OPTIMIZATION OF AMPLITUDE MODULATION FEATURES FOR LOW-RESOURCE ACOUSTIC SCENE CLASSIFICATION

Semih Ağcaer, Anton Schlesinger, Falk-Martin Hoffmann, and Rainer Martin

Institute of Communication Acoustics, Ruhr-Universität Bochum, Germany

ABSTRACT

We developed a new feature extraction algorithm based on the Amplitude Modulation Spectrum (AMS), which mainly consists of two filter bank stages composed of low-order recursive filters. The passband range of each filter was optimized by using the Covariance Matrix Adaptation - Evolution Strategy (CMA-ES). The classification task was accomplished by a Linear Discriminant Analysis (LDA) classifier. To evaluate the performance of the proposed acoustic scene classifier based on AMS features, we tested it with the publicly available dataset provided by the IEEE AASP Challenge 2013. Using only 9 optimized AMS features, we achieved $85\,\%$ classification accuracy, outperforming the best previously available approaches by $10\,\%$.

***Index Terms*—** evolutionary optimization, acoustic scene classification, acoustic feature extraction, amplitude modulation spectrum

1. INTRODUCTION

The goal of an acoustic scene classification system is to automatically assign sound signals to certain acoustic classes. Such a system can, for example, be used in mobile phones [1–3] to automatically match the acoustic notification profile to the currently detected acoustic environment. For instance, during a meeting the acoustic notification profile could switch into silent mode. However, an acoustic scene classifier is not only beneficial to mobile phones. Another application could be in the field of mobile robotics, using acoustic information in addition to vision. Also hearing aids [4–6] can benefit from acoustic classifiers, e.g. to automatically adjust the signal processing profile to the current acoustic scene. For example, in a noisy environment the noise reduction can be activated or, if speech is detected, the beamformer could be switched on. In order to make such an approach applicable for mobile battery-based low-power devices, the computational complexity of the classification system must be minimized.

Properly chosen features are essential for a reliable and accurate classification system. Many different audio features have been developed that are computed either in the time or the frequency domain [7]. Often, useful features are extracted by feature selection techniques [8], resulting in a heterogeneous collection of audio features and thus a more complicated implementation. More recently, AMS based features, which are inspired by the human auditory signal processing, gained some prominence [9–12]. However, using a large number of AMS features as suggested in [10] is not feasible in computationally limited scenarios.

In this paper, we propose a computationally efficient approach to acoustic signal classification that makes use of amplitude modulation spectrum features and an LDA classifier. As we strive for a low number of features, we optimize our features using the Covariance Matrix Adaptation Evolutionary Strategy (CMA-ES) [13]. Therefore, we extract a small and homogeneous set of features, which is well suited for hardware or software implementation with low resources and can be easily adapted to different classification tasks.

In the next section of this paper, we introduce the AMS feature extraction scheme and the CMA-ES optimization method. In section 3, we test our acoustic scene classifier with the IEEE AASP Challenge 2013 public dataset and present our results. Section 4 concludes this paper with a discussion of these results.

2. METHODS

2.1. Amplitude Modulation Spectrum Features

Choosing an adequate feature extraction procedure for a classification system is crucial and challenging. Since previous research has revealed the importance of amplitude modulations for recognition and processing of audio signals in human auditory perception, our features are extracted on the basis of AMS [11]. Anemüller et al. built a classifier based on AMS features with a recognition rate of up to 90% for speech in real acoustic backgrounds of non-speech sounds [9]. Moritz et al. [12] and Bach et al. [10] also used AMS features for speech detection in noisy environments and achieved similar results. The proposed algorithm in [9] employs two FFT stages and extracts 493 features. This could however be demanding for battery-operated low-power digital signal processors. Therefore, we introduce a less complex AMS algorithm, which utilizes parametric filter banks instead of FFTs and extracts only a small number of features. The parameters of these filter banks are optimized using an evolutionary opti-

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

mization method.

