HIERARCHICAL AUDIO CLASSIFICATION USING CEPSTRAL MODULATION RATIO REGRESSIONS BASED ON LEGENDRE POLYNOMIALS

Anil Nagathil, Peter Göttel, and Rainer Martin

Institute of Communication Acoustics, Ruhr-Universität Bochum, Bochum, Germany
email: {anil.nagathil,peter.goettel,rainer.martin}@rub.de

ABSTRACT

In this work we present a scalable feature set which is obtained by fitting orthogonal polynomials to the normalized modulation spectrum of cepstral coefficients and which can be easily adapted to different classification tasks. The performance of the feature set is investigated in a hierarchically structured audio signal classification experiment and compared with other approaches reported in the literature. For the root categories speech, music and noise a classification accuracy of 95% is achieved. Subclasses such as male and female speech or different noise types are classified with an accuracy of 95% and 85%, respectively. In a 10-category musical genre discrimination experiment the proposed features exhibit an accuracy of 61%.

Index Terms— cepstral analysis, pattern recognition

1. INTRODUCTION

The automatic organization and classification of audio signals in terms of their underlying acoustical categories can facilitate either the processing or the presentation of audio material in a number of applications. In hearing aids, for instance, an audio classifier can be utilized as a preprocessing tool in order to cope with different hearing scenarios and to choose optimal parameter settings for customized audio programs [1]. This requires hierarchically structured audio signal classification methods which first categorize audio signals in terms of general audio classes and subsequently divide them into subclasses. Typically, approaches for hierarchical audio classification rely on a number of low-level signal descriptors derived from the time, spectral or cepstral domain modeling timbral or rhythmic characteristics of the signal. By means of classification task-related feature selection the classification performance can be optimized further [2, 3]. Other approaches focus on specific classification tasks such as speech/music discrimination or musical genre classification and propose customized features or classification methods for this purpose [4, 5, 6].

In this work we present a generic and scalable feature set which is based on parameterizing the normalized modulation spectrum of cepstral coefficients by means of a polynomial fit. While in [7] a least-squares method was proposed for the polynomial approximation, in this paper the parameterization is performed by means of a Legendre polynomial fit which considerably reduces the computational complexity and facilitates the scaling of the feature set. Based on this feature set a hierarchical audio signal classification experiment is performed considering the root categories speech, music and noise and the subclasses male/female speech, musical genres as well as different noise types. The results show that the proposed method yields comparable classification accuracies as for competing approaches found in the literature.

The paper is organized as follows. In Section 2 we review the computation of cepstral modulation ratios and introduce the approximation based on Legendre polynomials. Section 3 describes the experimental setup followed by a presentation of the classification results for various tasks in Section 4. Our conclusions are summarized in Section 5.

2. CEPSTRO-TEMPORAL SIGNAL ANALYSIS

A section of an audio signal $x(n)$ of length N_T, sampled at the sampling frequency f_s, is segmented into λ_T (possibly overlapping) frames of length N using a tapered window function $w(n)$ such as the Hann window $w(n) = 0.5\,(1 - \cos\,(2\pi n/N))$, with $n = 0, 1, ..., N$. Then, the discrete Fourier transform (DFT) of the weighted frame

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

is computed, where λ, R and μ = 0,1,...,N-1 denote the frame index, the frame shift and the frequency bin, respectively.

A decomposition into coefficients describing the spectral envelope and the spectral fine structure, respectively, is achieved by applying the cepstral transform which is defined as

$$x_c(q, \lambda) = \frac{1}{N} \sum_{\mu=0}^{N-1} \ln(|X(\mu, \lambda)|^2)\, e^{j\frac{2\pi q \mu}{N}}, \tag{2}$$

where q = 0,1,...,N-1 is the index of the cepstral coefficients. Note, that since the log magnitude spectrum is real-valued, the cepstrum is symmetric with respect to the Nyquist bin. Therefore, only cepstral coefficients with indices $q \leq N/2 + 1$ are considered below.

In the final preprocessing stage the temporal evolution of the cepstrum is analyzed. By means of a sliding window DFT we compute the time-varying modulation spectrum of the cepstrum

$$X_c(\nu, q, \Lambda) = \sum_{\kappa=0}^{K-1} x_c(q, \Lambda S + \kappa)\, e^{-j\frac{2\pi \kappa \nu}{K}} \tag{3}$$

where K, S and Λ denote the modulation analysis window length, shift and index, respectively. The modulation frequency bin is specified by ν = 0,1,...,K-1.

