REVERBERATION-BASED FEATURE EXTRACTION FOR ACOUSTIC SCENE CLASSIFICATION

Miloš Marković, Jürgen Geiger

Huawei Technologies Düsseldorf GmbH, European Research Center, Munich, Germany

ABSTRACT

We present a system for acoustic scene classification, which is the task to classify an environment based on audio recordings. First, we describe a strong low-complexity baseline system using a compact feature set. Second, this system is improved with a novel class of audio features, which exploit the knowledge of sound behaviour within the scene – reverberation. This information is complementary to commonly used features for acoustic scene classification, such as spectral or cepstral components. For extracting the new features, temporal peaks in the audio signal are detected, and the decay after the peak reveals information about the reverberation properties. For the detected decays, statistics are extracted and summarized over time and over frequency bands. The combination of the novel features with features used in state-of-the-art algorithms for acoustic scene classification increases the classification accuracy, as our results obtained with a large in-house database and the DCASE 2016 database demonstrate.

Index Terms— Acoustic scene classification, feature extraction, reverberation

1. INTRODUCTION

Acoustic scene classification (ASC) is the technology which aims at recognising the type of an environment where the user is located only from the sound recorded at that place - the sound events occurring at the specific environment and/or the sounds that environments produce themselves. It is one of the tasks in the field of computational auditory scene analysis (CASA) [1, 2]. Over the last years, a lot of progress has been made. This was mainly fostered by the public DCASE challenges organised in 2013 and 2016 [3, 4]. The progress in the field is synchronised to the field of acoustic event detection [5], as the two tasks are closely related, and similar technologies are used. It was already shown how ASC technology could be integrated into real products, such as smartphones [6, 7].

Generally, the ASC process is divided into two phases: training and classification. The model training phase involves estimation of scene models in terms of suitable classifier (SVM, GMM, neural networks…). It is done by extracting audio features from each instance of the audio recording database, and by training the system with the known samples of all classes. The classification phase requires scene models obtained in the training phase and it involves extraction of the same features from an unknown audio sample. Based on these two inputs, the unknown audio sample is classified into the best matching class [8].

An important part of ASC is to define and extract properties that characterize a certain environment – audio features. Previous work on acoustic scene classification investigated the application of various spectral, energy and voicing-related features [9]. The most commonly used categories of features are cepstral [10], image processing [11], voicing [10] and spatial features [12]. A class of spectro-temporal audio features that was originally proposed for robust speech recognition [13] has been successfully used for acoustic event detection in [14]. Most of the previously proposed audio features for ASC are based on properties of the specific acoustic events occurring in the scene, or on the relation and dynamics of the events. The actual acoustic properties of the environment, such as the type and amount of reverberation have mostly been neglected so far.

In this paper, we want to investigate how the acoustic properties of an environment, in terms of reverberation, can be exploited for acoustic scene classification. We present a new category of features which is inspired by an approach to blind reverberation time (RT) estimation [15, 16]. The features are extracted by analyzing an audio signal in terms of sub-band energy decay rate [17] and by applying basic statistics for the decay rate distribution in time and over frequency bands. The proposed feature set is referred to as decay rate distribution (DRD) features within this paper.

The details of the algorithm for reverberation-based feature extraction are given in Section 2. In Section 3, an ASC system based on Support Vector Machine (SVM) classifier [18] and the new feature category is described. The results of the mentioned ASC system are compared with the state-of-the-art ASC solutions and presented in section 4. Finally, in Section 5 the main conclusions on the presented work are given.

2. REVERBERATION-BASED FEATURES

We define a new category of audio features for ASC which is based on reverberation properties of enclosures or open spaces. Conventional features (MFCC, spectral) model the occurring sounds and acoustic events within the scene while the novel proposed feature category captures properties of the acoustic environment itself. A graphical overview of the algorithm is given in Figure 1. The steps applied on an audio recording in order to obtain a feature vector are grouped in three main parts: transformation to frequency domain, decay rate calculation and decay rate distribution. In order to capture the reverberation properties of an acoustic scene, an automatic method is employed. Temporal peaks are detected, and the energy decay after the peaks is assumed to represent a reverberation tail. Collecting statistics over a number of peaks and corresponding decay rates leads to a reverberation signature.

The research leading to these results has received funding from the European Commission Union Seventh Framework Programme (FP7/2007/2013) under grant agreement 607480 LASIE.

