$
MeanCov	$(D+3)DW_s$
FC	$(4\log_2(W_s)+3)\,DW_s$
DAR	$\frac{D}{3}(P+1)^3+((P+6)(P+1)+3)\,DW_s$
MAR	$\frac{1}{3}(PD+1)^3+$ $\left((P+4+\frac{2}{D})(PD+1)+(D+2)\right)DW_s$

TABLE II
COMPUTATIONAL COMPLEXITY OF THE MUSIC GENRE SETUP USING THE OPTIMIZED VALUES FROM THE EXPERIMENTAL SECTION, HENCE $P=3$, $D=6$, AND $W_s=188, 268, 322, 295, 162$ FOR THE MeanVar, MeanCov, FC, DAR, AND MAR, RESPECTIVELY. NOTE THAT THE COMPLEXITY VALUES ARE NORMALIZED SUCH THAT THE MeanVar HAS COMPLEXITY 1

Model	Complexity relative to MeanVar
MeanCov	3.2
FC	15.6
DAR	27.2
MAR	32.2

criterion (AIC); see, e.g., [26]. The order selection methods are traditionally applied to a single time series; however, in the music genre setup, we are interested in finding a single optimal model order for a large set of time-series. Additionally, there is a tradeoff between model order and the dimensionality of the feature space and, hence, problems with overfitting of the subsequent classifier, see Fig. 1, Section I. Therefore, the optimal order of the time-series alone is normally not the same as the optimal order determined for the complete system.

E. Complexity Considerations

Table I shows the complete number of multiplications and additions for a window of all the examined temporal feature integration methods. The column "multiplications & additions" shows the number of calculated multiplications/additions of the particular method. D is the dimensionality of the feature space, P is the DAR/MAR model order, and W_s is the window-size in number of short-time feature samples. In the calculations, the effect of overlapping windows have not been exploited. Table II shows the computational complexity of our actual music genre setup. The complexities are scaled according to the MeanVar calculation.

IV. EXPERIMENTS

Simulations were designed to compare the baseline MeanVar features with the newly proposed DAR and MAR features. Additionally, the FC features and MeanCov features were included in the comparisons. The FC features performed very well in [9], and the MeanCov features were included for the sake of completeness.

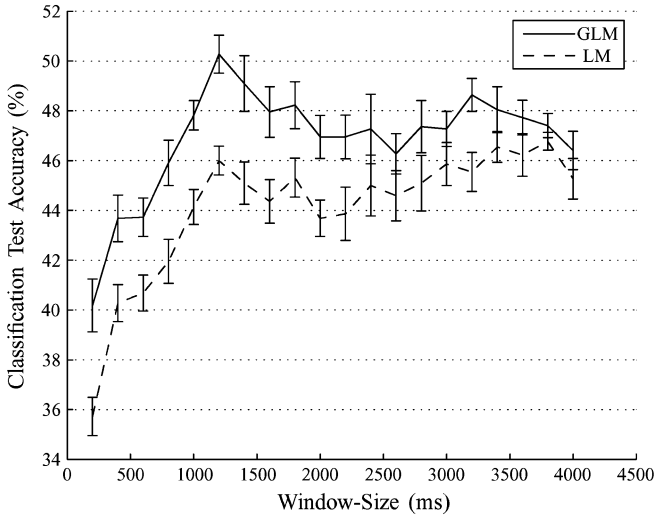

Fig. 5. Classification test accuracy is plotted against window-size for the MAR features using the LM and GLM classifiers. The hop-size was 200 ms in these experiments and data set B, Section IV-D, was used. The error-bars denote the standard deviation on the mean value. The importance of the window-size is clearly seen. The baseline classification accuracy by random guessing is ~9.1%.

The features were tested on two different data sets and four different classifiers to make the conclusions generalizable. In all of the experiments, tenfold cross-validation was used to estimate the mean and standard deviation of the mean classification test accuracy, which was used as the performance measure. Fig. 1 in Section I illustrates the complete classification system. The optimization of the system follows the data stream which means that the MFCC features were optimized first (choosing number of coefficients to use, whether to use normalization, etc.). Afterwards, the temporal feature integration part was optimized and so forth.

