



Reduction of reverberation effects in the MFCC modulation spectrum for improved classification of acoustic signals

Sebastian Gergen, Anil Nagathil, Rainer Martin

Institute of Communication Acoustics
Ruhr-Universität Bochum, Germany

{sebastian.gergen, anil.nagathil, rainer.martin}@rub.de

Abstract

The classification of acoustic signals is an important step in many audio signal processing algorithms, e.g. in the context of speech enhancement, speech recognition, and others. Signals which are captured for classification are often degraded by an unknown amount of reverberation in a real environment. If a classifier is trained on clean and anechoic data, a mismatch between training and test conditions results in a reduced classification accuracy. In this paper, we introduce a novel equalization gain matrix which can be applied to modulation domain audio features. This gain is designed to counteract the modifications which originate from reverberation such that the mismatch between clean training data and degraded test data is reduced. Experiments show that the classification accuracy can be increased significantly for reverberant signals.

Index Terms: Classification, Reverberation, MFCC

1. Introduction

The environmental conditions in which acoustic signal classification systems are used are in general not known and may change over time. One important factor which has to be considered in this context is reverberation. Reverberation results in a mismatch between the clean and anechoic training of a classification system and the test data. This mismatch leads to a reduced accuracy for signal classification or speech recognition [1, 2]. To improve the performance of an algorithm which suffers from the influence of reverberation, one can either account for the mismatch in the design of the classification model (back-end method), or try to reduce the degrading influence in the actual signal sample to be classified (front-end method), while training the classifier on clean and anechoic data. Signal dereverberation is often performed by means of iterative and adaptive acoustic channel estimation and equalization techniques [3]. In [4], these methods are applied for instance in an automatic speech recognition task. In contrast to the dereverberation of the actual signal, in our contribution we introduce a front-end approach that reduces reverberation related distortions in the extraction phase of audio features.

The feature set that we use in our classification task is based on the modulation spectrum of the Mel-frequency cepstral coefficients (MFCCs). Modulation domain features have been successfully used in several works before [5, 6, 7].

The modifications due to the transmission from the source to the receiver in a reverberant room are of convolutive nature and have an impact on the modulation characteristics of a signal. In earlier investigations on automatic speech recognition [8, 9], it was shown that filtering a reverberant signal in the spectral domain can be successfully used to restore temporal

characteristics of a speech signal which is degraded by reverberation. In our work, we extend this concept by applying a gain function dependent on the modulation frequency to the modulation spectrum of the MFCCs. In contrast to previous methods where one and the same high-pass or bandpass filter was used, this gain is optimized for each individual MFCC. The respective MFCC- and modulation frequency-dependent gains are estimated in a training step and averaged over several rooms and source-microphone placements.

This paper is structured as follows: in Section 2, we introduce the feature set that we use in our classification experiments. Then, in Section 3, we illustrate the effect of reverberation on the feature computation and introduce an equalization gain matrix which can be applied in the course of the feature extraction. The evaluation setup and classification experiments are described in Section 4. The classification results are presented and discussed in Section 5. Section 6 concludes the paper.

2. Mod-MFCC feature extraction

For classification tasks, captured audio data is usually transformed into a low-dimensional representation, which allows to distinguish different classes. Different types of audio features were introduced and evaluated in various classification experiments, e.g., for clean speech, speech in noise, noise and music classification [10, 11, 5, 12], or music genre classification [13, 14]. For our investigation we consider a cepstro-temporal representation of the signal [6] which is based on the MFCCs [15] of a signal. We call the feature set Modulation Mel-Frequency Cepstral Coefficients (Mod-MFCCs).