2.1. Transformation to a suitable frequency domain

Assuming that the input audio signal is given in a time domain (waveform), the first step is to make a suitable transformation into a frequency domain. The transformation is done using the short-Time Fourier transform (STFT). The logarithm of the magnitude of the resulting spectrum is calculated in order to obtain log-magnitude spectrum representation of the audio signal. Furthermore, a broadband spectrum is transformed to a perceptual scale by applying Mel-filterbank. The result is a log-magnitude spectrum in a number of frequency bands, with the number of bands N_b as defined by the Mel-filterbank.

2.2. Decay rate calculation

In each of the frequency bands, the log-magnitude spectrum is analyzed in terms of temporal peaks, where any standard well-known algorithm could be used. Peaks are detected according to a pre-defined threshold value which represents the difference between the magnitude of the sample of interest and the neighbouring local maxima. Sweeping over the whole length of the signal, peaks that fulfil the threshold criterion are obtained. A slope of each detected peak is calculated by applying the linear least square fitting algorithm to the set of points that starts at a peak sample and ends after a certain, pre-defined period of time. The calculated slope defines the decay for each peak; the number of decays (the same as number of detected peaks N_p) varies between frequency bands. Peak decays in each frequency bands define a vector per band (D_j), where j=(1,2,…, N_b).

The idea behind this step is that, as each peak corresponds to a short maximum in energy, ideally, the signal shortly after the peak corresponds to the energy decay (reverberation) which depends on the acoustic properties of the environment. In this way, an unknown acoustic environment is characterized by reverberation-related properties that help for classifying it to one of the pre-defined category. Although the approach used here is similar to the reverberation time estimation, it is important to distinguish the two; for the reverberation time estimation, the energy decay after the peak has to be 'clean' from the other audio events in order to capture only the properties of the enclosure while here such a condition is not required; the statistics applied later on the decay rate helps with obtaining the environment's properties related to the reverberation and not estimating the reverberation time values. Using the slope fitting, the reverberation properties are captured in the form of the decay slope.

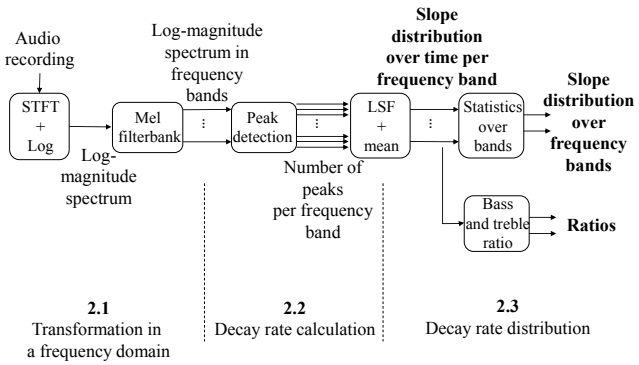

Figure 1: Reverberation-based feature extraction

2.3. Decay rate distribution

The decay distribution within each of the frequency bands is determined by terms of mean m_t,

$$m_t(j) = \frac{\sum_{i=1}^{N_p(j)} D_j(i)}{N_p(j)},\ j=(1,2,\ldots, N_b). \quad (1)$$

The result is a vector M_t of length equal to the number of frequency bands N_b, where each vector element represents the mean of decay distribution within bands over time m_t. The mean is used here as a well known statistical descriptor in order to characterize the distribution of the decay rates over time. Instead of the mean, other statistical parameters can be applied for obtaining the information of the decay rates population e.g. median, mode, variance etc. The resulting vector serves as a first part of a final DRD feature vector.

The second part of a final DRD feature vector is a result of decay distribution over frequency bands. For this purpose, mean m_b and skewness s_b of the vector obtained in the first step of the decay rate distribution (per band over time) are calculated,

$$m_b = \frac{1}{N_b}\sum_{j=1}^{N_b} m_t(j), \quad (2)$$

$$s_b = \frac{\frac{1}{N_b}\sum_{j=1}^{N_b}(m_t(j)-m_b)^3}{\left[\frac{1}{N_b-1}\sum_{j=1}^{N_b}(m_t(j)-m_b)^2\right]^{3/2}}. \quad (3)$$

The skewness parameter is added here in order to explore the asymmetry of the decay rate distribution over frequency bands. The idea behind the use of this parameter is that decay rate of different scenes shows different asymmetry of the distribution over frequency bands, e.g. more or less leaned towards low or high frequencies. This property of the decay rate distribution is shown in [15] where Wen *et al.* demonstrate the relationship between the skewness and the true decay rate. It was shown there that the distribution is 'skewed' more as the decay rate tends to zero.

