CEPSTRAL MODULATION RATIO REGRESSION (CMRARE) PARAMETERS FOR AUDIO SIGNAL ANALYSIS AND CLASSIFICATION

Rainer Martin and Anil Nagathil

Institute of Communication Acoustics, Ruhr-Universität Bochum, Bochum, Germany
email: firstname.lastname@rub.de

ABSTRACT

In this paper we propose a new set of parameters for audio signal analysis and classification. These parameters are regressions computed on the normalized modulation spectrum of high-resolution cepstral coefficients. The parameter set is scalable in its size and gives a compact representation of the modulation content of speech and other audio signals. These parameters as well as the regression approximation error are well suited for characterizing audio signals in a unified framework. In particular we use a set of eight parameters in a speech/music/noise classification task in which we achieve a classification accuracy which compares very well with other approaches including static and dynamic MFCCs.

Index Terms— cepstrum, modulation spectrum, signal classification

1. INTRODUCTION

Cepstral coefficients [1] and variations thereof have become an indispensable tool in many speech processing applications. Mel-frequency cepstral coefficients (MFCC) [2], for instance, are widely used as acoustic features in automatic speech recognition as they provide a compact representation of the log-spectral envelope of speech signals. Moreover, the first and second order temporal derivatives have been found to be useful to capture their temporal dynamics. Feature vectors composed of MFCCs (aka static features) and their derivatives (aka dynamic features) frequently serve as the basis for acoustic models in speech and speaker recognition tasks. It has been found in automatic speech recognition experiments that generalized dynamic features using a more comprehensive modulation representation [3] and temporal normalization based on the modulation spectrum of cepstral features [4] improve speech recognition performance. In fact, since cepstral coefficients encode the coarse and the fine structure of the log-magnitude spectrum, their modulation spectrum reveals interesting details of the signals under investigation.

In this paper we propose new parameters derived from the modulation spectrum of high-resolution cepstral coefficients. Unlike MFCCs we do not group the cepstral coefficients in mel-scaled frequency bands and thus preserve the harmonic fine-structure of the signal. We compute a sliding-window modulation spectrum of these fully resolved cepstral coefficients and model their average modulation content vs. quefrency by means of a low-order polynomial. Thus, these coefficients model the temporal and spectral modulations of the signal under investigation.

The newly derived features are then applied to a speech/music/-noise classification task. This task has become of interest in a variety of applications such as acoustic scene analysis in hearing instruments and broadcast content segmentation and transcription. In previous works a variety of features, such as four Hertz modulation energy, MFCCs, roll-off of the spectrum, spectral entropy and zero-crossing rate have been investigated [5, 6, 7, 8]. McKinney and Breebaart [9], for instance, compare several feature sets, including MFCCs, in a common feature extraction framework. While they emphasize the importance of features based on temporal modulations they achieve best results with a mixture of static and dynamic features. Meng et al. [10] model the temporal evolution of MFCCs in terms of univariate or multivariate autoregressive processes. Furthermore, the cepstrum of the modulation spectrum was investigated in [11] with good success for the purpose of discriminating between speech and environmental noise.

Unlike the above works we do not use a variety of different features but develop a unified and flexible framework on the basis of high-resolution cepstral coefficients. In our speech/music/noise discrimination task we achieve a performance that compares very well with the error rates reported in the literature. The modulation spectrum of the high-resolution cepstrum and its parametric approximation may also explain why MFCCs work fairly well in less complex speech/music discrimination tasks [6, 12] and where possible limitations of MFCC features may be found.

The remainder of this paper is organized as follows: In the next section we outline the computation of the cepstral modulation spectrum and the regression parameters. In Section 3 we describe our classification experiment. We provide experimental results for our new features in Section 4 and compare these to results obtained with standard static and dynamic MFCCs.