To compute the Mod-MFCC features, we use the short-time Fourier transform (STFT) representation $X(k, b)$ of the signal $x(t)$, where k and b denote the frequency bin and time frame index, respectively. Now, the squared magnitude spectrum is mapped onto the Mel scale [16], which results in the Mel-spectrum $X_{\text{mel}}(k', b)$, where $k' = 0, 1, \dots, K' - 1$ is the index of the Mel scale frequency bin. Then, the MFCCs $X_{\text{mfcc}}(\eta, b)$ with the cepstral coefficient index $\eta = 0, 1, \dots, K' - 1$ are calculated by computing the discrete cosine transform of the logarithm of the absolute Mel-spectrum. We compute the short-time MFCC modulation spectrum $\hat{X}_{\text{mfcc}}(\nu, \eta, c)$ using a sliding window discrete Fourier transform (DFT):

$$\hat{X}_{\text{mfcc}}(\nu, \eta, c) = \sum_{\ell=0}^{L-1} X_{\text{mfcc}}(\eta, cR + \ell) e^{-j \frac{2\pi \ell \nu}{L}}, \quad (1)$$

where, starting at sub-frame index $b = cR$, the sliding window considers L consecutive frames. The modulation frequency bin index is specified by $\nu = 0, 1, \dots, L/2$ and c and R denote the temporal modulation window index and shift, respectively,

10.21437/Interspeech.2015-438

with $c = 0, 1, \dots, C - 1$ [6, 1]. We aim to generate a feature representation which behaves rather stationary in comparison to short-time audio features. Therefore, the absolute values of the modulation spectra are averaged over all C frames (2) and cepstral modulation ratios (CMR) $\rho_{\nu_1|\nu_2}(\eta)$ are computed in (3) to summarize the modulation spectrum,

$$\tilde{X}_{\text{mfcc}}(\nu, \eta) = \frac{1}{C} \sum_{c=0}^{C-1} |\hat{X}_{\text{mfcc}}(\nu, \eta, c)| \quad (2)$$

$$\rho_{\nu_1|\nu_2}(\eta) = \frac{\sum_{\nu=\nu_1}^{\nu_2} \tilde{X}_{\text{mfcc}}(\nu, \eta)}{(\nu_2 - \nu_1 + 1) \tilde{X}_{\text{mfcc}}(0, \eta)}. \quad (3)$$

Note, that in (3) the average of several modulation frequency bands within $\nu_1 \leq \nu \leq \nu_2$ is normalized on the zeroth modulation frequency band. If $\nu_1 = \nu_2 \geq 1$, this reduces to a single modulation frequency band that is normalized by the zeroth band. In addition to CMRs, we further compute $\bar{X}_{\text{mfcc}}(\eta)$ for the feature vector, where

$$\bar{X}_{\text{mfcc}}(\eta) = \frac{1}{L} \sum_{\nu'=0}^{L-1} \tilde{X}_{\text{mfcc}}(\nu', \eta) \quad (4)$$

is the modulation spectrum averaged over all modulation frequencies ν for each MFCC bin η .

For the work that is presented here, we sample microphone signals at $f_s = 16$ kHz and process them in blocks of $T = 4$ seconds duration. We extract Mod-MFCC features and apply cepstral mean normalization (CMN) [17, 18], for the training and test data of a classification experiment. CMN reduces the influence of slowly varying characteristics of the transmission channel or overall variations of the power, as both result in a constant additive offset in the MFCC computation. For the spectral and cepstral analysis, the frame length is 512 and frame shift is 256 samples. For the cepstral modulation analysis the frame length and shift is set to $L = 16$ and $R = 8$, respectively. Feature vectors are computed for the first 13 MFCCs which are selected from the computation of the DCT with a constant number of 40 coefficients. The CMRs $\rho_{\nu_1|\nu_2}(\eta)$ for $\nu_1 = 1, \nu_2 = 1$ as well as for $\nu_1 = 2, \nu_2 = 8$ are computed, where $\nu_2 = 8$ is the highest modulation frequency in our analysis. The CMRs and the averaged modulation spectrum $\bar{X}_{\text{mfcc}}(\eta)$ are rewritten in vector notation (for $\eta = 1, \dots, 13$) as $\rho_{1|1}$, $\rho_{2|8}$ and \bar{X}_{mfcc} , and finally stacked into one feature vector

$$\mathbf{v} = (\bar{X}_{\text{mfcc}}^T, \rho_{1|1}^T, \rho_{2|8}^T)^T. \quad (5)$$

This results in 39 coefficients to summarize 4s of audio data. Note, that all vector components make use of the averaged MFCC modulation spectrum (2) to which the equalization gain matrix will be applied.

