INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Acoustic Scene Classification using a CNN-SuperVector system trained with Auditory and Spectrogram Image Features

Rakib Hyder[1], *Shabnam Ghaffarzadegan*[2], *Zhe Feng*[2], *John H.L. Hansen*[3], *Taufiq Hasan*[1]

[1]Bangladesh University of Engineering and Technology (BUET), Dhaka, Bangladesh.
[2]Robert Bosch Research and Technology Center (RTC), Palo Alto, CA.
[3]Center for Robust Speech Systems (CRSS), The University of Texas at Dallas, Richardson, TX.

rakib.hyder.bd@ieee.org, {shabnam.ghaffarzadegan,zhe.feng2}@us.bosch.com,
john.hansen@utdallas.edu, taufiq@bme.buet.ac.bd

Abstract

Enabling smart devices to infer about the environment using audio signals has been one of the several long-standing challenges in machine listening. The availability of public-domain datasets, e.g., Detection and Classification of Acoustic Scenes and Events (DCASE) 2016, enabled researchers to compare various algorithms on standard predefined tasks. Most of the current best performing individual acoustic scene classification systems utilize different spectrogram image based features with a Convolutional Neural Network (CNN) architecture. In this study, we first analyze the performance of a state-of-the-art CNN system for different auditory image and spectrogram features, including Mel-scaled, logarithmically scaled, linearly scaled filterbank spectrograms, and Stabilized Auditory Image (SAI) features. Next, we benchmark an MFCC based Gaussian Mixture Model (GMM) SuperVector (SV) system for acoustic scene classification. Finally, we utilize the activations from the final layer of the CNN to form a SuperVector (SV) and use them as feature vectors for a Probabilistic Linear Discriminative Analysis (PLDA) classifier. Experimental evaluation on the DCASE 2016 database demonstrates the effectiveness of the proposed CNN-SV approach compared to conventional CNNs with a fully connected softmax output layer. Score fusion of individual systems provides up to 7% relative improvement in overall accuracy compared to the CNN baseline system.

Index Terms: audio event detection, acoustic scene classification.

1. Introduction

Advances in automatic speech recognition (ASR) [1,2] and music information retrieval (MIR) [3, 4] have brought us systems that can transcribe speech signals and music chords in a variety of adverse conditions. However, speech and music are only two of many indoor and outdoor audio signals that surround us. With increasing users of hand-held devices listening in diverse environments, e.g., smartphones and tablets, the research question is whether *machine listening* can perform on par with humans in deciphering the environmental audio signal.

Audio Event Detection (AED) and Acoustic Scene Classification (ASC), with the goal of understanding the environment and detecting events and anomalies, are among the growing topics in the research community [5–7]. In particular, information about non-speech sounds can be used as a complementary source of information along with other modalities to analyze human and social activities, monitor machines and infrastructures, improve context awareness for speech applications [8], smart homes and industrial facilities, etc. However, there are many challenges to be addressed in AED and ASC, such as (i) lack of a fundamental set of acoustic units similar to phonemes and words, (ii) presence of non-stationary noises in the environment, (iii) rare occurrence of events, and (iv) unexpected occurrence of events.

AED and ASC methods can be categorized into two main approaches: *supervised* and *unsupervised* methods. For supervised methods, different features have been used in the literature. ASR inspired features such as: Mel-frequency Cepstral Coefficients (MFCC) [9], Zero Crossing Rate (ZCR) [10], and so on have been prominently used in the field [11]. However, the state-of-the-art results have been achieved using more complicated representations such as auditory images [12], Spectrogram Image Feature (SIF) [13], spectrogram-derived Sub-band Power Distribution (SPD) [14], etc. Moreover, feature vectors derived from these methods are used to train different machine learning models such as: Support vector Machine (SVM) [15], Gaussian Mixture Model (GMM) [16], Hidden Markov Model (HMM) [17], Deep Neural Networks (DNNs) [6,18,19], etc. On the other hand, very few studies have been focused on unsupervised event and scene detection with limited labeled data available. For instance, in [20], a semi-supervised method has been proposed in which an unusual event is derived from the usual event model at each step of an iterative process via Bayesian adaptation. In [21], a sound event detection system capable of handling overlapping multi-source environments is proposed. In [22], an unsupervised feature learning method has been explored for acoustic scene classification.