2. CEPSTRAL MODULATION REGRESSION PARAMETERS

In this section we describe the computation of cepstral modulation regression parameters. Before these low-order polynomial parameters can be computed, three preprocessing stages have to be performed. First, a section of an audio signal $x(n)$, henceforth called *frame*, is segmented into λ_T (possibly overlapping) *subframes* of length N using the Hann window $w(n) = 0.5\,(1 - \cos\,(2\pi n/N))$. Then, the discrete Fourier transform (DFT) of the weighted subframe

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

is computed, where λ, R and $\mu = 0,1,...,N\text{-}1$ denote the subframe index, the subframe shift and the frequency bin, respectively. Subsequently, the inverse discrete Fourier transform (IDFT) of the logarithmic squared magnitude spectrum, also known as the cepstrum, is

obtained

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

where q = 0,1,...,N-1 is the quefrency index of the cepstral coefficients. In the final preprocessing stage the spectro-temporal evolution of the cepstrum is analyzed. By means of a sliding window DFT we compute the time-varying modulation spectrum of the cepstrum, the short-time cepstral modulation spectrum. Starting at subframe-index $\lambda = \Lambda S$ the sliding window considers K consecutive subframes

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

with the window shift parameter S and the modulation frequency bin ν = 0,1,...,K-1. Its properties vary for speech, music and noise.

To compute the regression parameters we determine the temporal average of the magnitude of the modulation spectra in each cepstral bin across the whole frame. Subsequently, the mean of each modulation frequency band is normalized on the mean of the zeroth modulation frequency band

$$\begin{aligned} r_\nu(q) &= \frac{\sum_{\ell=1}^{\Lambda_T} |X_c(\nu,q,\ell)|}{\sum_{\ell=1}^{\Lambda_T} |X_c(0,q,\ell)|} \\ &\overset{!}{\approx} \sum_{k=0}^{p} t_{\nu,k}\, q^k \quad \forall \quad q \in \{1,...,N/2\}, \end{aligned} \tag{4}$$

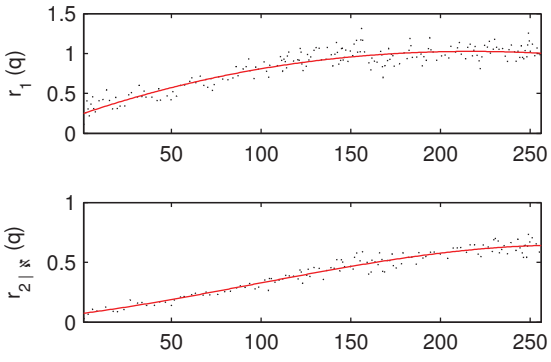
where Λ_T denotes the total number of magnitude modulation spectra. For $q = 1 \ldots N/2$ the cepstral modulation ratios (CMR) $r_\nu(q)$ can then be approximated using a polynomial of order p and a standard least-squares procedure which results in $p+1$ parameters $t_{\nu,0}$, $t_{\nu,1}$, $\ldots$ $t_{\nu,p}$. In (4) we favor the average magnitude over the average squared magnitude (power) because the squared magnitude would implicitly increase the order of the polynomial approximation. Furthermore, since we do not include the cepstral bin $q = 0$ the approximation is entirely independent of the signal's recording level.

Alternatively, it can be also of interest to compute the average of the means for several modulation bands $\nu_1 \ldots \nu_2$, e.g. bands $2 \ldots K/2$ and relate this average to the means of the zeroth band

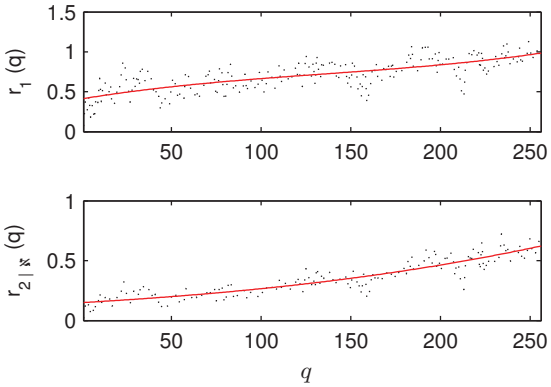
$$\begin{aligned} r_{\nu_1|\nu_2}(q) &= \frac{\sum_{\nu=\nu_1}^{\nu_2}\sum_{\ell=1}^{\Lambda_T} |X_c(\nu,q,\ell)|}{(\nu_2-\nu_1+1)\sum_{\ell=1}^{\Lambda_T} |X_c(0,q,\ell)|} \\ &\overset{!}{\approx} \sum_{k=0}^{p} t_{\nu_1|\nu_2,k}\, q^k \quad \forall \quad q \in \{1,...,N/2\}. \end{aligned} \tag{5}$$

Again, a polynomial fit of order p delivers coefficients $t_{\nu_1|\nu_2,0}$, $t_{\nu_1|\nu_2,1}$, $\ldots$ $t_{\nu_1|\nu_2,p}$. By varying the number of cepstral modulation bands and the order of the approximating polynomials these *Cepstral Modulation RAtio REgression* (CMRARE) features provide a unified yet flexible framework for characterizing audio signals.