3. Reduction of reverberation effects in the MFCC modulation spectrum

In this section, we illustrate the effect that reverberation has to the Mel-spectrum $X_{\text{mel}}(k', b)$, the MFCCs $X_{\text{mfcc}}(\eta, b)$, and to the averaged MFCC modulation spectrum (2) of a sample signal. Further, we propose an equalization gain that can be applied in the feature extraction stage to reduce the influence of environmental reverberation on the resulting feature vector.

3.1. Influence of reverberation onto the feature representation

In Fig. 1(a), the Mel-spectrum $X_{\text{mel}}(k', b)$ of a speech signal (i.e., the word *otherwise*), is illustrated. We convolved the sig-

nal with a (measured) room impulse response (RIR) ($T_{60} \approx 600$ ms) to create a reverberant instance of the signal, illustrated in Fig. 1(b). The temporal smearing is clearly observable and the general structure looks more noisy for the reverberant signal.

The respective MFCCs of the anechoic and the reverberant signals are illustrated in Fig. 2. For our feature vector computation, we consider the temporal evolution of the MFCCs. With this in mind, we observe rather slow variations with smooth transitions between high and low values for the clean signal MFCCs (Fig. 2(a)). In comparison, the MFCCs of the reverberant signal look more scattered (Fig. 2(b)). The transitions between high and low values are less smooth and the dynamic range is reduced in general.

Resulting differences are recognizable in the averaged MFCC modulation spectra (2) of both signals (Fig. 3), as well. In the zeroth and first MFCC of the clean signal MFCC modulation spectrum, the modulation energy is higher for nearly all modulation frequencies, compared to the respective reverberant counterpart. The property of slow and smooth variations for $\eta \in \{2, \dots, 9\}$ results in higher modulation amplitudes for these MFCCs in the low modulation frequencies $\nu \in \{2, 3\}$ for the anechoic signal (Fig. 3(a)). The noise-like fluctuations in MFCCs $\eta > 1$ of the reverberant signal (Fig. 3(b)) result in larger amplitudes for high modulation frequencies.

A straightforward processing strategy for reducing the mismatch in the MFCC modulation characteristics between anechoic and reverberant data therefore is to improve the degraded modulation spectrum of the reverberant MFCCs.

3.2. Equalization gain to reduce the influence of reverberation on the averaged modulation spectrum

Given the averaged MFCC modulation spectra of the clean and the degraded signal, as in Fig. 3, we can define an equalization gain matrix \mathbf{G} as

$$\mathbf{G} = \begin{bmatrix} g_{1,1} & \cdots & g_{1,\eta_{\max}} \\ \vdots & g_{\nu,\eta} & \vdots \\ g_{\nu_{\max},1} & \cdots & g_{\nu_{\max},\eta_{\max}} \end{bmatrix} \quad (6)$$

where $\nu_{\max} = L/2$. The values $g_{\nu,\eta}$ can be derived independently for each point in the (ν, η) -plane as the ratio between both averaged modulation spectra. Then \mathbf{G} can be used in a point-wise multiplication to modify the degraded signal MFCC modulation spectrum such that it then resembles the clean modulation spectrum. Of course, \mathbf{G} is in this case the matched solution for one special example of a signal and reverberation scenario.

However, when several audio signals of different types are used in combination with RIRs of different reverberant scenarios, we can estimate an average equalization matrix \mathbf{G} , e.g., by means of a least-squares (LS) problem. To do so, we simulate reverberated versions of signals of different signal types and scenarios, and compare the averaged modulation spectra of these signals with their equivalents for which only the direct path transmission is considered. This comparison is performed for κ simulated reverberated microphone signals. Thus, we formulate a LS problem for each combination of modulation frequency and MFCC as

$$e(g) = \|\tilde{X}_{\text{mfcc,dir}} - g\tilde{X}_{\text{mfcc,rev}}\|^2. \quad (7)$$

For every point in the (ν, η) -plane, we define a vector of the respective values of the averaged modulation spectra of the κ

