Cepstral Modulation Features for Classifying Audio Data

Anil Nagathil, Timo Gerkmann, and Rainer Martin
Institute of Communication Acoustics (IKA), Ruhr-Universität Bochum, Email: {firstname}.{lastname}@rub.de

Introduction

In the last decade music information retrieval has emerged as a new field of research. As an efficient storage and distribution of digital audio has become feasible since the advent of mp3 and the Internet, techniques for indexing and categorizing automatically the huge amount of available audio data are required.

One of the pioneer works addressing the automatic classification of musical genres is described in [1]. In this contribution the authors use a number of static low-level features but neglect the temporal characteristics of music. In [2] and [3], however, it is shown that dynamic features which are extracted by means of modulation analysis or auto-regressive modeling of static features improve the classification performance as opposed to using simply their means and standard deviations.

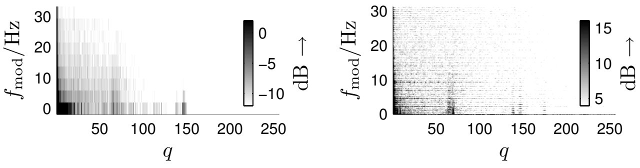
As these approaches are based on features which describe the underlying audio signal roughly, analyzing the temporal evolution of a highly resolved spectral representation first and then extracting dynamic features is likely to be the better strategy to preserve as much information as possible. In this paper we consider the modulation spectrum of a high resolution cepstrum as a common basis for deriving novel features for speech, music and noise discrimination as well as musical genre classification.

Cepstral Modulation Spectrum

After performing a short-term spectral analysis of a raw audio signal section by means of a discrete Fourier transform (DFT), a harmonic decomposition of the log-spectral magnitudes is achieved by the cepstral transformation [4]. Cepstral coefficients correspond to different degrees of spectral detail, i.e. low cepstral coefficients represent the log-spectral envelope whereas higher cepstral coefficients describe the spectral fine structure. Further, it can be assumed that strictly harmonic patterns are mapped on a single cepstral bin. Thus, a separation of spectral envelope and pitch characteristics is achieved.

In order to capture the cepstral dynamics a cepstral short-time modulation spectrum is obtained by means of a sliding window DFT. The complexity of this representation can be reduced by temporally averaging over the total number of short-time magnitude modulation spectra which yields the mean cepstral magnitude modulation spectrum (MCMMS).

In Figure 1 the MCMMS of an exemplary Electronic music section is depicted for two different modulation frequency resolutions. It can be observed that for low cepstral coefficients q (spectral envelope) strong modulations are confined to lower modulation bands. The modulations of pitch (higher cepstral coefficients) are observed for $q \approx 70$ and $q \approx 150$. Along the modulation frequency f_{mod} it can be seen that the high resolution modulation spectrum (Figure 1(b)) exhibits a harmonic modulation pattern which cannot be resolved in the low resolution modulation spectrum (Figure 1(a)).



(a) Low resolution (b) High resolution

Figure 1: MCMMS for Electronic music (Processor - *Nibtal*), with q: cepstral bin index and f_{mod}: modulation frequency

Audio Classification

In this section features derived from the cepstral modulation spectrum are outlined for which classification experiments were carried out. For discriminating audio categories a linear discriminant analysis (LDA) [5] was performed. For conducting the classification experiments we collected representative audio data from public and own sources. In a number of trials, the data sets were randomly split into disjoint training and test sets. For each trial the classifier was trained using the training sets and evaluated using the test data. The final classification results were obtained by averaging over all classification trials.

General Audio Discrimination

If a low resolution modulation spectrum is considered, the cepstro-temporal information content of an audio signal can be summarized by computing *cepstral modulation ratios* (CMR) [6]. Here, we normalize a modulation frequency bin on the zeroth modulation frequency bin. Additionally, higher bins can be averaged along the modulation frequency and normalized on zeroth modulation frequency bin as well. In doing so, more pronounced low frequency modulations are modeled more accurately than less pronounced high frequency modulations. Finally, these CMRs are approximated by a low order polynomial, respectively. The polynomial coefficients which are termed *Cepstral Modulation RAtio REgression* (CMRARE) parameters serve as dynamic audio features.

For a coarse discrimination of audio data into speech, music and environmental noise we applied eight CMRARE features following the classification procedure described above. We achieve detection rates over 95% for music and noise, respectively. For speech almost 99% of the test data is classified correctly. From these findings we can conclude that a small number of CMRARE features is indeed sufficient to discriminate between speech, music and noise.