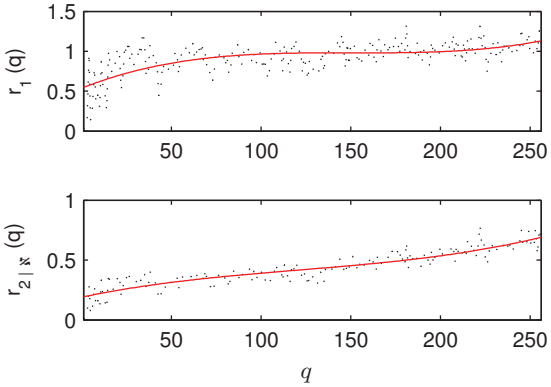
Figure 1 depicts the averaged modulation cepstrum ratios $r_1(q)$ and $r_{2|\aleph}(q)$ with $\nu_2 = \aleph = K/2$ and the corresponding polynomial approximation for a speech, music and noise signal. Here, $p = 3$ and $N = 512$ was chosen. Clearly, the third order polynomial approximation provides a good fit to the overall shape of the CMR vs. q curves and the signals can be discriminated on the basis of their polynomial parameters. Furthermore, we note that the variance of the approximation error, especially for $q < 100$, is significantly smaller for the speech signal than for the other signal classes. Thus, we may observe at this point that for speech signals a "smoothed" representation such as provided by MFCCs will work very well [12, 9]. This is not necessarily the case for music and noise.



(a) male speaker



(b) music (Oasis - "Live Forever")



(c) subway noise

Fig. 1. Cepstral modulation ratios (dotted) and the polynomial approximation of order $p = 3$ (solid) vs. quefrency index q for three audio examples. $N = 512$, $K = 16$, $S = 1$, $\Lambda_T = \lambda_T = 766$.

3. THE SIGNAL CLASSIFICATION EXPERIMENT

In this section we apply the CMRARE features in a speech/music/-noise classification task and compare results to classification using static and dynamic MFCCs. We describe the data set, the parameter settings for the feature extraction as well as the classification approach itself.

3.1. Data set

In order to establish heterogeneous speech, music and noise subsets we merge audio data from a variety of independent commercial, public and own sources (e.g., [13, 14, 15]) for each class. We consider Electronic, Metal, Pop, Punk, R&B, Rock, Classical, Jazz and World music, speech data from several hundred male and female speakers as well as various noise types such as babble, construction, household, office, car, subway and traffic noise. The sound files are available in the form of raw sampled data or compressed mp3 files.

3.2. Feature extraction

The CMRARE and MFCC features are obtained by analyzing frames of 50000 signal samples at a sampling rate of 16 kHz. For short-term spectral analysis these 50000 samples are subdivided into $\lambda_T = 766$ subframes which are then used for modulation analysis and smoothing. For music files only one frame (50000 signal samples) in the middle of the respective sound file is extracted, whereas for speech and noise data we consider all frames in a file.

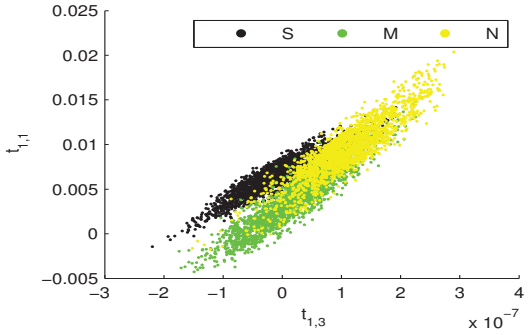
To compute CMRARE parameters a spectral analysis (1) is performed, where the DFTs of all subframes with $N = 512$ samples and a subframe shift of $R = 64$ samples are determined. Then, the cepstrum as defined in (2) is obtained. Subsequently, we compute the short-time cepstral modulation spectrum (3) for all subframes λ (i.e., $S = 1$) where we set $K = 16$. From the average modulation spectrum we compute eight regression parameters by fitting polynomials of order $p = 3$ to the normalized average modulation spectrum of the first modulation band and the normalized sum of the remaining modulation bands. Thus, we obtain the regression parameters $t_{1,0}$, $t_{1,1}$, $t_{1,2}$ and $t_{1,3}$ as well as $t_{2|\aleph,0}$, $t_{2|\aleph,1}$, $t_{2|\aleph,2}$ and $t_{2|\aleph,3}$, as defined in (4) and (5) and with $\nu_2 = \aleph = K/2$. Fig. 2 shows two examples of the empirical joint distributions of two parameters. It can be seen that there is a good degree of class separability.