A. Preliminary Investigations

Several investigations of preprocessing both before and after the temporal feature integration were made. Dimensionality reduction of the high-dimensional MAR and DAR features by PCA did not prove beneficial,[3] and neither did whitening (making the feature vector representation zero-mean and unit covariance matrix) or normalization (making each feature component zero-mean and unit variance individually) for any of the features. To avoid numerical problems, however, they were all normalized. Preprocessing, in terms of normalization of the short-time MFCC features did not seem to have an effect either.

B. Features

To ensure a fair comparison between the features, their optimal hop- and window-sizes were examined individually, since especially window-size seems important with respect to classification accuracy. An example of its importance is illustrated in Fig. 5.

[3]This is only true for the standard GLM and LM classifiers, that does not have significant overfitting problems.

For the short-time MFCC features, the first six coefficients (including the zero-order MFCC) where found to be adequate for the experiments on the two datasets. The optimal hop- and frame-size were found to be 7.5 and 15 ms, respectively. The optimal hop-size was 400 ms for the DAR, MAR, MeanVar, and MeanCov features and 500 ms for the FC features. The window-sizes were 1200 ms for the MAR features, 2200 ms for the DAR features, 1400 ms for the MeanVar, 2000 ms for the MeanCov, and 2400 ms for the FC features.

An important parameter in the DAR and MAR feature models is the model order parameter P. The optimal values for this parameter were found to be 5 and 3 for the DAR and MAR features, respectively. This optimization was based on the large data set B, see Section IV-D. Using these parameters, the resulting dimensions of the feature spaces become: MAR—135, DAR—42, FC—24, MeanCov—27, and MeanVar—12.

C. Classification and Postprocessing

Several classifiers have been tested such as a linear model trained by minimizing least squares error (LM), Gaussian classifier with full covariance matrix (GC), Gaussian mixture model (GMM) classifier with full covariance matrices and a Generalized Linear Model (GLM) classifier [28]. The LM and GLM classifiers are robust and have been used in all of the initial feature investigations.

The LM classifier is a linear regression classifier, but has the advantage of being fast and noniterative; the training essentially amounts to finding the pseudoinverse of the feature-matrix. The GLM classifier is the extension of a logistic regression classifier to more than two classes. It can also be seen as an extension of the LM classifier, but with inclusion of a regularisation term (prior) on the weights and a cross-entropy error measure to account for the discrete classes. They are both discriminative, which could explain their robust behavior in the fairly high-dimensional feature space. Tenfold cross validation was used to set the prior of the GLM classifier.

1) Postprocessing: Majority voting and the sum-rule were examined to integrate the c classifier outputs of all the windows into 30 s (the size of the song clips), whereas majority voting counts the hard decisions

$$\arg\max_c P(c|\mathbf{z}_k) \tag{19}$$

for $k = 1, \ldots, K$ of the classifier outputs, the sum-rule sums over the "soft" probability densities $P(c|\mathbf{z}_k)$ for $k = 1, \ldots, K$. The sum-rule was found to perform slightly better than majority voting.

D. Data Sets

Two data sets have been used in this investigation. Both of the data sets have been described in more detail in [16] and [2].

The first data set, denoted "data set A," consists of 100 sound clips distributed evenly among the five music genres: *Rock, Classical, Pop, Jazz*, and *Techno*. Each of the 100 sound clips, of length 30 s, are recorded in mono PCM format at a sampling frequency of 22 050 Hz.

The second data set denoted "data set B" consists of 1210 music snippets distributed evenly among the 11 music genres: *Alternative, Country, Easy Listening, Electronica, Jazz, Latin, Pop&Dance, Rap&HipHop, R&B Soul, Reggae*, and *Rock*. Each of the sound clips, of length 30 s, are encoded in the MPEG1-layer 3 format with a bit-rate of 128 kb/s. The sound clips were converted to mono PCM format with a sampling frequency of 22 050 Hz prior to processing.

