MUSICAL GENRE CLASSIFICATION BASED ON A HIGHLY-RESOLVED CEPSTRAL MODULATION SPECTRUM

Anil Nagathil, Timo Gerkmann, and Rainer Martin

Institute of Communication Acoustics, Ruhr-Universität Bochum, Bochum, Germany
{anil.nagathil,timo.gerkmann,rainer.martin}@rub.de

ABSTRACT

We propose new features for musical genre classification which are based on the modulation spectrum of cepstral coefficients, and investigate the impact of the modulation frequency resolution on the classification accuracy. We compare the performance of the novel feature set which is derived from a high-resolution modulation spectrum to that of two feature sets which are either based on a coarsely resolved modulation spectrum or roughly summarize the modulation energy in a few bands. From the results of a 5-class musical genre classification experiment it can be concluded that a high modulation frequency resolution is crucial for representing the harmonic modulation structure of Electronic music in particular. The proposed features outperform the two competing methods with an overall detection rate of 81%. After computing the cepstral modulation spectrum with efficient FFT operations, the computational complexity for feature extraction is fairly low as only 22 low-level features need to be computed.

1. INTRODUCTION

A variety of features for the discrimination of different music categories has been proposed throughout the last decade [1]. Typically, these are short-time features which are extracted from quasi-stationary signal segments. The most prominent and certainly most commonly used features are mel-frequency cepstral coefficients (MFCC) which are a perceptually motivated variant of linear-frequency cepstral coefficients (LFCC) [2]. MFCCs, which mainly represent the spectral envelope, were applied successfully in automatic speech recognition (ASR) [3] and were introduced in the field of music information retrieval as well, e.g [4, 5]. However, in [6] the suitability of MFCCs for music modeling is investigated and possible limitations thereof are discussed.

After extracting short-time features the question arises how they should be aggregated temporally in order to obtain a compact but representative feature set. The easiest way certainly is to assume a Gaussian distribution and express the properties of each feature time series by its mean and variance as it was done in [4]. This strategy, however, models the temporal evolution of the underlying music signal insufficiently. Therefore, other approaches such as feature modulation spectrum analysis [5, 7] or autoregressive modeling [8] were proposed which yield parameters frequently referred to as dynamic features and considerably improve the performance in an audio classification task.

The process of feature extraction and temporal aggregation, however, can also be reversed. In [9] first a low-resolution modulation spectrum of linear-frequency cepstral coefficients is computed from which the actual features are extracted in the final processing step. As opposed to conventional feature extraction strategies, this procedure aims at preserving as much information as possible before modeling the signal dynamics. In a speech, music and noise discrimination experiment it was shown that only eight features are required for obtaining detection rates above 95% where 39 static and dynamic MFCCs are needed to yield comparable results. Thus, instead of accumulating a large amount of features from various signal representations and selecting the most powerful ones subsequently as in e.g. [7, 10], we argue that a unified and yet flexible high-resolution spectro-temporal representation of the signal should be considered from which simple, but effective dynamic features can be extracted.

In this contribution, we use the features proposed in [9] for a musical genre classification task and evaluate their performance. Further, a novel set of features is derived which is based on a cepstral modulation spectrum with highly-resolved modulation frequencies as opposed to the low-resolution modulation spectrum used in [9]. The new features are more related to characteristics of music signals such as rhythm, timbre and pitch. The performances of these two feature sets are compared which will answer the question if the temporal fine structure of music signals which is captured by a high-resolution modulation spectrum bears distinctive genre-related properties and therefore needs to be represented. In addition, a competing method proposed in [5] is evaluated which uses static and dynamic MFCCs. In this approach the modulation energy of MFCCs is coarsely summarized in four bands.

After describing the signal processing stages required for obtaining the cepstral modulation spectrum in Section 2 and analyzing it for music signals in Section 3, the two different feature sets derived from this representation are outlined in Section 4. The performance of these features is studied in Section 5 where we assume different modulation frequency resolutions, and is compared to static and dynamic MFCCs.