In this study, we aim to examine different auditory and spectrogram image features for ASC using a CNN architecture. SIF with CNNs have been shown to yield the best performing independent submissions in the DCASE 2016 ASC challenge [6, 19]. We also investigate the performance of a GMM-SuperVector (SV) system [23] with a Probabilistic Linear Discriminant Analysis (PLDA) [24] classifier which closely follows our previous work on abnormal audio event detection [25]. We hypothesize that the CNN is superior in feature learning whereas the PLDA classifier's strength lies in classifying high dimensional vectors, as it was successfully used with I-vectors [26] in speaker recognition [27, 28]. In an attempt to combine the strengths of CNN and PLDA, we use the activation from the last layer of CNN and form a CNN-SuperVector which is then post-processed and classified by a PLDA model. Finally, a score-level fusion of multiple systems is investigated to achieve the best possible ASC on the DCASE 2016 dataset.

http://dx.doi.org/10.21437/Interspeech.2017-431

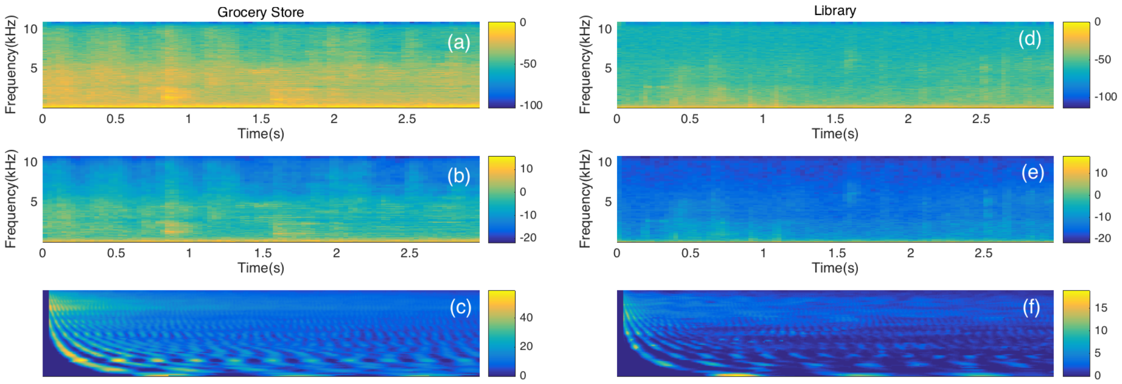

Figure 1: *Auditory and Spectrogram Image Features (SIF) obtained from two different acoustic scenes, namely: grocery store (a-c) and library (d-f). Features include: (a,d) Linear-scaled SIF, (b,e) Log-scaled SIF, and (c,f) SAI features.*

2. Dataset

In this work, we utilize the DCASE 2016 acoustic scene classification challenge data [29]. The dataset consists of audio samples from 15 (fifteen) different indoor and outdoor locations or environments. There are 1170 and 390 audio segments in the development and evaluation data, respectively. The 2-channel audio segments are 30s in duration and are recorded in a 24-bit PCM format at 44100Hz sampling rate. The ASC development dataset is designed as a four-fold cross-validation task with about 75% data used for training and the remaining 25% for testing. The average accuracy over the folds is used as the performance evaluation metric [29].

3. Baseline CNN system

3.1. Spectrogram Image Features (SIF)

We use a single channel spectrogram excerpt with a dimension of 149 × 149 as input to our CNN system. This setup closely follows the top-scoring system in DCASE 2016 [6]. First, the audio data is down-sampled to a rate of $22,050$ Hz and segmented at 31.25 frames/sec. Next, Short-Time Fourier Transform (STFT) is computed on 2048 sample time windows. In [6], a logarithmic filterbank was used with 24 bands per octave extracted within a passband of 20 Hz to 11.025 kHz with the MADMOM toolkit [30] which resulted in 149 frequency bins. Then the spectrogram was segmented into 149 × 149 frames with 25% overlap in the temporal dimension. These spectrogram excerpts are the inputs to the CNN system for an audio segment. To investigate the effect of changing the frequency scale of the filterbanks, we used linear-scaled, log-scaled and Mel-scaled filterbanks with the same number of frequency bins to generate corresponding SIF segments from each STFT window. Although such high-dimensional Mel-spectrograms are not used in traditional speech processing, this has been found to provide significant performance gains for acoustic scene classification [19]. In our experiments, we observed about 9% absolute improvement in average accuracy by increasing the Mel-spectrogram dimension from 13 to 149.

3.2. Stabilised Auditory Image (SAI)

Stabilized auditory image (SAI) features were first proposed in [31]. These features utilize an auditory nerve model to mimic the functionality of the inner ear and generate a two-dimensional image. SAI features have been used for audio event detection [32] and thus we intended to investigate its effectiveness for the CNN classifier on the DCASE 2016 ASC task. We used [33] to calculate SAI features from 3s windows without overlap . SAI calculation is known to be computationally demanding [31]. To reduce the computational burden, the audio data is further down-sampled to 8kHz. Finally, the extracted SAI features yield 75 × 280 dimensional image from each 3s audio segment. We then further downsample this image in temporal dimension to bring it down to 75 × 75 dimensional image which we pass to the CNN. Different SIFs along with SAI features are shown for two different acoustic scenes in Fig. 1. Here, we observe that while the different acoustic scenes affect distinct time-frequency regions of SIF features, the patterns in the entire auditory-image varies in case of SAI.