E. Human Evaluation

The level of performance in the music genre setups using various algorithms and methods only shows their relative differences. However, by estimating the human performance on the same data sets, the quality of automated genre classification systems can be assessed.

Listening tests have been conducted on both the small data set (A) and the larger data set (B). At first, subsets of the full databases were picked randomly with an equal amount from each genre (25 of 100 and 220 of 1210), and these subsets are believed to represent the full databases. A group of people (22 specialists and nonspecialists) were kindly asked to listen to 30 different sound clips of length 10 s from data set A[4] and classify each sound clip into one of the genres on a forced-choice basis. A similar setup was used for the larger data set B, but now 25 persons were asked to classify 33 sound clips of length 30 s.[5] No prior information except the genre names were given to the test persons. The average human accuracy on data set A lies in a 95% confidence interval of [0.97;0.99], and for data set B it is [0.54;0.61]. Another interesting measure is the confusion between genres which has been compared to the automated music system in Fig. 7.

F. Results and Discussion

The main classification results are illustrated in Fig. 6 for both the small and the large data set. The figures compares the cross-validated classification test accuracies of the FC and MeanCov features and the baseline MeanVar with the newly proposed DAR and MAR features. It is difficult to see much difference in performance between the features for the small data set A [see Fig. 6(a)], but note that it was created to have only slightly overlapping genres which could explain why all the features perform so well compared to the random guess of only 20% accuracy. The classification test accuracies of the different methods are not to far from the average human classification accuracy of 98%.

The results from the more difficult, large data set B are shown in Fig. 6(b). Here, the MAR features are seen to clearly outperform the conventional MeanVar features when the LM or GLM classifiers are used. Similarly, they outperform the MeanCov and DAR features. The DAR features only performed slightly better than the three reference features, but in a feature space of much lower dimensionality than the MAR features. The GMM classifier is the best for the low-dimensional MeanVar features, but gradually loses to the discriminative classifiers as the feature space dimensionality rises. This overfitting problem was

[4]These sound clips have been created by splitting each 30-s sound clip into five overlapping sound clips of 10 s. This results in 125 sound clips of 10 s.

[5]Hence, 33 songs from the subset of 220 were picked at random for each test person.