The magnitudes of these short-time cepstral modulation spectra are temporally averaged which yields the predominant cepstro-temporal modulation pattern within the time interval corresponding to $\Lambda = 0, 1, ..., \Lambda_T$ where Λ_T is the total number of modulation analysis windows

$$\overline{X}_c(\nu, q) = \frac{1}{\Lambda_T} \sum_{\Lambda=0}^{\Lambda_T - 1} |X_c(\nu, q, \Lambda)|. \tag{4}$$

2.1. Cepstral Modulation Ratios

The cepstral modulation spectrum specified in (4) can be represented approximately by means of cepstral modulation ratios (CMR) where the average of several modulation frequency bands within $\nu_1 \leq \nu \leq \nu_2$ is normalized on the zeroth modulation frequency band

$$r_{\nu_1|\nu_2}(q) = \frac{\sum_{\nu=\nu_1}^{\nu_2} \overline{X}_c(\nu, q)}{(\nu_2 - \nu_1 + 1)\,\overline{X}_c(0, q)}. \quad (5)$$

Note, that for $\nu_2 = \nu_1$, (5) is reduced to the case where a single modulation frequency band $\nu_1 \geq 1$ is normalized on the zeroth modulation frequency band.

2.2. Parameterization of Cepstral Modulation Ratios

CMRs can be parameterized efficiently by means of a polynomial fit. While in [7] it was proposed to compute least-squares regression coefficients, here we approximate CMRs using k-th order Legendre polynomials [8, (8.910.2)] which can be obtained by evaluating

$$P_k(q) = \frac{1}{2^k k!} \frac{d^k}{dx^k}(x^2 - 1)^k, \qquad \forall k \in \mathbb{N}. \quad (6)$$

Hence, a CMR can be represented by a sum of weighted Legendre polynomials up to an approximation order p and an approximation error $e(q)$

$$r_{\nu_1|\nu_2}(q) = \sum_{k=0}^{p} t_{\nu_1|\nu_2,k}\, P_k(q) \; + e(q) \quad (7)$$

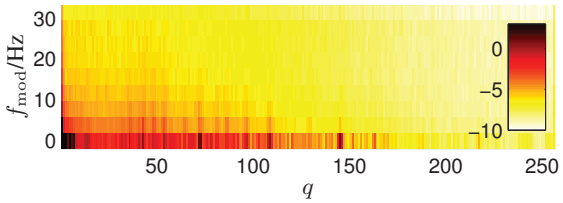
where for $k \in \{0, 1, ..., p\}$ the weights $t_{\nu_1|\nu_2,k}$ are specified by

$$t_{\nu_1|\nu_2,k} = \left(\frac{2k+1}{2}\right) \int_{-1}^{1} r_{\nu_1|\nu_2}(q) P_k(q)\, dq. \quad (8)$$

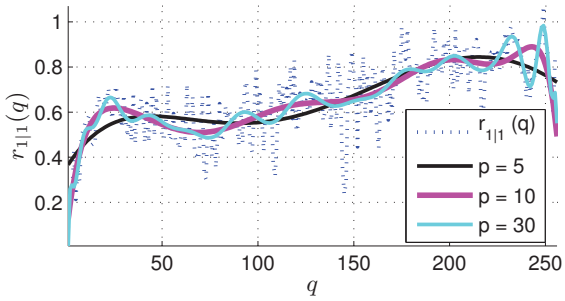
Note, that before applying (8) the interval $q \in \{0, N/2 + 1\}$ has to be shifted and scaled to the range $q \in \{-1, 1\}$. The expression can then be computed, e.g. using trapezoidal numerical integration. The Legendre polynomial coefficients can be computed beforehand and saved in a look-up table. The proposed CMR approximation of order p thus results in $p + 1$ *Legendre Polynomial-based Cepstral Modulation RAtio REgression* (LP-CMRARE) parameters $t_{\nu_1|\nu_2,k}$ which can be used as a set of audio signal descriptors.

Compared to the least-squares procedure [7, 9] the Legendre polynomial fit offers a number of advantages. The former relies on the inversion of a $p \times p$ matrix which on its own already has a computational complexity of $O(p^3)$ when the Gauss-Jordan elimination algorithm is used [10]. Computing the weights of Legendre polynomials, however, only requires a number of operations in the order of $O(p)$ which can be attributed to the intrinsic orthogonality property of Legendre polynomials [8, (7.221.1)]. This further implies that for increasing the approximation order from p to $p + 1$, only a single weighting factor $t_{\nu_1|\nu_2,p+1}$ has to be computed while the existing weights $t_{\nu_1|\nu_2,0}, t_{\nu_1|\nu_2,1}, ..., t_{\nu_1|\nu_2,p}$ can be retained. The least-squares method on the contrary would require solving a new system of equations which, furthermore, can be ill-conditioned if the approximation order p is set to a large value.