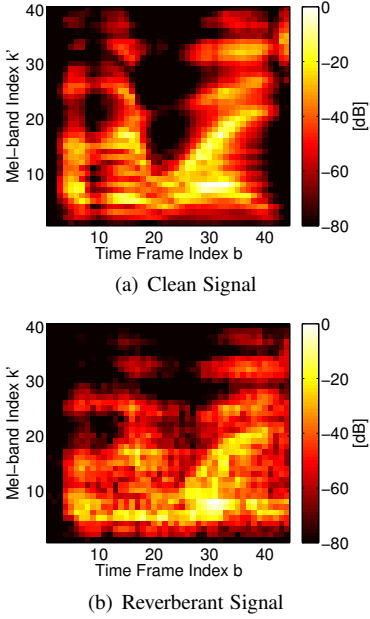


Figure 1: Absolute values of the Mel-spectrum of the word *otherwise*.

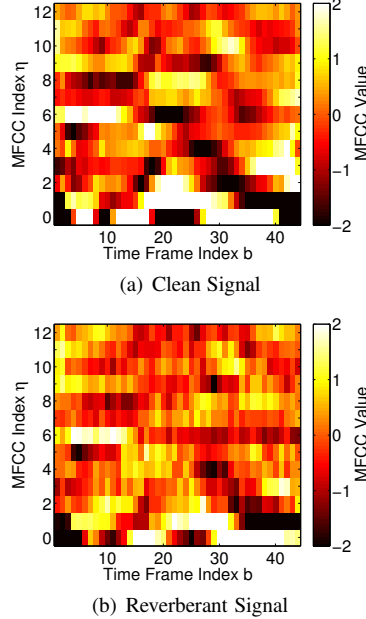


Figure 2: MFCCs of the word *otherwise*.

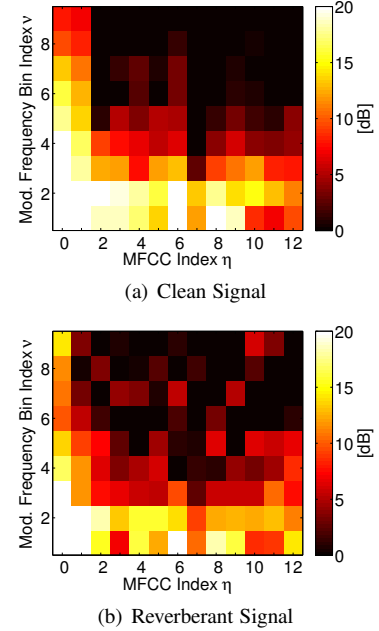


Figure 3: Time averaged MFCC modulation spectrum of the word *otherwise*.

simulated microphone signals for which only direct path transmission from a source to a microphone is considered $\tilde{\mathbf{X}}_{\text{mfcc,dir}} = (\tilde{X}_{\text{mfcc,dir},1}, \tilde{X}_{\text{mfcc,dir},2}, \dots, \tilde{X}_{\text{mfcc,dir},\kappa})^T$, and a vector of values of the averaged modulation spectra of the simulated microphone signals, based on the convolution with complete RIRs $\tilde{\mathbf{X}}_{\text{mfcc,rev}} = (\tilde{X}_{\text{mfcc,rev},1}, \tilde{X}_{\text{mfcc,rev},2}, \dots, \tilde{X}_{\text{mfcc,rev},\kappa})^T$. The dependency on (ν, η) is dropped in the definitions of $\tilde{\mathbf{X}}_{\text{mfcc,dir}}$, $\tilde{\mathbf{X}}_{\text{mfcc,rev}}$ and in (7) to improve readability. By computing the LS-solution

$$g = (\tilde{\mathbf{X}}_{\text{mfcc,rev}}^T \tilde{\mathbf{X}}_{\text{mfcc,rev}})^{-1} \tilde{\mathbf{X}}_{\text{mfcc,rev}}^T \tilde{\mathbf{X}}_{\text{mfcc,dir}} \quad (8)$$

for each point in the (ν, η) -plane, we obtain the weighting matrix \mathbf{G} . Details about the simulations used for this computation and the results are presented in the next section. After the estimations of the elements of an averaged matrix \mathbf{G} , we can apply the gain on the averaged modulation spectrum of the reverberant test signal and obtain an improved averaged modulation spectrum for the feature computation,

$$\tilde{X}_{\text{mfcc,imp}}(\nu, \eta) = \tilde{X}_{\text{mfcc}}(\nu, \eta) g_{\nu, \eta}, \quad (9)$$

which is now (due to the average computation of the gains over several transmission scenarios in different reverberant environments) an approximation of the original averaged modulation spectrum.