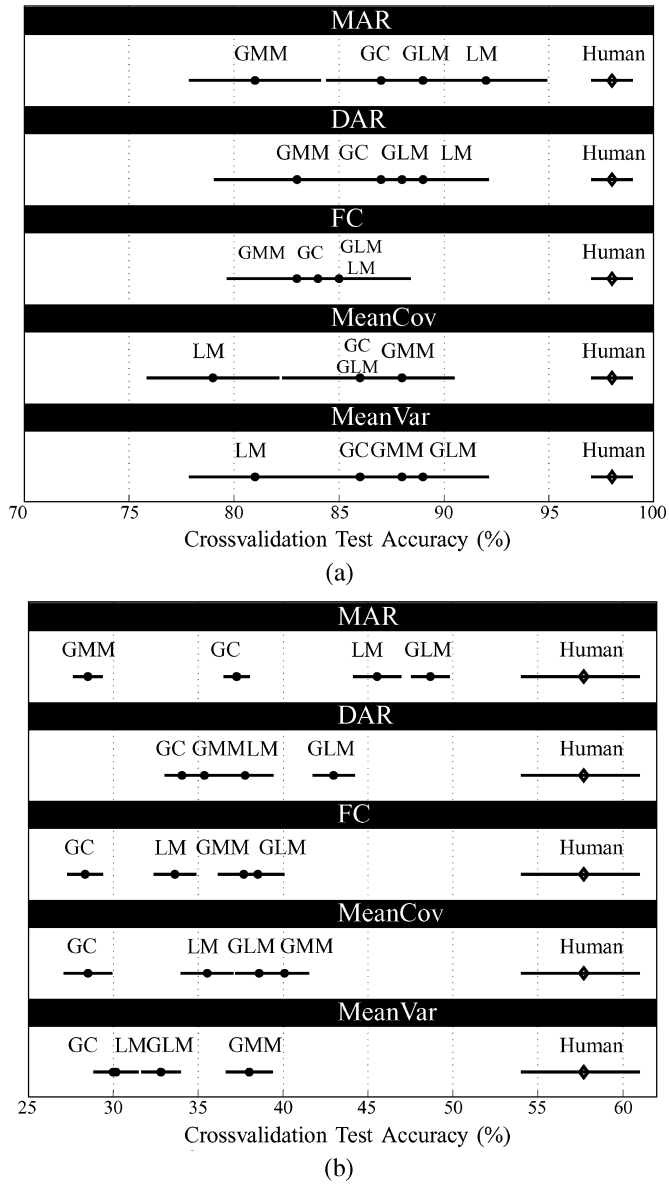

Fig. 6. Genre classification test accuracies for the GC, GMM, LM, and GLM classifiers on the five different integrated features. The results for the small data set A is shown in the upper panel of the figure and the results for the larger data set B in the lower panel. The mean accuracy of tenfold cross-validation is shown along with error bars, which are one standard deviation of the mean to each side. 95% binomial confidence intervals have been shown for the human accuracy. (a) Experiment on data set A. (b) Experiment on data set B.

obviously worst for the 135-dimensional MAR features and dimensionality reduction was necessary. However, a PCA subspace projection was not able to capture enough information to make the GMM classifier competitive for the MAR features. Improved accuracy of the GMM classifier on the MAR features was achieved by projecting the features into a subspace spanned by the $c-1$ weight directions of the partial least squares (PLS) [29], where c refers to the number of genres. The classification accuracy, however, did not exceed the accuracy of the GLM classifier on the MAR features.

The MAR features are still around 9% from the average human classification test accuracy of approximately 57%; however, it should be noted that only the initial six MFCCs were used. Furthermore, it should be noticed that random classification accuracy is only 9%.

The cross-validation paired t-test [30] was made on both data sets to test whether the best performances of the DAR and MAR features differed significantly from the best performances of the other features. Comparing the MAR features against the other four features gave t-statistics estimates all above 3.90—well above the 0.975 percentile critical value of $t_{9,0.975} = 2.26$ for tenfold cross-validation. Thus, the null hypothesis of similar performance can be rejected. The comparison between the DAR features and the three reference features gave t-statistics estimates of 2.67 and 2.83 for the FC and MeanVar features, but only 1.56 for the MeanCov features which means that the null hypothesis cannot be rejected for the MeanCov.

As described in Section IV-B, the window-sizes were carefully investigated and the best results were found using window-sizes in the range of 1200 to 2400 ms, followed by the sum-rule on the classifier decisions up to 30 s. However, in, e.g., music retrieval and regarding computational speed and storage, it would be advantageous to model the whole 30-s music snippet with a single feature vector. This approach have been followed by several authors, see, e.g., [17], [31], and [32]. In [31], primarily models with no closed-form solution of the parameters have been investigated.[6] When modeling at a music snippet time scale, the temporal correlations and the cross-correlations among the feature dimensions differs from the correlations extracted when modeling at the medium time scale; see [16]. Especially, the among feature dimensions correlation for the MFCCs tend to be small at the music snippet time scale, which is motivated from the fact that the discrete cosine transform (DCT) in the MFCC extraction stage is in fact decorrelating the feature dimensions.

Hence, experiments were made with the MAR features with a window size of 30 s, i.e. modeling the sound snippet with a single MAR model. The best mean cross-validated classification test accuracies on data set B were 44% and 40% for the LM and GLM classifiers, respectively, using a MAR model order of 3. In our view, this indicates that these MAR features could be used with success in, e.g., song similarity tasks. Additional experiments with a support vector machine (SVM) classifier [32] using a RBF type of kernel even improved the accuracy to 46%. The SVM classifier was used since it is less prone to overfitting. This is especially important when each song is represented by a single feature vector, which means that our training set only consists of $11 \cdot 99 = 1089$ samples in each cross-validation run.