2. CEPSTRAL MODULATION SPECTRUM

We consider a section of a sampled raw audio signal $x(\tilde{n})$, henceforth called *frame*, where $\tilde{n} = 0, 1, ..., N_\mathrm{T}$ is the discrete time index and N_T is the total number of samples. The sampling frequency is denoted by f_s. The audio frame is segmented into λ_T (possibly overlapping) *subframes* $x(\lambda R + n)$ of length N, i.e. $n = 0, 1, ..., N-1$, where λ and R denote the subframe index and subframe shift, respectively. After weighting each subframe with the Hann window $w(n) = 0.5(1 - \cos(2\pi n/N))$, we perform a short-term spectral analysis by means of a discrete Fourier transform (DFT)

$$X(\mu, \lambda) = \sum_{n=0}^{N-1} x(\lambda R + n)\, w(n) \exp\left(-j\frac{2\pi n \mu}{N}\right), \qquad (1)$$

where $\mu = 0, 1, ..., N-1$ denotes the frequency bin index. A harmonic decomposition of the spectral content results in coefficients which correspond to different degrees of spectral detail. This is achieved by applying the cepstral transform

$$x_\mathrm{c}(q, \lambda) = \frac{1}{N} \sum_{\mu=0}^{N-1} \ln\Big(|X(\mu, \lambda)|^2\Big) \exp\left(j\frac{2\pi \mu q}{N}\right), \qquad (2)$$

where $q = 0, 1, ..., N-1$ is the cepstral index. Low cepstral coefficients characterize the coarse spectral structure whereas higher cepstral coefficients represent the spectral details. Strictly harmonic structures in the spectrum are mainly mapped on single cepstral coefficients, i.e. the fundamental frequency f_0 of a periodic signal is mapped on a cepstral coefficient at $q_0 = f_\mathrm{s}/f_0$.

Cepstral dynamics can be captured by means of a sliding window DFT which results in a cepstral short-time modulation spectrum. Starting at subframe $\lambda = \Lambda S$ with the modulation analysis window index Λ and the modulation analysis window shift S the sliding window considers K consecutive subframes

$$X_{\mathrm{c}}(\nu, q, \Lambda) = \sum_{\kappa=0}^{K-1} x_{\mathrm{c}}(q, \Lambda S + \kappa) \exp\left(-j\frac{2\pi\kappa\nu}{K}\right) \tag{3}$$

where $\nu = 0, 1, ..., K-1$ is the modulation frequency bin index.

Temporally averaging over the total number of Λ_T magnitude modulation spectra yields a more compact representation of the cepstro-temporal structure

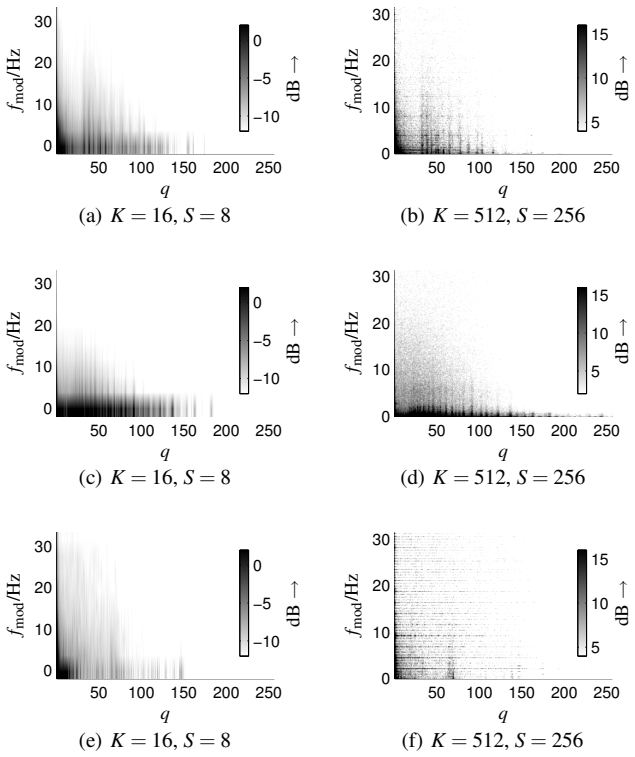
$$\overline{X}_{\mathrm{c}}(\nu, q) = \frac{1}{\Lambda_{\mathrm{T}}} \sum_{\Lambda=0}^{\Lambda_{\mathrm{T}}-1} |X_{\mathrm{c}}(\nu, q, \Lambda)|. \tag{4}$$

which we refer to as the mean cepstral magnitude modulation spectrum (MCMMS).