Figure 1a illustrates the structure of our feature extraction procedure. It consists of two successive filter bank stages and a final averaging stage. In the following, we describe the processing of an input signal $x(k)$ in more detail. First, the input signal $x(k)$ with length N passes through the time domain filter bank, which contains N_{TF} parallel bandpass filters. The first filter in this filter bank is a lowpass filter with the cutoff frequency f_{u1}. The cutoff frequency f_{u1} of the first filter is the lower edge frequency f_{l2} of the second bandpass filter and so on. The upper edge frequency of the last filter is set to 10000 Hz. The output $\mathbf{Y_{TF}}$ of this stage is an $N_{TF} \times N_R$ matrix with $N_R = \lfloor \frac{N}{R} \rfloor$, where R is a decimation factor and $\lfloor \cdot \rfloor$ is the floor operation. The inner structure of the time domain filter bank stage is depicted in Figure 1b. First, a highpass filter with a cutoff frequency of 25 Hz removes any DC component from the input signal. The highpass filtered signal is fed into N_{TF} parallel bandpass filters, which leads to N_{TF} subband signals $x_{T,i}(k)$ with different spectral content. In the next steps the subband signals are rectified, lowpass filtered and decimated. The lowpass filter is a fifth-order recursive Chebyshev II filter with 30 dB attenuation in the stopband. The cutoff frequency $f_{T,s}$ is determined by the highest bandpass filter upper edge frequency of the modulation filter bank plus an additional 40 Hz. The lowpass filter prior to the decimation avoids aliasing effects. The decimation factor is given by $R = \lfloor \frac{f_s}{2 \cdot f_{T,s}} \rfloor$, where f_s is the sampling frequency. The decimated subband signals $x_{R,i}(n)$ are then raised to the power of ϵ_1. The following logarithmic compression block is optional and can be selected by $I_\gamma \in \{0, 1\}$. The last block of the time domain filter bank stage is a simple smoothing filter with the smoothing parameter α_1. The last three blocks of this stage can be summarized by

$$y_{R,i}(n) = \alpha_1 y_{R,i}(n-1) + (1-\alpha_1)\gamma\left((x_{R,i}(n))^{\epsilon_1}\right) \quad (1)$$

with

$$\gamma(\cdot) = \begin{cases} \log_{10}(\cdot) & \text{if } I_\gamma = 1 \\ (\cdot) & \text{if } I_\gamma = 0\,. \end{cases} \quad (2)$$

The final output signals $\mathbf{y}_{R,i} = [y_{R,i}(1)\; y_{R,i}(2)\; \dots\; y_{R,i}(N_R)]$ are arranged into a matrix $\mathbf{Y}_{TF} = \left[\mathbf{y}_{R,1}^T\; \mathbf{y}_{R,2}^T\; \dots\; \mathbf{y}_{R,N_{TF}}^T\right]^T$, where each row corresponds to a certain frequency band. The total number of tunable parameters in the time domain filter bank stage equals $N_{TF} - 1 + 3$.

Figure 1c shows the inner structure of the modulation filterbank stage. The output $\mathbf{Y}_{TF}$ of the time domain filter bank stage (where each row corresponds to a certain frequency band) is first fed row by row to N_M parallel modulation filters. The modulation filters can be parameterized for each frequency band individually, yielding an overall number of $N_M \times N_{TF}$ filters. The ordering of the parallel bandpass filters for each frequency is analog to the parallel bandpass filter in the time domain stage as shown in Figure 2, except the upper edge frequency of the last filter is not predefined. The absolute values of the output signals $\hat{x}_{M,i}(n)$ with $i \in \{1, \dots N_{TF}, \times N_M\}$ of the modulation filter bank yields the corresponding envelope. The last three blocks of the modulation filter bank stage are similar to those of the time domain filter stage. Accordingly, the processing after the modulation filters can be summarized by

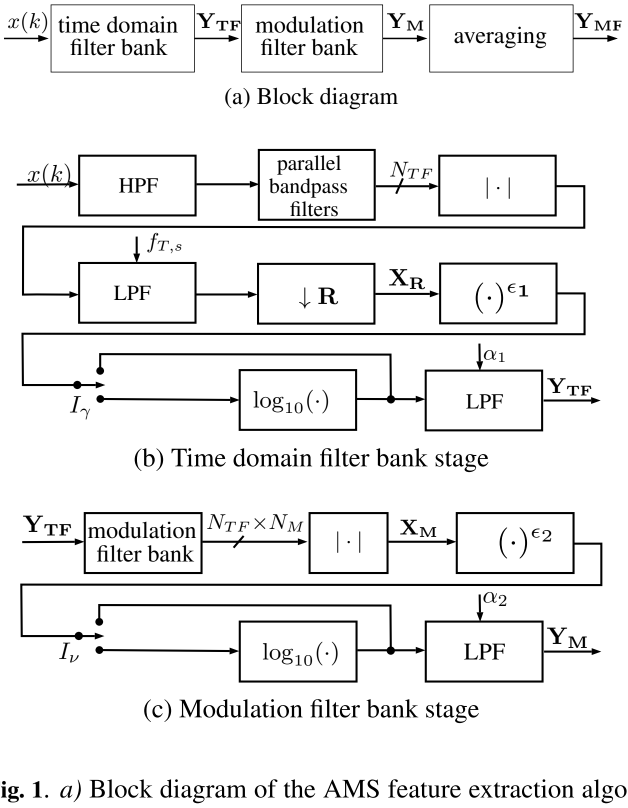
$$y_{M,i}(n) = \alpha_2 y_{M,i}(n-1) + (1-\alpha_2)\nu\left(|\hat{x}_{M,i}(n)|^{\epsilon_2}\right), \quad (3)$$