As an illustration of the proposed feature extraction method, the temporally averaged cepstral modulation spectrum (4) is shown in Figure 1(a) for a one minute section of an exemplary Pop music song. In Figure 1(b) the CMR $r_{1|1}(q)$, i.e. the ratio of the first and zeroth modulation frequency bands, is plotted along with Legendre polynomial fits of order $p = 5$, $p = 10$ and $p = 30$, respectively, modeling different degrees of cepstral modulation detail.



(a) Temporally averaged cepstral magnitude modulation spectrum (4)



(b) CMR (5) and Legendre polynomial fits of order $p = 5$, $p = 10$ and $p = 30$

Fig. 1. Cepstral modulation analysis of an exemplary Pop music section, with $f_s = 16$ kHz, $N_T = 960000$, $N = 512$, $R = 256$, $K = 16$, $S = 8$ and $\nu_2 = \nu_1 = 1$.

3. CLASSIFICATION EXPERIMENTS

In [7] least-squares CMRARE features were applied in a speech, music and noise discrimination task. In this work we perform a hierarchically structured speech, music and noise classification experiment using the proposed LP-CMRARE features. For that we consider subclasses such as male and female speech, different types of environmental noise as well as different musical genres.

3.1. Database

For performing the classification experiments audio data from publicly available and own sources are merged comprising e.g. [11] or [12] for speech and [13] or the commonly used set proposed in [5] for music. The noise types considered here are babble, household, car, office, subway and other kinds of environmental noise.

3.2. Parameter Settings for Feature Extraction

Throughout all classification tasks the audio analysis is performed at the sampling rate $f_s = 16$ kHz. If not stated elsewise, the duration of the audio section under consideration is $T = 3$ seconds corresponding to $N_T = 48000$ samples. The frame length and shift for the spectral (1) and cepstral (2) analysis are set to $N = 512$ and $R = 256$, respectively, implying a frame rate of $f_{\mathrm{s,mod}} = f_s/R = 62.5$ Hz and a modulation Nyquist frequency of $f_{\mathrm{c,mod}} = 0.5 f_{\mathrm{s,mod}} = 31.25$ Hz. The cepstral modulation analysis (3) and (4) is performed setting the window length and shift to $K = 16$ and $S = 8$, respectively, yielding the temporally averaged cepstral magnitude modulation spectrum. The CMRs $r_{1|1}(q)$ and $r_{2|8}(q)$ as defined in (5) are then approximated using a Legendre polynomial-based p-th order parameterization. The $2(p + 1)$ Legendre polynomial weights yield the LP-CMRARE features. Note, that the cepstral modulation spectrum as well as their CMRs can be considered as a

fixed basis for the hierarchically structured classification experiment as long as the section duration T is not varied. Only p is subjected to classification task-related variations where for an increase of p existing weights can be retained as argued in Section 2.2.

3.3. Classification Concept

Classification is performed in a supervised fashion based on a linear discriminant analysis (LDA) [14] with an underlying multivariate Gaussian feature model. In each classification task the data set consists of at least 100 examples per category out of which 75% are used for training and 25% are used for testing the classifier. The robustness of the results is ensured by means of a 10-fold cross-validation which repetitively allocates the data to disjoint training and test sets, trains the classifier and performs the classification using the test data. The performance of the LP-CMRARE features is evaluated based on the classification accuracy averaged over the 10 cross-validation iterations.

4. RESULTS

4.1. Speech/Music/Noise, Male/Female Speech and Noise Types

For the hierarchically structured classification task general audio classes such as speech, music and noise are considered as root categories. Figure 2 shows the category-averaged classification accuracy for this task as a function of the varying approximation order p based on a cepstral modulation spectrum of a 3 second audio signal section. It can be seen that for the discrimination of speech, music and noise an accuracy beyond 95% is achieved for $p \geq 5$. Moreover, at $p = 5$ which corresponds to 12 LP-CMRARE features the accuracy starts to saturate which makes a further increase of the approximation order unnecessary. The results obtained using LP-CMRARE features are comparable with those achieved in e.g. [3] in which timbral, MPEG-7 and rhythm features were used with a Gaussian mixture model.