3. ANALYSIS OF THE MCMMS FOR MUSIC SIGNALS

In Figure 1 the MCMMS of exemplary Rhythm & Blues (R&B), Classical and Electronic music files are illustrated, respectively. Here, we consider coarsely resolved modulation frequencies (Figures 1(a), 1(c) and 1(e)) and highly resolved modulation frequencies (Figures 1(b), 1(d) and 1(f)) of a cepstral modulation spectrum for modulation frequencies $f_{\mathrm{mod}} < 31.25$ Hz. Note, that the subframe shift R and the modulation analysis window length K determine the modulation frequency range of the MCMMS representation in such a way that $f_{\mathrm{c,mod}} = 0.5 f_{\mathrm{s}}/R$ and $\Delta f_{\mathrm{mod}} = f_{\mathrm{s}}/(RK)$ denote the modulation cut-off frequency and modulation frequency bin spacing, respectively.

The low-resolution MCMMS of R&B music (Figure 1(a)) and Classical music (Figure 1(c)) bear a striking resemblance with their high-resolution counterparts (Figures 1(b) and 1(d)). A difference can be noticed, however, in the case of Electronic music (Figures 1(e) and 1(f)). Here, the high-resolution representation shows fine horizontal lines which result from the highly regular cepstro-temporal structure of this music style.



(a) $K = 16, S = 8$ (b) $K = 512, S = 256$

(c) $K = 16, S = 8$ (d) $K = 512, S = 256$

(e) $K = 16, S = 8$ (f) $K = 512, S = 256$

Figure 1: LFCC based MCMMS for Timbaland - *Fantasy* (R&B) [upper plots], Mussorgsky - *Pictures at an exhibition 1* (Classical) [middle plots] and Processor - *Nibtal* (Electronic) [lower plots], with $f_{\mathrm{s}} = 16$ kHz, $N = 512$ and $R = 256$, q: cepstral index.

4. CEPSTRAL MODULATION FEATURES

4.1 Cepstral Modulation Ratio Regressions (CMRARE)

As proposed in [9], the information content of an MCMMS with coarsely resolved modulation frequencies can be aggregated by means of cepstral modulation ratios (CMR) which normalize higher modulation frequency bins on the zeroth modulation frequency bin. These CMRs can be parameterized, respectively, by means of a polynomial fit of order p using a standard least-squares procedure [11]. This results in $p+1$ polynomial coefficients per CMR which we term *Cepstral Modulation RAtio REgression* (CMRARE) parameters.

4.2 Cepstral Modulation Music (CMM) Features

For an increasing modulation frequency resolution, CMRARE features become inappropriate as they reverse the benefits of a high-resolution cepstral modulation spectrum. Therefore, a set of 22 *Cepstral Modulation Music* (CMM) features for musical genre classification is proposed which characterizes the details of a highly-resolved modulation spectrum.

4.2.1 Decomposition of the MCMMS

For a systematic analysis of the cepstro-temporal properties of music signals, the MCMMS is decomposed into three cepstral ranges which describe the modulation structure of the spectral envelope, high fundamental frequencies and low fundamental frequencies, respectively. After analyzing the MCMMS of representative music signals and assuming the parameters $f_{\mathrm{s}} = 16$ kHz, $N = 512$, $R = 256$, $K = 512$ and $S = 256$, we have defined three cepstral ranges.

The interval $q \in [q_1, q_2] = [1, 30]$ can be assumed to characterize the spectral envelope. In Figure 2 the short-time spectrum of an exemplary music signal frame (solid line) is cepstrally smoothed, i.e. all cepstral coefficients $q > 30$ are to set to zero before transforming back the cepstral representation into the spectral domain. This is also referred to as liftering [2]. The liftered spectrum is depicted as a dashed line which represents the spectral envelope of the signal section. Therefore, the limits of the spectral envelope range specified above are chosen reasonably. Further, we define $q \in [q_1, q_2] = [31, 130]$ as the high pitch range and $q \in [q_1, q_2] = [131, 256]$ as the low pitch range, i.e. the transition from high to low fundamental frequencies is assumed to be at $f = f_{\mathrm{s}}/130 \approx 123$ Hz.