Finally, the third part of a final DRD feature vector is created as a function of elements of the vector obtained in the first distribution step (per band over time). A function that defines ratio of decay rate distribution between low and mid frequency bands is bass ratio (BR), while treble ratio (TR) gives the ratio between high and mid frequency bands,

$$BR = \frac{M_t(b\rightarrow 125Hz)+M_t(b\rightarrow 250Hz)}{M_t(b\rightarrow 500Hz)+M_t(b\rightarrow 1kHz)} \quad (4)$$

$$TR = \frac{M_t(b\rightarrow 2kHz)+M_t(b\rightarrow 4kHz)}{M_t(b\rightarrow 500Hz)+M_t(b\rightarrow 1kHz)}. \quad (5)$$

The advantage of including the ratios is to reveal furthermore the differences of the scenes in terms of frequency band dependent content regarding decay rates. Bass and treble ratios are defined as the relative contribution of respectively low and high frequencies to the overall spectral energy. They are related to the subjective impressions of *warmth* and *brilliance* and they contribute to human ability to make a distinction between different acoustic environments [19].

3. ASC SYSTEM

The proposed feature extraction algorithm was tested against two different databases of acoustic scenes. The first database is our

non-public, in-house database, and the second is the official DCASE 2016 database. A state-of-the-art algorithm for ASC is implemented, based on Support Vector Machine (SVM) class of machine learning algorithms.

3.1 Baseline system

A system similar to the one proposed in [10] is used as a baseline system. A binary SVM classifier is used with complexity C=1; we used the radial basis function kernel (for the in-house dataset) with gamma $g=1/N_f$, where N_f is the number of audio features. For the DCASE database, a linear kernel was chosen, using pair-wise SVMs and majority voting for the multi-class problem. The first set of baseline audio features is made of 12 standard Mel-frequency cepstral coefficients (MFCC), with a window time of 20 ms and hop time of 10 ms, together with their delta coefficients. MFCCs are a generally accepted baseline feature set which has proven to be successful in many different audio analysis tasks [20]. The low-level features are summarized over each 6 s (in-house database) or 4 s (DCASE) window using four statistical functionals. As a first simple baseline, we use only mean and standard deviation as functional, on MFCCs and MFCC deltas, resulting in 48 features. This system is denoted as "MFCC baseline 1" in this paper. For a second baseline feature set, in addition, the mean, standard deviation, skewness and kurtosis are computed for the raw MFCCs. MFCC deltas use flatness, standard deviation, skewness, and percentile range as functional. Thus, in total, this feature set contains 96 features and it is used in "MFCC baseline 2" ASC system. A third baseline set is considered which, in addition to the 96 MFCC features, contains 140 features based on Mel filterbank coefficients. 26 Mel coefficients are computed, and post-processed with RASTA filtering [21], auditory weighting and liftering. In addition, the average of these coefficients and the average of the unprocessed Mel coefficients are used, resulting in 28 low-level descriptors. Five functionals are applied, which are the inter-quartile range 1-2 and 2-3, uplevel-time 25, uplevel-time 75 and rise-time. Thus, the third baseline feature set contains 236 features and it is used in "MFCC+Mel baseline" ASC system. All baseline feature sets were designed with the goal of low complexity in mind, aiming at a small feature set. The implementation of the features was inspired by the implementations in the openSMILE toolkit [22].

3.2 Reverb-based feature extraction implementation

The log-magnitude spectrum representation of an audio file is obtained by applying STFT with the window length of 64 ms and 16 ms hop size. The spectrum is calculated with a resolution of 1024 frequency bins. A perceptual filterbank based on 26 Mel frequency bands and 0-8kHz frequency range is used to split the spectrum into 26 frequency bands. For each frequency band, a peak detection algorithm with the magnitude threshold of 10dB was applied and a number of peaks per band are acquired. For each peak, a linear regression is done on a set of consecutive points from the peak to the end of 5 ms time window by terms of a linear least-square fitting. In this way, a slope of a fitted line for each peak defines a decay rate. By calculating a mean of the decays over time per frequency band, a first part of a DRD feature vector is obtained and it consists of 26 values where each represents decay rate distribution (mean over time) per frequency band (26 features). These 26 values are statistically analyzed by terms of mean and skewness and a second part of DRD feature vector is created with these two numbers (2 features). Finally, a third part of a DRD feature vector is calculated and it also consists of two numbers – BR and TR calculated as explained in Eq. (4) and (5) in the previous section (2 features). The ratios are obtained considering 2nd and 3rd band as low, 12th and 13th as mid and 24th and 25th as high frequency bands.