where

$$\nu(\cdot) = \begin{cases} \log_{10}(\cdot) & \text{if } I_\nu = 1 \\ (\cdot) & \text{if } I_\nu = 0\,. \end{cases} \quad (4)$$

Again, the logarithmic compression is optional and can be selected by the parameter $I_\nu \in \{0, 1\}$. The output of this stage is a $(N_{TF} \times N_M) \times N_R$ matrix $\mathbf{Y_M}$. In a final step, the amplitude modulation spectrum is obtained by averaging over N_R signal samples. The result is a feature vector Y_{MF} with dimensions $N_{TF} \times N_M$.

Therefore, the total number of adjustable parameters equals to $N_{TF} - 1 + 3 + (N_{TF} \times N_M) + 3 = N_{TF} \cdot (N_M + 1) + 5$. The CMA-ES tunes these parameters and determines the most proper filter structure, e.g. if expansions and/or compressions should be applied.



(a) Block diagram

(b) Time domain filter bank stage

(c) Modulation filter bank stage

Fig. 1. *a)* Block diagram of the AMS feature extraction algorithm *b)* the inner structure of time domain filter bank stage and *c)* the inner structure of the modulation filter bank

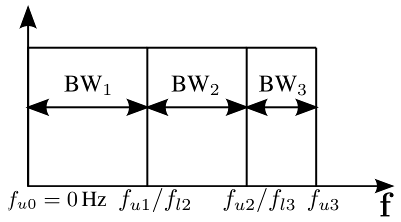


Fig. 2. Filter band structure for three filters. f_{li} is the lower edge frequency of the i-th filter, f_{ui} is the upper edge frequency of the i-th filter, and BW_i is the bandwidth of the i-th filter

2.2. Classification

We choose a Linear Discriminant Analysis (LDA) classifier as our classification method. LDA assumes that each class density is a multivariate Gaussian and the classes have a common covariance matrix $\hat{\Sigma}$ [14]. For a two class classification task, with classes 1 and 2, the decision rule for class 2 is [14]

$$x^T\hat{\Sigma}^{-1}(\hat{\mu}_2 - \hat{\mu}_1) > \frac{1}{2}(\hat{\mu}_2 + \hat{\mu}_1)^T\hat{\Sigma}^{-1}(\hat{\mu}_2 - \hat{\mu}_1) - \log\frac{N_2}{N_1} \tag{5}$$

where x is the feature vector, $\hat{\Sigma}$ the covariance matrix, $\hat{\mu}_i$ the mean of class i and N_i the number of class-i members in the training set. For a multi-class classification task, the one-vs-one approach was employed. The LDA classifier was trained and tested with a k-fold cross-validation method, with $k = 5$.

2.3. CMA Evolution Strategy

As mentioned previously our proposed AMS algorithm has $N_{TF} \cdot (N_M + 1) + 5$ independently tunable parameters. For $N_{TF} = 3$ time domain filter and $N_M = 3$ modulation filters, this leads to 17 parameters, which have to be chosen properly to minimize the classification error. Finding optimal parameters using an exhaustive search would not be feasible due to the high dimensionality. A gradient descent algorithm would also not be suitable because our multimodal cost function (classification error) is not differentiable. Thus we chose a Covariance Matrix Adaptation Evolution Strategy (CMA-ES) [13] based method in order to find an ideal parameter set for our feature extraction step. ES is a subclass of evolutionary algorithms (EA) and shares the basic idea to imitate evolution, for instance by mutation and selection, and it does not require the computation of any derivatives [15]. The optimal parameter set is approximated iteratively by evaluating a fitness function after each step. Here, the fitness function or cost function is the classification error (the ratio of the number of misclassified objects to the number of all objects) of the LDA classifier as a function of the independently tunable parameters.