Besides the classification test accuracy, an interesting measure of performance is the confusion matrix. Fig. 7 illustrates the confusion matrix of the MAR system with highest classification test accuracy and shows the relation to the human genre confusion matrix on the large data set. It is worth noting that the three genres that humans classify correctly most often, i.e., Country, Rap&HipHop, and Reggae are also the three genres that our classification system typically classifies correctly. To get an insight in the confusion among the different genres, dendrograms were created from the confusion matrices in Fig. (7). The dendrogram of the human and MAR confusion

[6]Gaussian mixture models and hidden Markov models (HMMs).

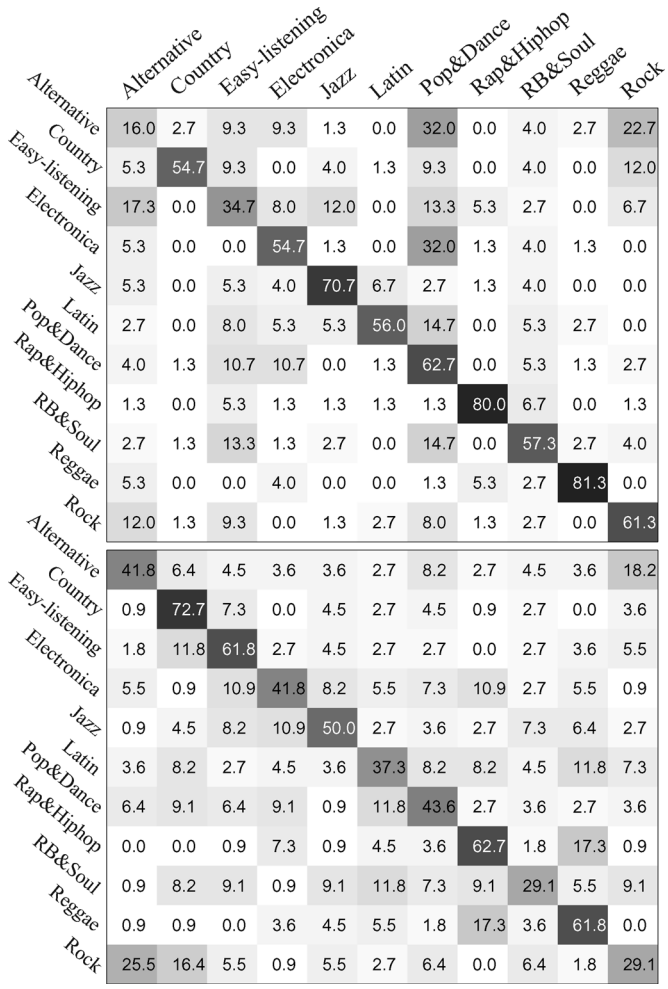

Fig. 7. The above confusion matrices were created from data set B. The upper figure shows the confusion matrix from evaluations of the 25 people, and the lower figure shows the average of the confusion matrices over the ten cross-validation runs of the best performing combination (MAR features with the GLM classifier). The "true" genres are shown as the rows, which each sum to 100%. The predicted genres are then represented in the columns. The diagonal illustrates the accuracy of each genre separately.