The final DRD feature vector of 30 elements is then combined with the MFCC baseline 2 and MFCC+Mel baseline feature sets resulting in 126 and 266 element feature vectors, respectively. The new feature vectors now containing DRD features are used with the SVM classifier for the purpose of ASC.

3.3 Audio databases for ASC

Experiments were carried out with two different databases of acoustic scenes. ASC models were trained using a training set, and the performance is evaluated on an independent test set, using the (weighted) average accuracy over all classes as an objective measure.

The first experiments used a Huawei in-house database, which contains audio recordings of two different classes: *car* and *other*, where *other* consists of bus and subway recordings. All classes correspond to moving vehicles and the recordings were made with the same smartphone in different conditions, e.g. device in the bag, in the hand, etc. The recordings are available as single-channel audio signals with a sampling rate of 16 kHz and 32 bit resolution. Overall, the database contains around 100 hours of recordings, recorded in many sessions of several minutes each. The two classes are equally represented in the database. The database is divided into a training set and test set, whereas recordings of one recording session cannot be in both sets. The training set and test set were both further split into small windows of 6 seconds. This way, the training set contains ca. 76,300 samples, and the test set contains ca. 22,000 samples.

The second set of experiments is performed with the publicly available database for the D-CASE 2016 challenge [23]. This dataset contains recordings of 15 different classes: lakeside beach, bus, cafe/restaurant, car, city center, forest path, grocery store, home, library, metro station, office, urban park, residential area, train, and tram, recorded with a high-quality binaural microphone. The recordings are split into segments of 30 seconds, and for each scene, 78 such segments are available. The classification decision should be made over a 30 second segment, and the system is evaluated using 4-fold cross validation, following the official protocol for the development set. We used the development set, since the test set labels are not yet publicly available. Training and test recordings are further segmented into segments of 4 seconds, with an overlap of 2 seconds. For the test recordings, the majority vote over all windows within the 30 seconds is used.

4. RESULTS

The results on the car-other dataset are shown in Table 1, for different combinations of the tested feature sets, i.e. for the three baseline feature sets, for the DRD feature set alone, and for the combinations of the baseline sets with the DRD features (for the combinations, the MFCC baseline 2 and MFCC+Mel baseline with 96 and 236 features are used). Results are shown separately for *car* and *other*, as well as the average accuracy; the table also lists the number of features (N_f) extracted for each case. Compared to the first baseline features (48 features, average 84.4% accuracy), our extended feature sets manage to increase the accuracy to 87.7%

and 90.0%, while keeping the number of features low. DRD improves the accuracy of MFCC baseline 2 system from 87.7% to 89.7%, and the accuracy of the large baseline system from 90.0% to 90.3%.

The results obtained with the publicly available DCASE 2016 dataset are given in Table 2. Here, we included the official baseline system given by the organizers of the challenge, two of our own implementations with different feature sets, as well as the results of some state-of-the-art methods published for the purpose of the challenge. We included a rough estimate of the system complexity, based on the number of features, training and test complexity of the classifier, and overall system complexity (e.g., fusion of several systems). All results are obtained with the official development set. The baseline system has a medium complexity and reaches 72.5% accuracy for the 15 classes. Using our ASC system based on SVM approach (described in Section 3.3), we achieve a result of 75.9%. This result is further improved by adding introduced DRD features, reaching 77.8%.