3.3. CNN architecture

We followed [19] with some modifications to build our deep CNN architecture. A block diagram representation of our model is presented in Fig. 2. The first layer performs a convolution over the input spectrogram with 128 kernels with 3 × 3 kernel size and unitary depth and stride in both dimensions. The obtained feature maps are then sub-sampled using a max-pooling layer operating over 3×3 non-overlapping squares. The second convolutional layer is very similar to the first one except with a higher number of kernels (256 instead of 128). The second and last sub-sampling operation is performed aiming to remove the temporal axis. Therefore, we use a max-pooling layer which operates over the entire sequence length and, on the frequency axis, only over 3 non-overlapping frequency bands. Rectified linear unit (ReLU) [34] activation functions are used for the kernels in both convolutional layers. Finally, to classify the audio segment in 15 classes, the output layer consists of a 15-node fully-connected neural network with a softmax activation function.

3.4. Regularization, Optimization and Model Training

We utilized batch normalization [35] as an intermediate layer after each of the two convolutional layers. We also used a dropout layer for regularization (dropout $= 0.25$) after each of the two max-pooling layers. The CNN system is implemented with the Keras Python library. We utilized the categorical cross-entropy loss function and the Adaptive Momentum (ADAM) [36] optimization approach. We set exponential decay rate for the moment estimates of the ADAM algorithm as $\beta_2 = 0.99999$ and $\beta_1 = 0.9$ to reduce the weights on the previous time stamps. We kept the default value for the parameter, ϵ (10^{-8}). Finally,

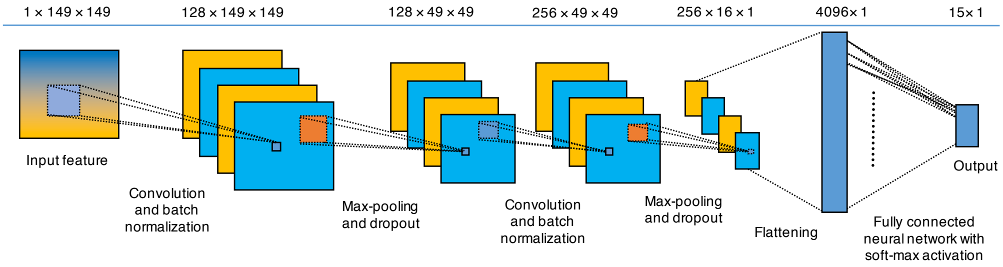

Figure 2: *A flow-diagram of the utilized CNN architecture for the experiments using auditory and spectrogram image features. The proposed CNN-SuperVectors are formed from the flattening layer activations which are input to softmax output layer.*

we used a learning rate of 0.0001 with 200 training epochs. The classification decision for an audio segment is calculated by averaging the prediction scores obtained from the short segments.

4. GMM-SuperVector system

The GMM-SuperVector framework was first utilized in speaker recognition [23]. This method generates a high-dimensional vector by concatenating the GMM mean vectors that model specific audio segments using the Maximum *A Posteriori* (MAP) adaptation [37]. These SuperVectors (SV) are then used as features for other classifiers, e.g. SVM.

4.1. Acoustic features

For the GMM-SV system, we extract 60 dimensional Mel-frequency cepstral coefficients (MFCC) [9], where 19 static coefficients are computed including C_0, and the velocity (Δ) and acceleration ($\Delta + \Delta$) coefficients are appended.

4.2. GMM training and MAP adaptation

Initially, a GMM is trained on the DCASE training data for the corresponding fold. This is a generic acoustic scene independent model, known as the Universal Background Model (UBM) in speaker recognition literature [37, 38]. Next, audio segment dependent GMM parameters are estimated using MAP adaptation. For features $\mathbf{x}_n$, a GMM-UBM model λ_0 with M Gaussian components is represented as:

$$f(\mathbf{x}_n|\lambda_0) = \sum_{g=1}^{M} \pi_g \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g), \tag{1}$$

where, $\mathcal{N}(\cdot,\cdot)$, π_g, $\boldsymbol{\mu}_g$ and $\boldsymbol{\Sigma}_g$ indicate a normal distribution, the g-th component weight, mean, and covariance matrix, respectively. Next, we define $\gamma_n(g) = p(g|\mathbf{x}_n, \lambda_0)$ and calculate the sufficient statistics from features $\mathcal{X}_s$ in audio segment s:

$$N_s(g) = \sum_{n\in\mathcal{X}_s} \gamma_n(g) \quad \text{and} \quad \mathbf{F}_s(g) = \sum_{n\in\mathcal{X}_s} \gamma_n(g)\mathbf{x}_n. \tag{2}$$