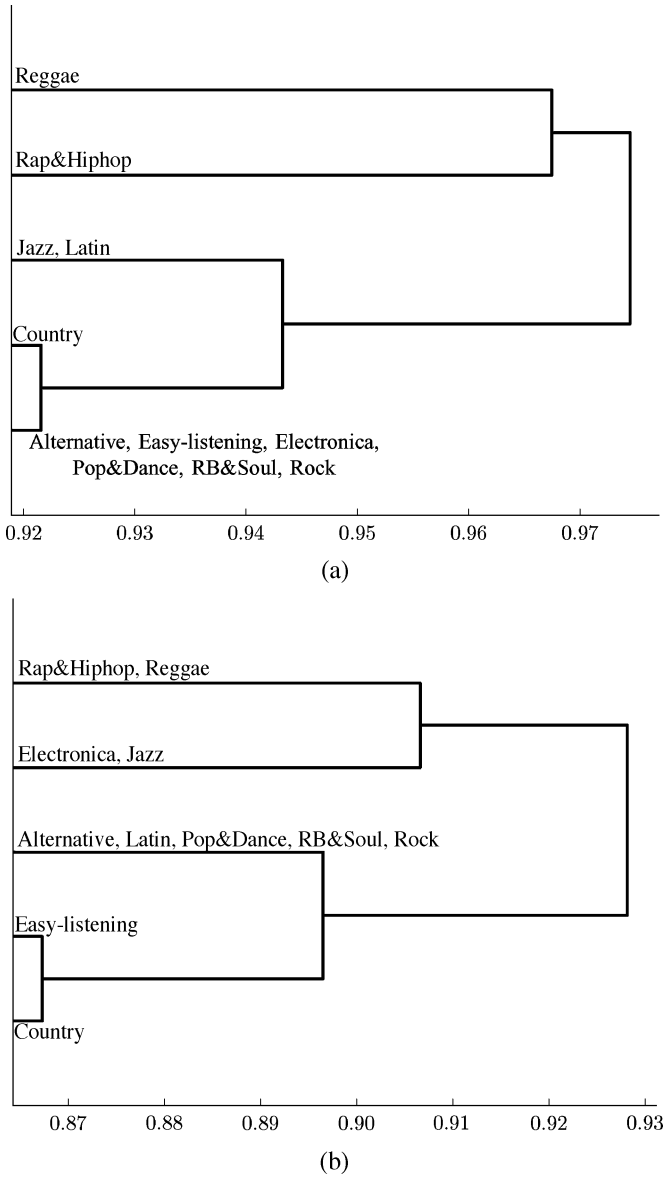

Fig. 8. Dendrograms illustrating the groupings of genres determined from the confusion matrices in Fig. 7. (a) Dendrogram created from the human confusion matrix. (b) Dendrogram created from MAR confusion matrix.

matrices have been illustrated in Fig. 8(a) and (b), respectively. The confusion matrices were symmeterized before creating the dendrograms. Furthermore, a scaled exponential distance measure were applied for creation of the five-cluster dendrograms. Different distance measures were investigated; however, no big differences were observed in the resulting clusters. The dendrograms illustrate that the larger clusters of the human and MAR confusion matrices shares the music genres: Alternative, Pop&Dance, Rb&Soul, and Rock. Furthermore, the MAR model with a GLM classifier tend to confuse Rap&Hiphop and Reggae more than humans do.

A small-scale analysis was conducted to test the robustness of the MAR-model to MP3 encoding. The best performing setup for data set A, which was a MAR model with a LM classifier was investigated. Each music snippet was encoded to 128, 64, 32, and 16 kb/s, respectively, using the `LAME version 3.96.1` encoder. Similarily, the music snippets were decoded using the `LAME` decoder prior to the MFCC extraction stage. The different music snippets were resampled to 16 kHz when extracting the MFCCs, to ensure a common ground for comparison. The classification test accuracy was assessed with tenfold cross-validation. In each fold, the training set consisted of the PCM samples and test accuracies were obtained from the different MP3 encodings and the PCM encoding. The mean cross-validation test accuracies obtained have been illustrated in Table III.

The combination of a MAR model and LM classifier is robust in the given setup to encodings of 32 kb/s and above. It should be noticed, however, that since we are modeling the short-time features, the robustness of the complete system is dictated by the robustness of the short-time features towards the different encoding schemes. Still, the investigation indicate that the MAR features are not over-sensitive to small changes in the short-time features.

TABLE III
MEAN CROSS-VALIDATION TEST ACCURACIES OF THE LM CLASSIFIER ON THE MAR FEATURES ON DATA SET A USING DIFFERENT MP3 ENCODING RATES. TRAINING HAVE BEEN PERFORMED WITH THE RAW PCM SAMPLES

Encoding	Mean test accuracy ± Std. deviation of the mean
PCM	$93.3\% \pm 1.8$
128 kb/s	$92.2\% \pm 2.4$
64 kb/s	$91.1\% \pm 2.2$
32 kb/s	$94.4\% \pm 1.8$
16 kb/s	$28.9\% \pm 4.7$