Table 1: ASC system accuracy on the internal dataset

Features	N_f	Car [%]	Other [%]	Average [%]
MFCC baseline 1	48	76.3	92.4	84.4
MFCC baseline 2	96	82.4	93.0	87.7
MFCC+Mel baseline	236	85.7	94.3	90.0
DRD	30	75.8	72.9	74.3
MFCC baseline 2 +DRD	126	85.3	94.1	89.7
MFCC+Mel baseline +DRD	266	86.2	94.4	90.3

The other results are obtained from participants of the 2016 DCASE challenge. We included some of the top-performing results in order to compare the accuracy of the proposed ASC system with other state-of-the-art methods in terms of feature number, complexity and accuracy. The best-performing system in the challenge reaches 89.9% accuracy and is based on fusing a system with i-vectors and a convolutional neural network (CNN) classifier. Using only the i-vector system, 80.8% can be obtained. Both systems make use of binaural multi-channel audio features. Using an NMF classifier enabled a result of 86.2%. A result of 81.4% was obtained using a DNN system in combination with a large feature space. This is only slightly more than our 77.8%, however it comes with a much higher system complexity. One participant achieved 79% with a tuned CNN system, which gives slightly better accuracy than our internal system but has a higher complexity.

5. CONCLUSIONS

We presented a strong, but low-complexity baseline system for acoustic scene classification, which is also improved with a novel class of audio features. The goal of involving new features is to improve the existing ASC algorithms in terms of accuracy by keeping at the same time the computational speed and number of the additional features low. We showed that adding the proposed reverberation-based (DRD) features to the baseline ASC system, the accuracy is increased for both internal and public databases. Additionally, the computation of the DRD features is fast, as the algorithmic complexity is low. The number of features is small compared to the baseline feature sets, which can help to keep the complexity of the classifier low.

With the internal database, the results in Section 4 show that MFCC features represent a very good baseline system, with an average accuracy of 87.7%. Adding Mel features results in an improvement, leading to up to 90.0%. This comes at the cost of a higher number of features, up to 236 instead of only 96 with MFCCs. Higher number of features means that the complexity for feature extraction is higher, as well as for classification. Furthermore, the memory size of the trained models will become larger. By adding only 30 DRD features to the MFCC baseline 2 system, the accuracy is increased by 2% and it is comparable with a more complex system that includes 236.

As for the public DCASE database it is shown again that adding the DRD features to the baseline MFCC features improves the accuracy of the classifier. The results show that the DRD features are complementary to the baseline feature set and can contribute to improving the accuracy of an ASC system. When compared to the other state-of-the-art solutions, it is concluded that most of the systems have a very high complexity, in terms of the employed algorithms, training time, model size, feature extraction, and classification. Furthermore, most of the top-performing challenge results are obtained by fusion. This means that different independent systems are built, and the final result is obtained from a combination of the independent system predictions. This adds a lot to the complexity.

Future work will involve further development of the described ASC system in order to increase accuracy while keeping the low-complexity of both the feature extractor and system classifier. The proposed DRD feature extractor is going to be broadened to the multichannel case, where we can exploit the spatial recording setup and binaural features of audio signals in order to get a more sophisticated measure of the acoustic properties in terms of reverberation. Another classifier types (GMM, DNN...) will be considered and a potential usage of DRD feature extractor for a signal pre-processing in combination with them will be analyzed and explored.

Table 2: ASC accuracy for the DCASE 2016 dataset, for various feature sets and state-of-the-art methods

Origin	Features	Class-ifier	Compl-exity	Average accuracy
Official Baseline	MFCC	GMM	medium	72.5%
Huawei Media GRC	MFCC	SVM	low	75.9%
Huawei Media GRC	MFCC + DRD	SVM	low	77.8%
University Marche, Ancona, Tampere University [24]	spectrogram	CNN	high	79.0%
University Passau, audEERI-NG [25]	spectral, cepstral, energy, voicing, auditory	DNN, subspace learning, fusion	very high	81.4%
Telecom ParisTech [26]	spectrogram	NMF	high	86.2%
J. Kepler University of Linz [27]	i-vectors, binaural	LDA, WCCN scoring	high	80.8%
J. Kepler University of Linz [27]	i-vectors, binaural, spectrogram	CNN and system fusion	very high	89.9%

6. REFERENCES

[1] D. Wang and G. J. Brown, *Computational auditory scene analysis: Principles, algorithms, and applications*, Wiley interscience, 2006.

[2] L. Ma, B. Milner, and D. Smith, "Acoustic environment classification", ACM Transactions on Speech and Language Processing (TSLP), 3(2):1–22, 2006.

[3] 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 (pp. 1-4). 2013.

[4] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," In 24th European Signal Processing Conference, 2016.