Furthermore, we consider three equally spaced modulation frequency intervals $\nu \in [\nu_1, \nu_2] = [2, 86]$ (slow modulations), $\nu \in [\nu_1, \nu_2] = [87, 171]$ (medium modulations) and $\nu \in [\nu_1, \nu_2] = [172, 256]$ (fast modulations). The cepstral and modulation frequency intervals specified above define nine partitions which are illustrated in Figure 3.

4.2.2 Timbral and Pitch-Related Features

For each cepstral range we compute the cepstral average of the MCMMS

$$Y_{q_1|q_2}(\nu) = \frac{1}{q_2 - q_1 + 1} \sum_{q=q_1}^{q_2} \overline{X}_{\mathrm{c}}(\nu, q). \tag{5}$$

This quantity exhibits a hyperbolic roll-off along the modulation frequency axis which is illustrated by solid lines in Figure 4(a) for

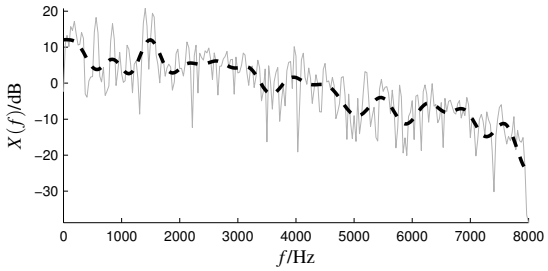


Figure 2: Short-time magnitude spectrum of an exemplary music signal segment (solid) and corresponding liftered spectrum (dashed).

the spectral envelope range of exemplary Rock and R&B music signals. These shapes indicate a temporal correlation of cepstral coefficients. As R&B music often relies on synthetic instruments and looped samples, cepstral coefficients are likely to be more temporally correlated than for Rock music where instruments are played manually. This results in different slopes for (5).

As indicated by dashed lines in Figure 4(a), $Y_{q_1|q_2}(\nu)$ can be modeled well by means of a hyperbolic function $\widehat{Y}_{q_1|q_2}(\nu) = \alpha_{\mathrm{q_1|q_2}}/\nu + \beta_{\mathrm{q_1|q_2}}$, where $\alpha_{\mathrm{q_1|q_2}}$ describes the shape of the hyperbola and therefore influences its genre-related slope. The parameter $\beta_{\mathrm{q_1|q_2}}$ is the vertical shift. We can estimate $\alpha_{\mathrm{q_1|q_2}}$ by transforming (5) into an approximately linear relation $Y^{\mathrm{lin}}_{q_1|q_2}(\nu) = \nu Y_{q_1|q_2}(\nu)$ (Figure 4(b)) and performing a first order least-squares polynomial fit [11]. Since we are not interested in the vertical shift $\beta_{\mathrm{q_1|q_2}}$ of the hyperbola function we only extract the hyperbolic shape parameter

$$\alpha_{\mathrm{q_1|q_2}} = \frac{\sum_{\kappa=0}^{K/2}\kappa^2\sum_{\nu=0}^{K/2}Y^{\mathrm{lin}}_{q_1|q_2}(\nu) - \sum_{\kappa=0}^{K/2}\kappa\sum_{\nu=0}^{K/2}\nu Y^{\mathrm{lin}}_{q_1|q_2}(\nu)}{\frac{K}{2}\sum_{\nu=0}^{K/2}\nu^2 - \left(\sum_{\nu=0}^{K/2}\nu\right)^2}. \quad (6)$$

After simplifying (6) and omitting constant, genre-independent factors we obtain one *modulation roll-off* feature

$$t_{\mathrm{mr}}(q_1,q_2) = \sum_{\nu=0}^{K/2}(K+1-3\nu)\,Y^{\mathrm{lin}}_{q_1|q_2}(\nu) \quad (7)$$

for each cepstral range, with $t_{\mathrm{mr}}(q_1,q_2) \sim \alpha_{\mathrm{q_1|q_2}}$.