The basic equation for CMA-ES is the sampling equation of new search points [13]

$$\mathbf{x}_k^{g+1} \sim \mathbf{m}^{(g)} + \sigma^{(g)}\mathcal{N}\left(\mathbf{0}, \mathbf{C}^{(g)}\right) \qquad \text{for } k = 1, \ldots, \lambda \tag{6}$$

where g is the index of the current generation (iteration), $\mathbf{x}_k^{g+1}$ k-th offspring from generation $g+1$, λ the number of offspring, $\mathbf{m}^{(g)}$ mean value of the search distribution at generation g, $\mathcal{N}\left(\mathbf{0}, \mathbf{C}^{(g)}\right)$ a multivariate normal distribution with the covariance matrix $\mathbf{C}^{(g)}$ of generation g and $\sigma^{(g)}$ the step-size of generation g. From the λ sampled new solution candidates, the μ best points (in terms of minimal cost function) are selected and the new mean of generation $g+1$ is determined by a weighted average according to

$$\mathbf{m}^{(g+1)} = \sum_{i=1}^{\lambda} w_i \mathbf{x}_{i:\lambda}^{g+1}, \tag{7}$$

$$\sum_{i=1}^{\lambda} w_i = 1, \qquad w_1 \geq w_2 \geq \cdots \geq w_\mu > 0. \tag{8}$$

In each iteration of the CMA-ES, the covariance matrix $\mathbf{C}$ and the step-size σ are adapted according to the success of the sampled offspring. The shape of the multivariate normal distribution is formed in the direction of the old mean $\mathbf{m}^{(g)}$ towards the new mean $\mathbf{m}^{(g+1)}$. The sampling, selection and recombination steps are repeated until either a predefined threshold on the cost function or a maximum number of generations is reached. We constricted the allowed search space of the parameters to intervals as described by Colutto et al. in [16]. For a more detailed description of CMA-ES, in particular on how the covariance matrix $\mathbf{C}$ and the step-size σ is adapted in each step, as well as a Matlab implementation, please refer to [13].

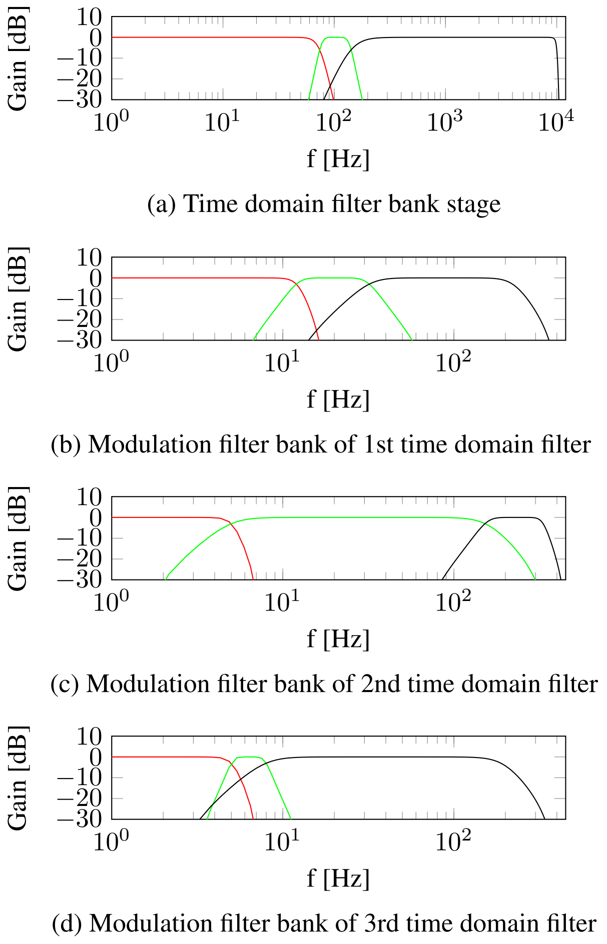
3. RESULTS

3.1. Data

We used the public dataset for scene classification provided in the context of the IEEE AASP Challenge 2013 [17]. The dataset consists of 10 acoustic scenes: busy street, quiet street, supermarket/store, restaurant, office, park, bus, tube/metro, tube station and open market. For each scene there exist 10 recordings of 30 seconds each. The original stereo recordings are decimated from 44100 Hz to 22050 Hz and only one channel was used. The resolution of 16 bit remained unchanged.

3.2. AMS Parameters Obtained from CMA-ES

Based on preliminary results, the number of sixth-order time-domain filters N_{TF} was set to 3 and the number of sixth-order modulation filters N_M for each time-domain filter was also set to 3. Thus, we obtained 9 features from the AMS feature extraction stage and had 17 tunable parameters to set.