3.3. Estimation of Mod-MFCC equalization gain weights

To obtain \mathbf{G} rather independent from a single signal or a specific signal type, we simulate reverberant signals of speech, music and noise sounds. Further, to obtain independence from a specific environment, we create virtual RIRs of three different rooms sizes and reverberation times ($T_{60} = 340, 490, 630\text{ms}$) with randomly placed microphones and sources (details about these rooms are given in [1]) with an improved version of the image source model that we introduced in [19]. In this way, we create $\kappa = 5400$ different instances which contribute to the LS-solution for every point in the (ν, η) -plane. The resulting equalization matrix is depicted in Fig. 4.

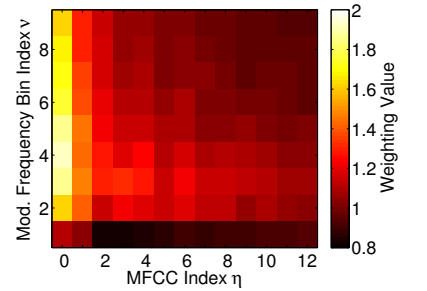


Figure 4: Equalization matrix \mathbf{G} , estimated using simulations of different signal types and varying reverberation scenarios.

The results are in line with the observations based on our simple example for the word *otherwise* before. The modulation energy of the first two MFCCs is increased. Further, we obtain a bandpass-like behavior for larger MFCCs. The amplitudes of the low modulation frequencies $\nu \in \{2, 3\}$ are increased, as the gain is larger than 1. For the zeroth modulation frequency as well as for modulation frequencies $\nu > 3$, the gain is lower than 1, and thus, the amplitudes of the modulation analysis is reduced. Interestingly, the modulation frequencies $\nu \in \{2, 3\}$ correspond to modulations between 4 – 8Hz, which are particularly important for speech [20, 21].

4. Evaluation

We evaluate the influence of the weighting in the modulation spectrum by means of classification experiments with reverberant signals. For this, we create RIRs, again using [19], for the transmission of a source signal to three microphones, for the setup that is illustrated in Fig. 5 (note, that the RIRs of this setup are not used in the estimation of \mathbf{G}). The distance d_i between the source and the microphone i is: $d_1 = 0.4\text{m}$, $d_2 = 2.0\text{m}$, $d_3 = 5.0\text{m}$ and the reverberation time is $T_{60} \approx 600\text{ms}$. Thus, microphone 1 receives the signal with the largest ratio between direct path and reverberant sound energy (DRR). Further,

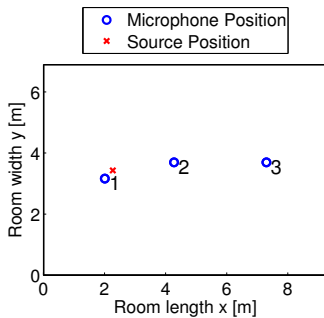


Figure 5: Simulation setup for the classification of microphone signals.

we create two additional microphone signals using two measured RIRs from an office (Mic. 4, $d_4 = 3\text{m}$, $T_{60} \approx 450\text{ms}$) and a lecture hall (Mic. 5, $d_5 = 2.2\text{m}$, $T_{60} \approx 700\text{ms}$) [22]. Note, that for the estimation of \mathbf{G} , the RIRs from [22] are not used. Thus, different RIRs are used for the gain estimation and the testing with artificial and measured RIRs.