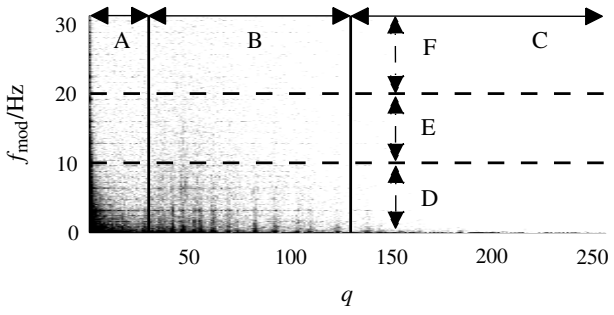
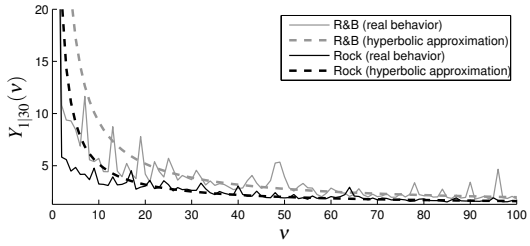
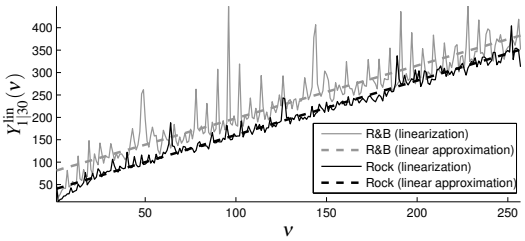


Figure 3: Decomposition of an MCMMS in (A) spectral envelope range, (B) high pitch range, (C) low pitch range, (D) interval of slow modulations, (E) interval of medium modulations and (F) interval of fast modulations for a section of an exemplary Pop music piece.



(a) Modulation roll-off (solid) and hyperbolic approximation (dashed)



(b) Linearized modulation roll-off (solid) and linear approximation (dashed)

Figure 4: Modulation roll-off in the spectral envelope range $1 \leq q \leq 30$ for exemplary R&B (Aaliyah - *More than a woman*) and Rock (Disturbed - *Land of confusion*) songs with $N_{\mathrm{T}} = 960000$, $f_{\mathrm{s}} = 16$ kHz, $N = 512$, $R = 256$, $K = 512$ and $S = 256$.

The nine partitions which are shown in Figure 3 can be summarized, respectively, by means of a two-dimensional average which account for the slow, medium and fast modulation strength. Hence, the *modulation strength* feature is defined as

$$t_{\mathrm{ms}}(q_1,q_2,\nu_1,\nu_2) = \frac{1}{\Delta\nu}\frac{1}{\Delta q}\sum_{\nu=\nu_1}^{\nu_2}\sum_{q=q_1}^{q_2}\overline{X}_{\mathrm{c}}(\nu,q), \quad (8)$$

for each cepstral and modulation frequency range, with $\Delta\nu = \nu_2 - \nu_1 + 1$ and $\Delta q = q_2 - q_1 + 1$.

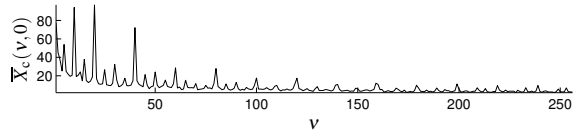
Note, that the features defined above coarsely average cepstral and modulation frequency intervals. We argue that modeling the pitch modulation structure too precisely may lead to overfitting effects and is inevitably connected with a higher number of features to be extracted which increases the computational complexity of the feature extraction and classification process.

4.2.3 Rhythm-Related Features

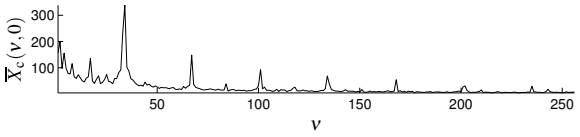
For deriving rhythm-related features from the MCMMS representation we analyze the modulation spectrum of the zeroth cepstral coefficient which is an energy-dependent quantity. Since any onset of musical notes and percussive beats temporarily increases the signal power, the modulation spectrum of the zeroth cepstral coefficient $\overline{X}_{\mathrm{c}}(\nu,0)$, which is depicted in Figure 5 for exemplary Pop and Electronic music sections, is a good measure from which rhythmic information about the underlying music signal can be extracted. Hence, we derive features corresponding to the peaks in Figure 5. For that we consider a beats-per-minute (BPM) range of 40 to 160 which encompasses a variety from slow to uptempo music. This range corresponds to the modulation frequency bin interval $V_1 \in [5,22]$ where we assume $f_{\mathrm{s}} = 16$ kHz, $R = 256$ and $K = 512$. Further, we define three octave bands of this interval $V_2 \in [11,44]$, $V_3 \in [22,87]$ and $V_4 \in [44,175]$ to extract possible harmonics of the fundamental peak. Note, that these values are rounded.