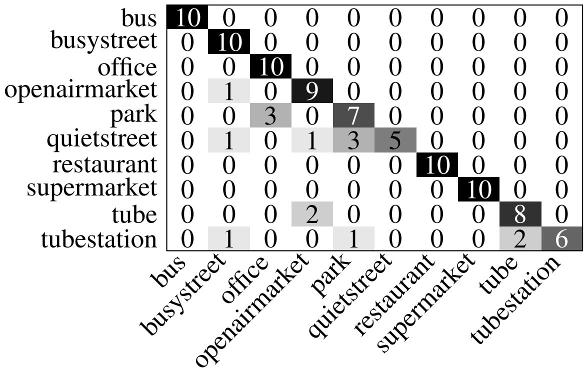
(a) Time domain filter bank stage

(b) Modulation filter bank of 1st time domain filter

(c) Modulation filter bank of 2nd time domain filter

(d) Modulation filter bank of 3rd time domain filter

Fig. 3. Frequency responses of the filters found by the CMA-ES Optimization



	bus	busystreet	office	openairmarket	park	quietstreet	restaurant	supermarket	tube	tubestation
bus	10	0	0	0	0	0	0	0	0	0
busystreet	0	10	0	0	0	0	0	0	0	0
office	0	0	10	0	0	0	0	0	0	0
openairmarket	0	1	0	9	0	0	0	0	0	0
park	0	0	3	0	7	0	0	0	0	0
quietstreet	0	1	0	1	3	5	0	0	0	0
restaurant	0	0	0	0	0	0	10	0	0	0
supermarket	0	0	0	0	0	0	0	10	0	0
tube	0	0	0	2	0	0	0	0	8	0
tubestation	0	1	0	0	1	0	0	0	2	6

Fig. 4. Confusion matrix for the AASP Challenge Database [17]

Parameter	Value
Filter order time domain filters	6
Filter order modulation filters	6
Passband 1. time domain filter	0 - 74 Hz
Passband 2. time domain filter	74 - 142 Hz
Passband 3. time domain filter	142 - 10000 Hz
Passband 1. modulation filter of 1. time domain filter	0 - 12 Hz
Passband 2. modulation filter of 1. time domain filter	12 - 32 Hz
Passband 3. modulation filter of 1. time domain filter	32 - 220 Hz
Passband 1. modulation filter of 2. time domain filter	0 - 5 Hz
Passband 2. modulation filter of 2. time domain filter	5 - 152 Hz
Passband 3. modulation filter of 2. time domainD filter	152 - 332 Hz
Passband 1. modulation filter of 3. time domain filter	0 - 5 Hz
Passband 2. modulation filter of 3. time domain filter	5 - 8 Hz
Passband 3. modulation filter of 3. time domain filter	8 - 188 Hz
α_1	0.806
α_2	0.707
ϵ_1	1.184
ϵ_2	0.317
I_γ	1
I_ν	0

Table 1. Parameter set found by the CMA-ES algorithm

These 17 features were found by the CMA-ES optimization, which was limited to 60 generations (22 offspring were sampled in each iteration) and took 3 hours on a PC with Intel(R) Core(TM) i5-3470 CPU @ 3.20GHz and Matlab 2012 64-Bit. The filter parameters found by the CMA-ES optimization are shown in Table 1 and the frequency responses of the corresponding filters are depicted in Figure 3.

3.3. Classification

Prior to the feature extraction step, the 30 s long audio files are divided into 6 frames, each being 5 s long, and the extracted features for each frame are fed into the classifier. Thus, we got 6 classification results for each audio file. The classification result for the whole audio file was determined by a majority vote. The achieved classification accuracy, which is defined as the ratio of the number of correctly classified scenes to the total number of scenes, is 85 % with a standard deviation of 3.54 %. The corresponding confusion matrix of the 5-fold cross-validation is depicted in Figure 4. Most misclassification occurred in the class *quietstreet*, which is confused 3 times with the class *park*.

4. DISCUSSION AND CONCLUSION

In this paper, we proposed an acoustic scene classification system based on a new AMS feature extraction algorithm, which is inspired by the signal processing in the human auditory cortex. It is computationally significantly less complex than other known AMS-based feature extraction algorithms [9–12]. The number of extracted features is determined by the number of filters used in each step. As a conventional gradient descent method (or other comparable deterministic optimization methods) was not considered suitable to find the ideal passband ranges for each filter, we used the CMA-ES optimization method to solve this problem.