[5] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, "Acoustic event detection in real life recordings", Proceedings of Signal Processing Conference, pp. 1267-1271, 2010.

[6] H. Lu, W. Pan, N. Lane, T. Choudhury and A. Campbell, "Soundsense: Scalable Sound Sensing for People-centric Applications on Mobile Phones", MobiSys '09, pp. 165–178, 2009.

[7] N. Lane, P. Georgiev and L. Qendro, "DeepEar: robust smartphone audio sensing in unconstrained acoustic environments using deep learning", Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 283-294, 2015.

[8] D. Barchiesi, D. Giannoulis, D. Stowell and M. D. Plumbley, "Acoustic Scene Classification", IEEE Signal Processing Magazine, pp. 16-34, 2015.

[9] Z. Liu, Y. Wang, and T. Chen, "Audio feature extraction and analysis for scene segmentation and classification," Journal of VLSI signal processing systems for signal, image and video technology, vol. 20, no. 1-2, pp. 61–79, 1998.

[10] J. T. Geiger, B. Schuller, and G. Rigoll, "Large-scale audio feature extraction and SVM for acoustic scene classification," 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2013.

[11] A. Rakotomamonjy and G. Gasso, "Histogram of gradients of time-frequency representations for audio scene classification", IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events, 2013.

[12] G. Roma, W. Nogueira and P. Herrera, "Recurrence quantification analysis features for auditory scene classification", IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events, 2013.

[13] R. M. Schädler, B. T. Meyer, and B. Kollmeier. „Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition". The Journal of the Acoustical Society of America, 131(5), 4134-4151. 2012.

[14] J. Geiger and K. Helwani, "Improving event detection for audio surveillance using gabor filterbank features", EUSIPCO, 2015.

[15] J. Wen, E. Habets and P. Naylor, "Blind estimation of reverberation time based on the distribution of signal decay rates", IEEE International Conference on Acoustics, Speech and Signal Processing, Las Vegas, NV, 2008.

[16] Sampo Vesa and Aki Härmä, "Automatic estimation of reverberation time from binaural signals," in Proceedings of the 30th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005), Philadelphia, PA, Mar. 2005, vol. 3, pp. 281-284.

[17] T. M. Prego, A. A. de Lima, R. Z. Lopez, and S. L. Netto, "Blind estimators for reverberation time and direct-to-reverberant energy ratio using sub-band speech decomposition", IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2015.

[18] H. Jiang, J. Bai, S. Zhang, and B. Xu, "Svm-based audio scene classification", in Proc. Natural Language Processing and Knowledge Engineering (NLP-KE), IEEE, pp. 131–136, 2005.

[19] H. Kuttruff, "Room Acoustics", Elsevier Applied Science, 1991.

[20] D. Stowell, D. Giannoulis and E. Benetos, "Detection and Classification of Acoustic Scenes and Events", IEEE Transactions on multimedia, vol. 17, no. 10, pp. 1733-1746, 2015.

[21] H. Hermansky, N. Morgan, "RASTA Processing of Speech", IEEE Transactions on speech and audio processing, vol. 2. no. 4, pp.578-589, 1994.

[22] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the Munich open-source multimedia feature extractor", Proceedings of the 21st ACM international conference on Multimedia, pp. 835-838, 2013.

[23] http://www.cs.tut.fi/sgn/arg/dcase2016/

[24] M. Valenti, A. Diment, G. Parascandolo, S. Squartini and T. Virtanen, "DCASE 2016 acoustic scene classification using convolutional neural networks", Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016), pp. 95-99, 2016.

[25] E. Marchi, D. Tonelli, X. Xu, F. Ringeval, J. Deng, S. Squartini & B. Schuller, "Pairwise decomposition with deep neural networks and multiscale kernel subspace learning for acoustic scene classification", Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016), pp. 65-69, 2016.

[26] V. Bisot, R. Serizel, S. Essid and G. Richard, "Supervised nonnegative matrix factorization for acoustic scene classification", Workshop on Detection and Classification of Acoustic Scenes and Events 2016 (DCASE2016), technical report, 2016.

[27] H. Eghbal-Zadeh, B. Lehner, M. Dorfer, and G. Widmer, "CP-JKU submissions for dcase-2016: a hybrid approach using binaural i-vectors and deep convolutional neural networks", Workshop on Detection and Classification of Acoustic Scenes and Events 2016 (DCASE2016), technical report, 2016.