Then, we can extract the position of the fundamental peak

$$t_{\mathrm{rr},1} = \underset{\nu\in V_1}{\arg\max}\,\overline{X}_{\mathrm{c}}(\nu,0) \quad (9)$$



(a) Pop (Kelly Clarkson - *A moment like this*)



(b) Electronic (2Raumwohnung - *Ich und Elaine*)

Figure 5: MCMMS of two exemplary songs for $q = 0$ and $\nu \geq 1$, with $N_T = 960000$, $f_s = 16$ kHz, $N = 512$, $R = 256$, $K = 512$ and $S = 256$.

and the normalized positions of the peaks in the octave bands

$$t_{\mathrm{rr},i} = \frac{1}{t_{\mathrm{rr},1}} \arg\max_{\nu \in V_i} \overline{X}_c(\nu, 0). \tag{10}$$

for $i \in \{2,3,4\}$. These features represent the *rhythmic regularity* of the music signal. The magnitude of the fundamental peak

$$t_{\mathrm{rs},1} = \overline{X}_c(t_{\mathrm{rr},1}, 0) \tag{11}$$

as well as the normalized magnitudes of the other three peaks

$$t_{\mathrm{rs},i} = \frac{\overline{X}_c(t_{\mathrm{rr},i}, 0)}{t_{\mathrm{rs},1}}, \tag{12}$$

with $i \in \{2,3,4\}$, account for the relative *rhythmic strength*.

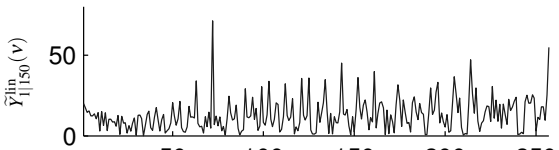
A feature corresponding to the cepstro-temporal regularity of music signals can be found by averaging the MCMMS over a wide cepstral range

$$Y_{1|q_{\max}}(\nu) = \frac{1}{q_{\max}} \sum_{q=1}^{q_{\max}} \overline{X}_c(\nu, q), \tag{13}$$

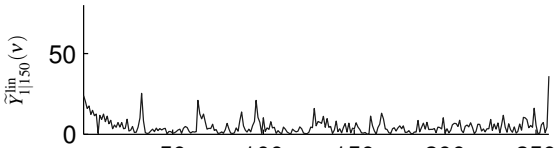
where we set $q_{\max} = 150$. This results in a curve with a hyperbolic trend. After linearizing (13) by $Y^{\mathrm{lin}}_{1|q_{\max}}(\nu) = \nu Y_{1|q_{\max}}(\nu)$ the linearized trend $\widehat{Y}^{\mathrm{lin}}_{1|q_{\max}}(\nu)$ is estimated and removed which yields the deviation from the hyperbolic model $\widetilde{Y}^{\mathrm{lin}}_{1|q_{\max}}(\nu) = Y^{\mathrm{lin}}_{1|q_{\max}}(\nu) - \widehat{Y}^{\mathrm{lin}}_{1|q_{\max}}(\nu)$. This deviation which shows the amount of harmonic modulations is more pronounced for Electronic music than for other music styles which is illustrated in Figure 6. Its strength can be expressed by the mean and variance of its absolute value.

5. CLASSIFICATION OF MUSICAL GENRES

In this section we present the results of a musical genre classification experiment using CMRARE and CMM features and compare their performance to static and dynamic MFCCs as proposed by McKinney and Breebaart in [5] where the latter are referred to as MB_MFCC hereafter. All of these feature sets are based on a modulation analysis of the underlying signal. However, they differ in the resolution of the modulation spectrum. Therefore, this experiment will answer the question if a low-resolution modulation spectrum is sufficient for discriminating musical genres or if it requires highly-resolved modulation frequencies. Only in the latter case a distinctive harmonic modulation structure can be resolved for Electronic music which was demonstrated in Section 3.