V. CONCLUSION

In this paper, we have investigated temporal feature integration of short-time features in a music genre classification task and a novel multivariate autoregressive feature integration scheme was proposed to incorporate dependencies among the feature dimensions and correlations in the temporal domain. This scheme gave rise to two new features, the DAR and MAR, which were carefully described and compared to features from existing temporal feature integration schemes. They were tested on two different data sets with four different classifiers, and the successful MFCC features were used as the short-time feature representation. The framework is generalizable to other types of short-time features. Especially, the MAR features were found to perform significantly better than existing features, but also the DAR features performed better than the FC and baseline MeanVar features on the large data set and in a much lower dimensional feature space than the MAR. Furthermore, it was illustrated that the MAR features are robust towards MP3 encoding for bitrates of 32 kb/s and above.

Human genre classification experiments were made on both data sets and we found that the mean human test accuracy was less than 18% better relative to the best performing MAR features approach on the 11 music genre dataset.

Possible directions for future research include investigation of other types of indexes for the general multivariate AR formulation, hence, allowing a more flexible modeling of short-time features at larger time scales and to consider methods for handling of nonstationary windows.

As a closing remark, it should be noticed that the considered framework of temporal feature integration is open to other areas of MIR.

REFERENCES

[1] F. Gouyon, S. Dixon, E. Pampalk, and G. Widmer, "Evaluating rhythmic descriptors for musical genre classification," in *Proc. 25th Int. AES Conf.*, London, U.K., 2004, pp. 196–204.

[2] P. Ahrendt, "Music genre classification systems—A computational approach," Ph.D. dissertation, Informatics and Mathematical Modeling, Technical Univ. Denmark,, Lyngby, Denmark, 2006.

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

[4] G. Tzanetakis, "Manipulation, analysis, and retrieval systems for audio signals," Ph.D. dissertation, Dept. Comput. Sci., Princeton Univ., Princeton, NJ, 2002.

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

[6] H. Soltau, T. Schultz, M. Westphal, and A. Waibel, "Recognition of music types," in *Proc. ICASSP*, Seattle, WA, May 1998, vol. 2, pp. 1137–1140.

[7] S. H. Srinivasan and M. Kankanhalli, "Harmonicity and dynamics-based features for audio," in *Proc. ICASSP*, 2004, pp. 321–324.

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

[9] A. Meng, P. Ahrendt, and J. Larsen, "Improving music genre classification using short-time feature integration," in *Proc. ICASSP*, 2005, pp. 497–500.

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

[11] P. Herrera, A. Yeterian, and F. Gouyon, "Automatic classification of drum sounds: A comparison of feature selection and classification techniques," in *Proc. 2nd Int. Conf. Music Artif. Intell.*, 2002, pp. 79–91.

[12] L. R. Rabiner and B. Juang, *Fundamental of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[13] S. Sigurdsson, K. B. Petersen, and T. Lehn-Schiøler, "Mel frequency cepstral coefficients: An evaluation of robustness of MP3 encoded music," in *Proc. Int. Conf. Music Inf. Retrieval*, 2005, pp. 286–289.

[14] B. Logan, "Mel frequency cepstral coefficients for music modeling," in *Proc. Int. Symp. Music Inf. Retrieval*, Plymouth, MA, Oct. 2000.

[15] J.-J. Aucouturier and F. Pachet, "Representing music genre: A state of the art," *J. New Music Res.*, vol. 32, no. 1, pp. 83–93, Jan. 2003.

[16] A. Meng, "Temporal feature integration for music organisation," Ph.D. dissertation, Informatics and Mathematical Modeling, Technical Univ. Denmark, Lyngby, Denmark, 2006.

[17] E. Pampalk, "Computational models of music similarity and their application to music information retrieval," Ph.D. dissertation, Vienna Univ. Technology, Vienna, Austria, Mar. 2006.

[18] K. Martin, "Sound-source recognition: A Theory and computational model," Ph.D. dissertation, Mass. Inst. Technol., Cambridge, MA, Jun. 1999.

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

[20] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 293–302, Jul. 2002.

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

[22] E. Terhardt, "On the perception of periodic sound fluctuations (roughness)," *Acustica*, vol. 30, no. 4, pp. 201–213, 1974.