In order to compare our AMS-based acoustic scene classifier system with other proposed methods, we evaluated it on the basis of the publicly available dataset for scene classification provided by the IEEE AASP Challenge 2013. With 85 % classification accuracy, our classifier significantly outperforms the best two algorithms submitted to the challenge (75 % [18] and 75 % [19]), with only 9 features and less complex methods. Furthermore, we use a uniform feature extraction scheme and not a collection of entirely different features, which facilitates the implementation of the method. Obviously, the parameters of these features depend on the training data, and with different training samples and different acoustic scenes we will obtain different parameters. However, the flexible yet regular structure of the feature extraction process allows an easy adaptation to other classification tasks.

REFERENCES

[1] In-Cheol Kim, Joo-Hee Kim, and Seok-Jun Lee, "MobileSense: A robust sound classification system for mobile applications," in *2014 International Conference on Systems, Signals and Image Processing (IWSSIP)*, May 2014, pp. 147–150.

[2] M. Rossi, S. Feese, O. Amft, N. Braune, S. Martis, and G. Troster, "AmbientSense: A real-time ambient sound recognition system for smartphones," in *2013 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*, March 2013, pp. 230–235.

[3] A.J. Eronen, V.T. Peltonen, J.T. Tuomi, A.P. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi, "Audio-based context recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 321–329, Jan 2006.

[4] M. Büchler, 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. 2005, no. 18, pp. 2991 – 3002, 2005.

[5] S. Ravindran and D.V. Anderson, "Audio classification and scene recognition for hearing aids," in *IEEE International Symposium on Circuits and Systems, 2005. ISCAS 2005.*, May 2005, pp. 860–863 Vol. 2.

[6] P. Nordqvist and A. Leijon, "An efficient robust sound classification algorithm for hearing aids," in *Journal of Acoustical Society of America*, June 2004, pp. 3033–3033 Vol. 115.

[7] O. Lartillot, P. Toiviainen, and T. Eerola, "Matlab toolbox for musical feature extraction from audio," *International Conference on Digital Audio Effectes*, 2007.

[8] E. Alexandre, L. Cuadra, M. Rosa, and F. Lopez-Ferreras, "Feature selection for sound classification in hearing aids through restricted search driven by genetic algorithms," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2249–2256, Nov 2007.

[9] J. Anemüller, D. Schmidt, and J.H. Bach, "Detection of speech embedded in real acoustic background based on amplitude modulation spectrogram features," in *Interspeech 2008 9th Annual Conference of the International Speech Communication Association*. Interspeech, 2008, pp. 2582–2585.

[10] J. Bach, J. Anemüller, and B. Kollmeier, "Robust speech detection in real acoustic backgrounds with perceptually motivated features," *Speech Communication*, vol. 53, no. 5, pp. 690 – 706, 2011, Perceptual and Statistical Audition.

[11] G. Langner, M. Sams, P. Heil, and H. Schulze, "Frequency and periodicity are represented in orthogonal maps in the human auditory cortex: evidence from magnetoencephalography.," *J Comp Physiol A*, vol. 181, no. 6, pp. 665 – 676, 1997.

[12] N. Moritz, J. Anemüller, and B. Kollmeier, "Amplitude modulation spectrogram based features for robust speech recognition in noisy and reverberant environments," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2011, pp. 5492–5495.

[13] N. Hansen, "The CMA evolution strategy: A comparing review," in *Towards a new evolutionary computation. Advances in estimation of distribution algorithms*. Springer, 2006, pp. 75–102.

[14] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer, 2009.

[15] H. Beyer, *Theory of Evolution Strategies*, Springer, 2001 edition.

[16] S. Colutto, F. Frühauf, M. Fuchs, and O. Scherzer, "The CMA-ES on Riemannian manifolds to reconstruct shapes in 3-D voxel images," *IEEE Transactions on Evolutionary Computation*, vol. 14, no. 2, pp. 227–245, April 2010.

[17] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange, and M.D. Plumbley, "Detection and classification of acoustic scenes and events: An IEEE AASP challenge," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct 2013, pp. 1–4.

[18] Kyogu Lee, Ziwon Hyung, and Juhan Nam, "Acoustic scene classification using sparse feature learning and event-based pooling," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct 2013, pp. 1–4.

[19] A. Rakotomamonjyd and G. Gasso, "Histogram of gradients of time frequency representations for audio scene classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 142–153, Jan 2015.