(a) Electronic (Prodigy - *Breathe*)



(b) Rock (Nirvana - *Smells like teen spirit*)

Figure 6: Deviation from the hyperbolic model of the cepstrally averaged MCMMS in the cepstral range $1 \leq q \leq 150$ for exemplary Electronic and Rock music.

5.1 Data Set

In this study music files from own and public sources belonging to the musical genres Classical, Electronic, Pop, R&B and Rock are considered where the number of files per genre range between 120 and 180. We decided to merge music from different sources since we observed that publicly available data sets such as [12] do not encompass the large musical variety as in commercially available music. Further, the signal segments provided by Tzanetakis and Cook for carrying out their work in [4] are not long enough to extract features which reliably represent the underlying song.

5.2 Feature Extraction

For each music file we analyze a signal section of one minute taken from the center of the file. This section is sampled at $f_s = 16$ kHz, i.e. we analyze $N_T = 960000$ signal samples. After setting the subframe length and shift to $N = 512$ and $R = 256$, respectively, a spectral analysis (1) is performed. A spectral decomposition is achieved by obtaining the cepstrum (2).

For computing the cepstral modulation spectrum (3) and the MCMMS representation (4) we set $K = 512$ and $S = 256$ if the 22 CMM features as outlined in Section 4.2 are computed subsequently.

For extracting CMRARE features, which are briefly described in Section 4.1, we adopt the parameters settings from [9], i.e. we set $K = 16$ and $S = 8$ to obtain a low-resolution modulation spectrum and compute a CMR for $\nu = 1$ and $2 \leq \nu \leq 8$, respectively. The order of the polynomial is chosen as $p = 10$ which yields 22 features.

Further, the MB_MFCC features are extracted. Here, the modulation spectrum of the first 13 MFCCs is computed as in e.g. [13], based on which the energy in the modulation frequency ranges 1-2 Hz (musical beat rates), 3-15 Hz (speech syllabic rates) and 20-43 Hz (perceptual roughness) is obtained alongside the DC values of the static MFCCs. This approach results in a set of 52 modulation energy features.

5.3 Classification Procedure

The classification of musical genres using the feature sets described in the previous section is performed by means of a linear discriminant analysis (LDA) [14]. We randomly select 100 files per genre and compute the corresponding feature vectors out of which 75 are used for training and 25 for testing. The disjoint training and test sets are cross-validated 50-foldly which ensures that training and

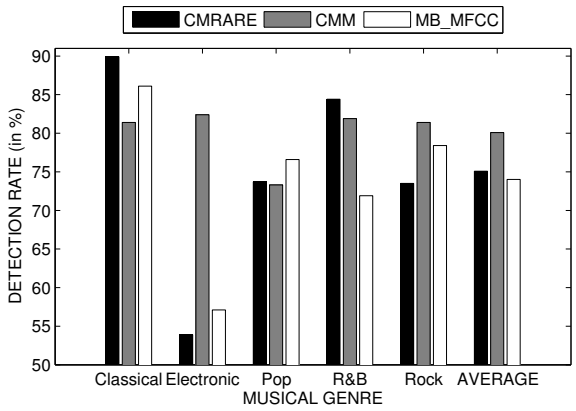


Figure 7: Detection rates for CMRARE features, CMM features as well as static and dynamic MFCCs [5].

test files are exchanged and consequently leads to more significant results. In order to account for all available files, the whole procedure, i.e. random selection of 100 files, training and testing, is repeated 20 times. Finally, we evaluate the average detection rates.

5.4 Classification Results

The detection rates resulting in the classification experiments are depicted in Figure 7 for the five musical genres and the different feature sets. The class averages are shown as well.

On average, CMM features outperform the competing features sets with a detection rate of 81.0%. For CMRARE and MB_MFCC features we obtain 75.1% and 74.0%, respectively.

CMRARE features work better for Classical music and perform equally well for Pop music as CMM and MB_MFCC features. For R&B music their performance is comparable to that of CMM features (84.4% vs. 81.9%). Here, MB_MFCC features perform worse (71.9%). In the case of Rock music, CMM features (81.4%) and MB_MFCC features (78.4%) show better results than CMRARE features (73.5%).