We use 100 signals of each class, speech [23], music (private database, different genres such as rock and pop, instrumental classic and jazz, labelled using the *allmusic* guidelines [24]) and noise [25] and extract the Mod-MFCC features using the parameters described earlier. For the classification we use a linear discriminant analysis (LDA) [26] which is trained with 75% of the data of each class. For the baseline experiment, we use the remaining 25% of the data of each class as clean test samples for the classification experiment (the gain in the modulation spectrum is not applied in this case). For the classification based on reverberant signals, the remaining 25% are used as source signal and convolved with the RIRs, to create the simulated microphone signals from which the audio features are extracted. The classification accuracy is averaged over 10 cross-validation iterations in which the allocation of training and test data is randomized and thus, new reverberant microphone signals are created in each iteration. For the classification experiments based on the simulated microphone signals, the feature vectors are computed for both cases, when \mathbf{G} is not applied (*original* modulation spectrum) and when \mathbf{G} is applied (*improved* modulation spectrum).

5. Results

The results of the baseline experiment in Tab. 1 show that the selected feature set in combination with the LDA classifier is capable to discriminate between the three signal classes with an accuracy of over 90% when no mismatch between training and test data is present. The classes speech and music are classified correctly in more than 96% of the test cases.

When feature vectors from the reverberant microphone signals are used as input for the classifier and the gain in the modulation spectrum is *not* applied, only the signals of microphone 1 (which is located very closely to the source) are classified with a similar accuracy (Tab. 2) for all classes. For the other microphones, the classification accuracy of speech suffers most from the influence of reverberation. The classification accuracy of noise is reduced as well. The modifications due to reverberation in the data lead to a misclassification of both classes as music. When the proposed equalization gain in the averaged modulation spectrum is applied for the feature computation of the microphone signals, the classification is much more accurate for the speech classification. The accuracy for music classification does not change notably. Noise is classified slightly

Table 1: Classification accuracy (in %), for clean data.

	classified as		
	Speech	Music	Noise
Speech	99.5	0.5	0.0
Music	1.7	96.2	2.1
Noise	0.0	8.2	91.8

Table 2: Accuracy of the classification (in %) based on simulated microphone signals using the original Mod-MFCCs.

Source	Classified as	Sim. RIRs			Meas. RIRs	
		Mic.-Index			4	5
		1	2	3		
Speech	Speech	97.6	48.4	40.0	76.4	56.4
	Music	2.4	51.6	60.0	23.6	43.6
	Noise	0.0	0.0	0.0	0.0	0.0
Music	Speech	1.4	1.2	1.8	1.6	0.0
	Music	98.6	98.8	97.8	97.6	99.6
	Noise	0.0	0.8	0.4	0.8	0.4
Noise	Speech	0.0	0.0	0.0	0.0	0.0
	Music	11.2	13.2	17.2	17.3	14.7
	Noise	88.8	86.8	82.8	82.7	85.3

Table 3: Accuracy of the classification (in %) based on simulated microphone signals. The modulation spectrum improvement is applied in the feature extraction.

Source	Classified as	Sim. RIRs			Meas. RIRs	
		Mic.-Index			4	5
		1	2	3		
Speech	Speech	99.8	98.8	96.8	99.6	98.0
	Music	0.2	1.6	3.2	0.4	2.0
	Noise	0.0	0.0	0.0	0.0	0.0
Music	Speech	2.4	2.0	3.0	3.2	3.2
	Music	97.2	97.2	96.6	96.4	95.6
	Noise	0.4	0.8	0.4	0.4	1.2
Noise	Speech	0.0	0.0	0.0	0.0	0.0
	Music	7.6	14.4	15.6	15.3	13.3
	Noise	92.4	85.6	84.4	84.7	86.7

more accurate, mainly for microphone 1 which is very close to the source. The accuracy improvement is observable for both, simulated and measured RIRs.

6. Conclusions

In this paper we proposed a multiplicative gain function in the MFCC modulation spectral domain which minimizes the mismatch between anechoic training data and reverberant test data in the least-square sense. The gain can be applied to obtain an enhanced feature representation as input for a classification system which is trained on clean and anechoic data. Several microphone-source scenarios in different reverberant rooms are used for the training of the gains. Tests are conducted on simulated microphone signals which are created with different artificial RIRs as well as with measured RIRs. Due to the application of the gain, the classification accuracy particularly for speech classification can be increased in the presence of reverberation. The improvement, however, may be less strong if the reverberation time in the test case diverges drastically from those used in the gain estimation. The proposed method can easily be adapted to other forms of modulation-based features.