Musical Genre Classification

While distinguishing speech, music and noise is a relatively easy task, the discrimination of musical genres is obviously more challenging which is due to their fuzzy nature. We applied 22 CMRARE features in a musical genre classification task where we considered Classical, Electronic, Pop, R&B and Rock music. We obtained detection rates between 73% and 90% for all genres except for Electronic music which was only detected correctly in 54% of all cases. This can be attributed to the fact that CMRARE features are not capable of resolving the fine modulation structure of audio signals. The fine modulation structure, however, exhibits more distinct properties for different musical genres. Therefore, a new feature set is introduced which is derived from a high resolution modulation spectrum.

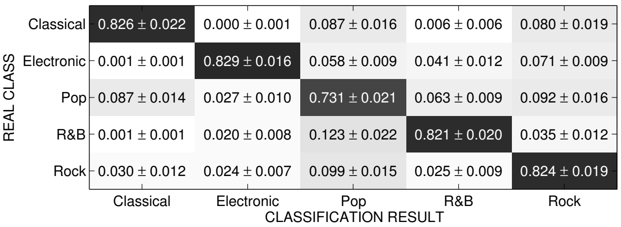
As music can be described in terms of e.g. rhythm, timbre and pitch, the cepstral modulation spectrum is analyzed in order to derive simple but powerful *Cepstral Modulation Music* (CMM) features which describe these characteristics.

As the zeroth cepstral coefficient is a power-related measure, its modulation spectrum captures the dynamic changes which correspond to the rhythmic properties of the music signal. Therefore, we can observe pronounced peaks in the modulation spectrum whose positions and amplitudes can be extracted as descriptors for the rhythmic regularity and strength of the music signal. Further, the strength of the harmonic cepstral modulation pattern of highly regular and repetitive musical styles such as Electronic music (Figure 1(b)) can be described by computing the mean and variance of the absolute value of a cepstrally averaged modulation spectrum.

Since characteristics of the spectral envelope are mapped on low cepstral coefficients, these coefficients are a good representation of timbral properties. For increasing modulation frequencies the modulation spectrum shows a hyperbolic roll-off whose slope is a genre-related measure and thus a feature reflecting the dynamic properties of the spectral envelope. Further, the modulation spectrum of low cepstral coefficients can be coarsely averaged in a few modulation bands which provides information about the strength of slow and fast modulations of the spectral envelope.

The signal processing steps performed for low cepstral coefficients can also be repeated for high cepstral coefficients which reflect pitch properties.

The total number of 22 CMM features were applied in a musical genre discrimination experiment considering the genres and the classification procedure described above. This yields the classification results presented in the confusion matrix in Figure 2. It can be observed that we obtain a detection rate of around 73% for Pop music and more than 82% for all remaining genres. Thus, CMM features which exploit the fine modulation structure of cepstral coefficients outperform CMRARE features on average. Due to the fuzzy nature of musical genres, the performance of musical genre discrimination cannot be expected as being as excellent as in the case of speech, music and noise classification. Particularly for Pop music which contains elements of many different contemporary musical styles it is difficult to find a commonly accepted definition.



REAL CLASS \ CLASSIFICATION RESULT	Classical	Electronic	Pop	R&B	Rock
Classical	0.826 ± 0.022	0.000 ± 0.001	0.087 ± 0.016	0.006 ± 0.006	0.080 ± 0.019
Electronic	0.001 ± 0.001	0.829 ± 0.016	0.058 ± 0.009	0.041 ± 0.012	0.071 ± 0.009
Pop	0.087 ± 0.014	0.027 ± 0.010	0.731 ± 0.021	0.063 ± 0.009	0.092 ± 0.016
R&B	0.001 ± 0.001	0.020 ± 0.008	0.123 ± 0.022	0.821 ± 0.020	0.035 ± 0.012
Rock	0.030 ± 0.012	0.024 ± 0.007	0.099 ± 0.015	0.025 ± 0.009	0.824 ± 0.019

Figure 2: Confusion matrix for five musical genres using 22 CMM features and an LDA classifier.

Conclusions

In this paper an alternative way of extracting dynamic features from audio signals was proposed. First the modulation spectrum of the cepstrum is computed based on which CMRARE or CMM features are obtained for a low or high resolution modulation spectrum, respectively. While CMRARE features excellently discriminate speech, music and noise with detection rates ranging between 95% and 99%, they do not perform well for classifying Electronic music in a musical genre discrimination task as they cannot resolve its highly distinct fine modulation structure. This deficiency is overcome by introducing CMM features for classifying musical genres which yield an overall detection rate of 81%.

References

[1] 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

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

[3] 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

[4] 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

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

[6] 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