For classification with MFCCs we compute 13 averaged static MFCCs and their normalized average modulation content similar to (3), (4) and (5). In this case a polynomial approximation is not necessary as we already have a small number of smoothed features.

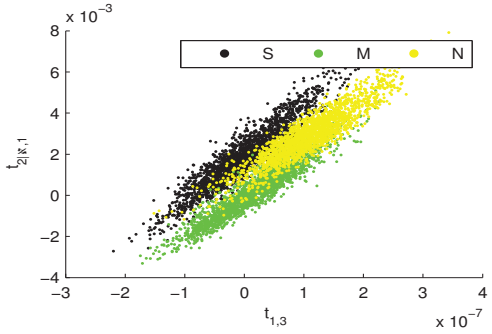
3.3. Classification

Once we have computed the feature vectors (CMRARE or MFCC) for all files of our set of speech, music and noise data, a linear discriminant analysis (LDA) [16] is performed in order to classify the audio material. We also investigated the use of a quadratic discriminant analysis (QDA) but found no advantage when CMRARE features based on the average magnitudes as defined in (4) and (5) are employed.

To keep the number of frames in balance for all classes within a classification experiment, we selected randomly 2000 frames of each class from the available data. In the next step the randomly selected frames are divided into disjoint training and test sets with 1500 and 500 frames, respectively. Under these conditions a 50-fold cross-validation (CV) procedure is performed to examine the robustness



(a) $t_{1,1}$ vs. $t_{1,3}$



(b) $t_{2|\aleph,1}$ vs. $t_{1,3}$

Fig. 2. Exemplary scatter plots of two CMRARE parameters for speech (S), music (M) and noise (N).

of the detection rates for the given subset of signal frames. In order to investigate the dependence of the classification results on the random preselection of 2000 frames per class the procedure above is repeated many times.

4. CLASSIFICATION RESULTS

The procedure, i.e. preselection and CV, is iterated 100 times. We found standard deviations in the order of 0.005 to 0.01 from which we conclude that the detection rates hardly depend on how the data is allocated to the training and test sets. Even more important to note is that apparently the detection rates are also robust against the random preselection of frames which are included in the classification experiment, i.e. they hardly vary across the 100 iterations. Thus, even if heterogeneous frames are selected, i.e. frames which correspond to different noise situations or speakers, the classification is not degraded. This oberservation can also be made by considering the values in Table 1, where the averaged confusion matrix for all 100 iterations is depicted. The low standard deviations hint at the robustness of our features against heterogeneous training and test material. The confusion matrix shows that detection rates between 96% for noise and almost 99% for speech are achieved.

Frames with the length of 50000 samples, which correspond to about three seconds at a sample rate of 16 kHz, obviously contain enough information to perform a reliable classification of speech, music and noise on the basis of the CMRARE features. Furthermore, only the combination of the two quadruples of regression parameters in (4) and (5) together provide good classification results. If classification is performed only using one of these quadruples, the

		classification result		
		S	M	N
real class	S	**0.986 ± 0.002**	0.000 ± 0.000	0.014 ± 0.002
	M	0.006 ± 0.001	**0.970 ± 0.002**	0.024 ± 0.001
	N	0.006 ± 0.001	0.031 ± 0.003	**0.963 ± 0.003**

Table 1. Confusion matrix for speech (S), music (M) and noise (N) using 8 CMRARE features, mean and standard deviations of 100 iterations, 50000 signal samples per frame.

detection rates degrade significantly.