[23] D. Ellis and K. Lee, "Features for segmenting and classifying long-duration recordings of personal audio," in *Proc. ISCA Tutorial Research Workshop Statistical Perceptual Audio Process. SAPA-04*, Jeju, Korea, Oct. 2004, pp. 1–6.

[24] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, Apr. 1975.

[25] H. Lütkepohl, *Introduction to Multiple Time Series Analysis*, 2nd ed. New York: Springer, 1993.

[26] A. Neumaier and T. Schneider, "Estimation of parameters and eigenmodes of multivariate autoregressive models," *ACM Trans. Math. Softw.*, vol. 27, no. 1, pp. 27–57, Mar. 2001.

[27] F. R. Bach and M. I. Jordan, "Learning graphical models for stationary time series," *IEEE Trans. Signal Process.*, vol. 52, no. 8, pp. 2189–2199, Aug. 2004.

[28] I. Nabney and C. Bishop, "Netlab Package," 1995 [Online]. Available: http://www.ncrg.aston.ac.uk/netlab/index.php

[29] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[30] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Comput.*, vol. 10, no. 7, pp. 1895–1923, Oct. 1998.

[31] J.-J. Aucouturier, "Ten experiments on the modelling of polyphonic timbre," Ph.D. dissertation, Univ. Paris, Paris, France, 2006.

[32] A. Meng and J. Shawe-Taylor, "An investigation of feature models for music genre classification using the support vector classifier," in *Proc. Int. Conf. Music Inf. Retrieval*, 2005, pp. 604–609.

Anders Meng received the M.Sc. and Ph.D. degrees in applied mathematics from the Technical University of Denmark (DTU), Lyngby, in 2003 and 2006. His thesis was entitled "Temporal feature integration for music organization."

From 1999 to 2001, he worked as a Research Assistant in a leading company within mobile communication. He is currently a Postdoctoral Researcher working in the Digital Signal Processing Group, Informatics and Mathematical Modeling, DTU . His current research interests include machine learning, audio signal processing, and the combination of the latter in large-scale webmining.

Peter Ahrendt received the M.Sc. degree from the University of Southern Denmark, Odense, in 2001 with specialization in applied mathematics and physics and the Ph.D. degree from the Technical University of Denmark, Lyngby, with the thesis "Music genre classification systems: A computational approach."

From 2001 to 2002, he worked as Research Scientist in a major pharmaceutical company and is currently with Widex, Vaerloese, Denmark, with a main focus on audiology research. His current research interests include machine learning, audio signal processing as well as areas related to the human perception of sound.

Jan Larsen (SM'03) received the M.Sc. and Ph.D. degrees in electrical engineering from the Technical University of Denmark (DTU), Lyngby, in 1989 and 1994, respectively.

He is currently an Associate Professor of Digital Signal Processing, Informatics and Mathematical Modeling, DTU. He has authored and coauthored around 100 papers and book chapters within the areas of nonlinear statistical signal processing, machine learning, neural networks, and datamining with applications to biomedicine, monitoring systems, multimedia, and webmining. He has participated in several national and international research programs and has served as reviewer for many international journals, conferences, publishing companies, and research funding organizations.

Dr. Larsen took part in conference organizations, among others, the IEEE Workshop on Machine Learning for Signal Processing (formerly Neural Networks for Signal Processing) 1999–2006. Currently, he is chair of the IEEE Machine Learning for Signal Processing Technical Committee of the IEEE Signal Processing Society and Chair of the IEEE Denmark Section's Signal Processing Chapter. He is an Editorial Board Member of Signal Processing, Elsevier, 2006–2009. He is a Guest Editor of the IEEE TRANSACTIONS ON NEURAL NETWORKS, the *Journal of VLSI Signal Processing Systems*, and *Neurocomputing*.

Lars Kai Hansen is a Professor of digital signal processing at the Technical University of Denmark, Lyngby. He is head of the THOR Center for Neuroinformatics and leader of the project http://www.intelligentsound.org. His research concerns adaptive signal processing and machine learning with applications in biomedicine and digital media.