The most striking result is obtained for Electronic music. Here, the classification performance using CMRARE or MB_MFCC features is poor as the detection rates fall below 60%. CMM features, however, significantly improve the detection rate in this case (82.4%). This set of features characterizes the distinctive harmonic modulation structure of Electronic music which can only be represented in a modulation spectrum with highly-resolved modulation frequencies. The results indicate that low-resolution modulation features as the CMRAREs or MB_MFCCs do not bear enough discriminative information to detect Electronic music or other kinds of music with a highly regular cepstro-temporal structure. Therefore, exploiting the fine modulation structure of music signals yields more distinctive features.

6. CONCLUSIONS

The cepstral modulation spectrum is a well-suited common basis from which features are extracted for a musical genre classification task. CMRARE features which are based on a low-resolution modulation spectrum as well as static and dynamic MFCCs [5] which coarsely summarize the modulation energy of MFCCs in four bands do not work well for the detection of Electronic music. It is important to choose a highly-resolved modulation frequency range which ensures that the fine modulation structure of regular and repetitive music styles such as Electronic music is represented. Therefore, the newly proposed CMM features which include characteristics describing the fine modulation structure perform better than CMRARE features as well as static and dynamic MFCCs. As opposed to conventionally used features in music classification which are often derived from different signal domains, CMM features are all derived from a single, unified cepstro-temporal representation. The features are simple, but effective as they are music-related. This guarantees that, after the cepstral modulation spectrum is obtained by means of efficient FFT operations, the computational complexity associated with the feature extraction is fairly low. While static and dynamic MFCCs contain 52 feature values per music file, only 22 CMM features are required to obtain a detection rate of 81% in a 5-class musical genre classification experiment.

REFERENCES

[1] H.-G. Kim, N. Moreau, and T. Sikora, *MPEG-7 Audio and Beyond - Audio Content Indexing and Retrieval*. Wiley, 2005.

[2] B.P. Bogert, M.J.R. Healy, and J.W. Tukey, "The Quefrency Alanysis of Time Series for Echoes: Cepstrum, Pseudo-Autocovariance, Cross-Cepstrum and Saphne Cracking," in *Proc. Symposium on Time Series Analysis*, pp. 209-243, 1963

[3] S.B. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357-366, 1980

[4] G. Tzanetakis and P. Cook, "Musical Genre Classification of Audio Signals," *IEEE Trans. Speech and Audio Processing*, vol. 10, no. 5, pp. 293-302, 2002

[5] M.F. McKinney and J. Breebaart, "Features for Audio and Music Classification," in *Proc. Int. Conf. Music Information Retrieval (ISMIR)*, 2003

[6] B. Logan, "Mel Frequency Cepstral Coefficients for Music Modeling," *Proc. International Symposium on Music Information Retrieval*, 2000

[7] C.-H. Lee, J.-L. Shih, K.-M. Yu, and H.-S. Lin, "Automatic Music Genre Classification Based on Modulation Spectral Analysis of Spectral and Cepstral Features," *IEEE Trans. Multimedia*, vol. 11, no.4, pp. 670-682, 2009

[8] A. Meng, P. Ahrendt, J. Larsen, and L.K. Hansen, "Temporal Feature Integration for Music Genre Classification," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no.5, pp. 1654-1664, 2007

[9] R. Martin and A. Nagathil, "Cepstral Modulation Ratio Regression (CMRARE) Parameters for Audio Signal Analysis and Classification," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, 2009

[10] F. Mörchen, A. Ultsch, M. Thies, and I. Löhken, "Modeling Timbre Distance With Temporal Statistics From Polyphonic Music," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, no.1, pp. 81-90, 2006

[11] J.F. Böhme, *Stochastische Signale: Eine Einführung in Modelle, Systemtheorie und Statistik*. Teubner, 1998

[12] K. West, "Genre Classification from Polyphonic Audio," http://www.music-ir.org/mirex/2005/index.php/ Audio_Genre_Classification

[13] M. Slaney, "Auditory Toolbox: A MATLAB Toolbox for Auditory Modeling Work," *Technical Report 45*, Apple Computer, 1994

[14] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification*. John Wiley & Sons, 2nd edition, 2001