A subdivision of speech in terms of male and female speech by means of LP-CMRARE features is performed with an accuracy of around 95% as well. Saturation is achieved for $p = 7$, i.e. 16 LP-CMRARE features. For $p \geq 30$ a decrease in terms of the accuracy is noticeable which can possibly be attributed to an overfitting effect. Note, that as we still consider an audio signal section of 3 seconds duration, the LP-CMRARE features obtained for speech, music and noise discrimination can be retained. The performance of LP-CMRAREs is in line with the results in [15] where normalized audio spectrum envelope features and PCA or ICA projections thereof yield an accuracy ranging from 98.4% to 100% in an HMM framework. In [16] first and second order statistics of male/female speech spectra result in a detection rate of 91% using neural networks.

In case noise was detected as a root category, it can be subclassified in terms of the noise type. The results show an increasing classification accuracy as the approximation order is increased. It begins to saturate at $p = 37$, i.e. 76 LP-CMRARE features. On average a detection rate of 85% can be achieved when the classifier is trained and tested for household, babble, car, office and subway noise. In [17] babble, car, bus, factory and street noise signals were classified based on line spectral frequencies using a quadratic Gaussian classifier which show a comparable accuracy of 86%.

A detailed performance evaluation of the noise type categorization is provided by the confusion matrix presented in Figure 3 for the case $p = 37$ which is marked by a circle in Figure 2. While for

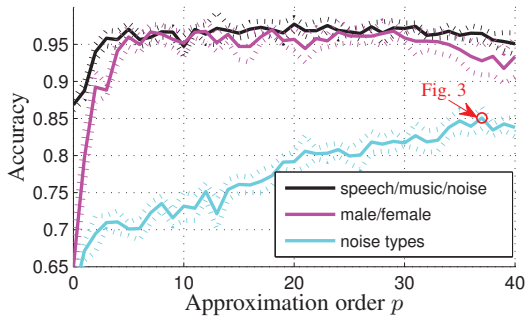


Fig. 2. Class-averaged accuracy (solid) and standard variations (dotted) in speech/music/noise, male/female speech and household/babble/car/office/subway noise discrimination tasks vs. approximation order.

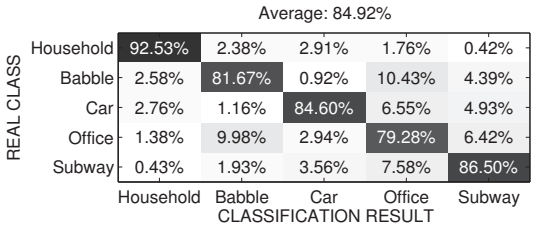


Fig. 3. Confusion matrix for five different noise types using 76 LP-CMRARE features, with the approximation order $p = 37$.

speech, music and noise classification as well as for male and female speech discrimination Figure 2 suggests a low amount of misclassification, the noise types are confused more often with each other. This hints at the complexity of the classification task given that categories such as babble or office noise which exhibit a considerable confusion indeed bear a strikingly audible resemblance.

4.2. Discrimination of Musical Genres

While different noise types often can still be discriminated well by human listeners, musical genres are inherently of a more fuzzy nature. In Figure 4 the classification accuracy for a 10-category musical genre discrimination task based on the data set introduced in [5] is shown as a function of the LP-CMRARE approximation order. These results were obtained considering a signal section of duration $T = 3$ seconds, which was the setting in Section 4.1, $T = 10$ seconds and $T = 30$ seconds, respectively. For $T = 3$ seconds we observe a poor performance with an average accuracy below 45%. However, when T is increased up to maximal duration of $T = 30$ seconds, which requires a re-computation of the cepstral modulation spectrum, the performance of LP-CMRAREs considerably improves as can be seen in Figure 4. For $p = 12$ the average detection rate is 61%. This result is comparable to that obtained in [5] in which a number of different low-level features were utilized or [6] which processes these features using diffusion maps. A closer look at the confusion matrix in Figure 5 for a set of LP-CMRARE features with $p = 12$ indicates a good classification accuracy for classical (93%) and metal music (83%). Yet the full-fledged musical genre classification task may require a more detailed spectro-temporal representation [18] or features tailored to music signals [9].

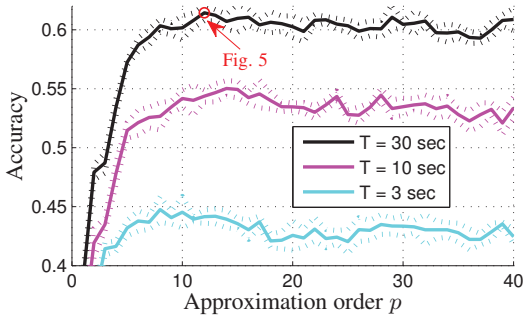


Fig. 4. Class-averaged accuracy (solid) and standard variations (dotted) in a 10-category musical genre discrimination task vs. approximation order for different signal durations.