To compare these results to well established procedures Tables 2 and 3 show classification scores using 13 static MFCCs only and 39 static and dynamic MFCCs, respectively. Both the static and the complete MFCC feature sets give excellent results for speech. The static MFCCs, however, do not work well for music detection. Even using all 39 MFCC features, the performance for music is slightly worse than when using the eight CMRARE features.

		classification result		
		S	M	N
real class	S	**0.981 ± 0.002**	0.019 ± 0.002	0.000 ± 0.000
	M	0.085 ± 0.002	**0.799 ± 0.004**	0.116 ± 0.004
	N	0.005 ± 0.001	0.06 ± 0.004	**0.935 ± 0.004**

Table 2. Confusion matrix for speech (S), music (M) and noise (N) using 13 static MFCC features. Mean and standard deviations of 100 iterations, 50000 signal samples per frame.

		classification result		
		S	M	N
real class	S	**0.999 ± 0.001**	0.001 ± 0.001	0.000 ± 0.000
	M	0.035 ± 0.001	**0.943 ± 0.002**	0.022 ± 0.002
	N	0.002 ± 0.001	0.034 ± 0.003	**0.964 ± 0.003**

Table 3. Confusion matrix for speech (S), music (M) and noise (N) using 39 static and dynamic MFCC features. Mean and standard deviations of 100 iterations, 50000 signal samples per frame.

5. CONCLUSIONS

The proposed CMRARE feature extraction method provides interesting insights into the temporal variations of cepstral coefficients and hints towards limits of MFCC representations in the context of music and noise analysis. The procedure computes a modulation spectrum on highly resolved cepstral coefficients and derives smoothed features only in the final step. We show that these features lead to a high signal classification accuracy in a speech/music/noise classification task and, compared to much larger feature sets based on both static and dynamic MFCCs, to similar detection rates. As there are no "static" features in our feature set we hereby emphasize the importance of temporal modulations. Furthermore, these features are entirely independent of the long-term signal power which is an important requirement for any audio classification scheme. Future work will focus on the extension to more classes as it is required, e.g., for genre classification.

6. REFERENCES

[1] 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 Saphe Cracking," in *Proc. of the Symposium on Time Series Analysis*, 1963, pp. 209–243.

[2] 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.

[3] N. Kanedera, H. Hermansky, and T. Arai, "Desired Characteristics of Modulation Spectrum for Robust Automatic Speech Recognition," in *Proc. IEEE Intl. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, 1998, vol. 2, pp. 613–616.

[4] C.-A. Pan, C.-C. Wang, and J.-W. Hung, "Improved Modulation Spectrum Normalization Techniques for Robust Speech Recognition," in *Proc. IEEE Intl. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, 2008, pp. 4089–4092.

[5] J. Saunders, "Real-time Discrimination of Broadcast Speech/Music," in *Proc. IEEE Intl. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, 1996, pp. 993–996.

[6] J.T. Foote, "Content-based Retrieval of Music and Audio," in *Multimedia Storage and Archiving Systems II, Proc. of SPIE*, 1997, pp. 138–147.

[7] M.J. Carey, E.S. Parris, and H. Lloyd-Thomas, "A Comparison of Features for Speech, Music Discrimination," in *Proc. IEEE Intl. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, 1999, pp. 149–152.

[8] A. Pikrakis, T. Giannakopoulos, and S. Theodoridis, "A Speech/Music Discriminator of Radio Recordings Based on Dynamic Programming and Bayesian Networks," *IEEE Trans. on Multimedia*, vol. 10, no. 5, pp. 846–857, 2008.

[9] M.F. McKinney and J. Breebaart, "Features for Audio and Music Classification," in *Proc. of the International Conference on Music Information Retrieval (ISMIR2003)*, 2003.

[10] 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.

[11] T. Miyoshi, T. Goto, T. Doi, T. Ishida, T. Arai, and Y. Murahara, "Modulation Cepstrum Discriminating between Speech and Environmental Noise," *Acoust Sci Technol*, vol. 25, no. 1, pp. 66–69, 2004.

[12] B. Logan, "Mel frequency cepstral coefficients for music modeling," in *Proc. International Symposium on Music Information Retrieval*, 2000.

[13] 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*, 1986, pp. 93–99.

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

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

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