7. References

- [1] S. Gergen, A. Nagathil, and R. Martin, "Classification of reverberant audio signals using clustered ad hoc distributed microphones," *Signal Processing*, vol. 107, pp. 21–32, 2015.
- [2] D. Kolossa and R. Haeb-Umbach, *Robust Speech Recognition of Uncertain or Missing Data*. Springer, 2011.
- [3] P. A. Naylor and N. D. Gaubitch, *Speech dereverberation*. Springer Science & Business Media, 2010.
- [4] D. Schmid, G. Enzner, S. Malik, D. Kolossa, and R. Martin, "Variational Bayesian inference for multichannel dereverberation and noise reduction," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 22, no. 8, pp. 1320–1335, 2014.
- [5] M. McKinney and J. Breebaart, "Features for audio and music classification," in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, 2003.
- [6] R. Martin and A. Nagathil, "Cepstral modulation ratio regression (CMRARE) parameters for audio signal analysis and classification," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2009, pp. 321–324.
- [7] J. Bach, B. Kollmeier, and J. Anemüller, "Modulation-based detection of speech in real background noise: Generalization to novel background classes," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2010, pp. 41–44.
- [8] H. G. Hirsch, "Die Enthaltung von Sprache zur Verbesserung einer automatischen Spracherkennung in Räumen," *Acta Acustica united with Acustica*, vol. 67, no. 3, pp. 216–221, Jan. 1989.
- [9] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn, "Rasta-plp speech analysis technique," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, 1992, pp. 121–124.
- [10] M. Buehler, S. Allegro, S. Launer, and N. Dillier, "Sound classification in hearing aids inspired by auditory scene analysis," *EURASIP Journal on Advances in Signal Processing*, vol. 18, pp. 2991–3002, 2005.
- [11] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1997, pp. 1331–1334.
- [12] A. Nagathil, P. Götzel, and R. Martin, "Hierarchical audio classification using cepstral modulation ratio regressions based on legendre polynomials," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011, pp. 2216–2219.
- [13] A. Meng, P. Ahrendt, J. Larsen, and L. Hansen, "Temporal feature integration for music genre classification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1654–1664, 2007.
- [14] A. Holzapfel and Y. Stylianou, "Musical genre classification using nonnegative matrix factorization-based features," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 424–434, 2008.
- [15] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, p. 357, 1980.
- [16] S. Furui, *Digital Speech Processing, Synthesis, and Recognition*. Marcel Dekker, Inc., 2000.
- [17] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. O.-G. D., Petrovskadelacrétaz, and D. A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP Journal on Applied Signal Processing*, vol. 2004, pp. 430–451, Jan 2004.
- [18] P. N. Garner, "Cepstral normalisation and the signal to noise ratio spectrum in automatic speech recognition," *Speech Communication*, vol. 53, no. 8, pp. 991–1001, October 2011.
- [19] S. Gergen, C. Borß, N. Madhu, and R. Martin, "An optimized parametric model for the simulation of reverberant microphone signals," in *Proceedings of the International Conference on Signal Processing, Communications and Computing (ICSPCC)*, 2012.
- [20] R. Drullman, J. M. Festen, and R. Plomp, "Effect of reducing slow temporal modulations on speech reception," *The Journal of the Acoustical Society of America*, vol. 95, no. 5, pp. 2670–2680, 1994.
- [21] N. Kanedera, H. Hermansky, and T. Arai, "On properties of modulation spectrum for robust automatic speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2. IEEE, 1998, pp. 613–616.
- [22] M. Jeub, M. Schäfer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Proceedings of the 16th International Conference on Digital Signal Processing*, 2009, pp. 1–5.
- [23] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "Timit acoustic-phonetic continuous speech corpus," 1993, Linguistic Data Consortium, Philadelphia.
- [24] "Allmusic," <http://www.allmusic.com/>.
- [25] B. Nimens, "The general series 6000 sound effects library," <http://www.sound-ideas.com/6000.htm>.
- [26] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York: Wiley, 2001.