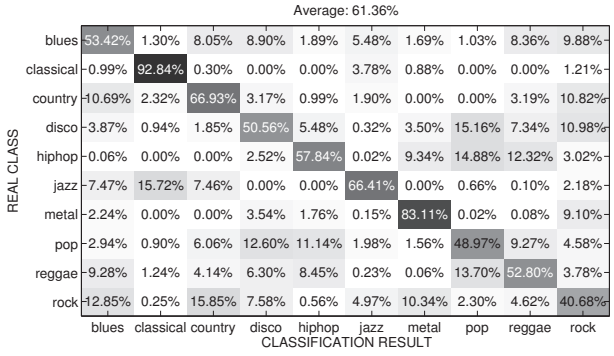


Fig. 5. Confusion matrix for 10 musical genres using 26 LP-CMRARE features, with approximation order $p = 12$.

5. CONCLUSIONS

LP-CMRARE parameters which are based on a Legendre-polynomial fit of cepstral modulation ratios are introduced as a generic and scalable feature set for audio signal classification. Compared to least-square CMRARE features [7] they require a considerably reduced number of operations which narrows down the complexity mainly to the computation of the cepstral modulation spectrum. The latter, however, can be obtained by efficient FFT algorithms. Moreover, for a hierarchically structured classification task which considers the root categories speech, music and noise as well as the subclasses male and female speech or different noise types it was shown that classification accuracies can be achieved which are comparable to those reported in the literature. For these problems the cepstral modulation spectrum only needs to be computed once while only the LP-CMRARE approximation order needs to be varied. Musical genre classification with many genres is a much more sophisticated classification problem which puts LP-CMRARE features at their limits when only three seconds of a music signal are considered. By means of increasing the duration of the analyzed signal section the classification accuracy can be improved and meets the results stated in the literature.

6. ACKNOWLEDGMENT

This work is funded by the German Research Foundation (DFG), Sonderforschungsbereich 823, Teilprojekt B3.

7. REFERENCES

[1] V. Hamacher, J. Chalupper, J. Eggers, E. Fischer, U. Kornagel, H. Puder, and U. Rass, "Signal Processing in High-End Hearing Aids: State of the Art, Challenges, and Future Trends," *EURASIP Journal on Applied Signal Processing*, vol 18, pp. 2915-2929, 2005.

[2] T. Zhang and C.-C. Jay Kuo, "Hierarchical Classification of Audio Data for Archiving and Retrieving," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 1999.

[3] J.J. Burred and A. Lerch, "A Hierarchical Approach to Automatic Musical Genre Classification," in *Proc. 6th Int. Conf. Digital Audio Effects (DAFx)*, 2003.

[4] E. Sheirer and M. Slaney, "Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 1997

[5] 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.

[6] M. Genussov and I. Cohen, "Musical Genre Classification of Audio Signals Using Geometric Methods," in *Proc. European Signal Processing Conference (EUSIPCO)*, pp. 497-501, 2010

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

[8] I. Gradshteyn and I. Ryzhik, *Table of Integrals, Series, and Products*, 6th edition, Academic Press, 2000.

[9] I. Vatolkin, W. Theimer, and M. Botteck, "Amuse (Advanced Music Explorer) - A Multitool Framework for Music Data Analysis," in *Proc. 11th International Society for Music Information Retrieval Conference (ISMIR)*, 2010.

[10] T.J. Dekker and W. Hoffmann, "Rehabilitation of the Gauss-Jordan algorithm," *Numerische Mathematik*, vol. 54, pp. 591-599, 1989.

[11] W.M. Fisher, G.R. Doddington, and K.M. Goudie-Marshall, "The DARPA Speech Recognition Research Database: Specifications and Status," in *Proc. of DARPA Workshop on Speech Recognition*, pp. 93-99, 1986.

[12] P. Kabal, "TSP Speech Database," http://www-mmsp.ece.mcgill.ca/ Documents/Data/index.html.

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

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

[15] H.-G. Kim, N. Moreau, and T. Sikora, "Audio Classification Based on MPEG-7 Spectral Basis Representations," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 14, no.5, pp. 716-725, 2004.

[16] H. Harb and L. Chen, "Gender Identification Using a General Audio Classifier," in *Proc. IEEE Int. Conf. on Multimedia and Exposition (ICME)*, vol. 2, pp. 733-736, 2003.

[17] K. El-Malah, A. Samouelian, and P. Kabal, "Frame-level Noise Classification in Mobile Environments," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, pp. 237-240, 1999.

[18] A. Nagathil, T. Gerkmann, and R. Martin, "Musical Genre Classification Based on a Highly-resolved Cepstral Modulation Spectrum," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2010.