We compute the posterior means given $\mathcal{X}_s$ for component g as:

$$E_g[\mathbf{x}_n|n \in \mathcal{X}_s] = \frac{\mathbf{F}_s(g)}{N_s(g)} \tag{3}$$

MAP [37, 39] adapted mean vectors for $\mathcal{X}_s$ are given by:

$$\hat{\boldsymbol{\mu}}_{g,s} = \alpha_g E_g[\mathbf{x}_n|n \in \mathcal{X}_s] + (1 - \alpha_g)\boldsymbol{\mu}_g \tag{4}$$

where, the α_g is the MAP adaptation parameter computed as: $\alpha_g = N_s(g)/(N_s(g) + r)$ where $r > 0$ is defined as the relevance factor [37]. Finally, the GMM-SV from segment s is obtained by concatenating the adapted mean vectors as:

$$\mathbf{m}_s = \begin{bmatrix} \hat{\boldsymbol{\mu}}_{1,s} \\ \hat{\boldsymbol{\mu}}_{2,s} \\ \vdots \\ \hat{\boldsymbol{\mu}}_{M,s} \end{bmatrix}. \tag{5}$$

We use $M = 64$ component GMMs trained using an Expectation Maximization (EM) algorithm with 5 iterations per mixture split. For MAP adaptation, $r = 14$ is used. Finally, 3840 (64×60) dimensional SVs are extracted from each training and test segment.

4.3. SV post-processing

We first perform mean normalization across the training SVs. Next, we divide each vector by it's own L^2 norm for length normalization [40]. The resulting vectors are then reduced in dimension to 14 using a Linear Discriminant Analysis (LDA) projection and normalized using the Within Class Covariance Normalization (WCCN) [26]. The parameters required in the post-processing steps are learned from the training data only and applied on the evaluation data.

4.4. Probabilistic Linear Discriminant Analysis (PLDA)

We utilize a Gaussian PLDA classifier with a full-covariance residual noise [40]. In this model, an R dimensional post-processed SV extracted from audio segment s is expressed as:

$$\mathbf{m}_s = \mathbf{m}_0 + \boldsymbol{\Phi}\beta + \mathbf{n}. \tag{6}$$

Here, $\mathbf{m}_0 \in \mathbb{R}^R$ is the acoustic scene independent mean vector, $\boldsymbol{\Phi}$ is an $R \times N$ low rank matrix representing the scene dependent basis functions, $\beta \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is an $K \times 1$ hidden vector, and $\mathbf{n} \in \mathbb{R}^R$ is a random vector representing the full covariance residual noise. We train our model using the averaged post-processed SV obtained from each class and set $K = 14$. The DCASE training data of the corresponding fold was used to train the PLDA model. The scoring is performed as described in [41]. To determine if $\mathbf{m}_i$ and $\mathbf{m}_j$ belong to the same class (H_1) or not (H_0), we use the following likelihood ratio:

$$\mathcal{L}_{i,j} = \frac{P(\mathbf{m}_i, \mathbf{m}_j|H_1)}{P(\mathbf{m}_i|H_0)P(\mathbf{m}_j|H_0)}. \tag{7}$$

This comparison is performed across all training and test segments to determine the highest scoring class for each test.

5. Proposed CNN-SuperVector system

In order to combine the feature learning strength of CNN with the SV back-end, we formed a high dimensional vector concatenating the activations from the flattening layer (Fig. 2) of a

Table 1: *Performance evaluation of the proposed CNN, GMM-SV and CNN-SV systems with different auditory image and spectrogram features on the DCASE 2016 ASC development dataset. %Accuracy for each fold and their average values are reported.*

ID	Feature	System	% Accuracy				
			Fold1	Fold2	Fold3	Fold4	Average
Sys. 1	75 × 75 SAI Features	Baseline CNN	57.24	61.65	60.82	61.52	60.31
Sys. 2	149 × 149 Linear-scaled filterbank SIF	Baseline CNN	74.13	72.45	73.87	72.75	73.30
Sys. 3	149 × 149 Log-scaled filterbank SIF	Baseline CNN	81.32	78.07	79.40	83.69	80.62
Sys. 4	149 × 149 Mel-scaled filterbank SIF	Baseline CNN	84.79	**79.84**	76.87	78.12	79.90
Sys. 5	60D MFCC+Δ+$\Delta\Delta$	GMM-SV-PLDA	81.73	75.54	81.47	85.82	81.14
Sys. 6	149 × 149 Linear-scaled filterbank SIF	CNN-SV-PLDA	83.42	73.82	80.94	77.27	78.86
Sys. 7	149 × 149 Log-scaled filterbank SIF	CNN-SV-PLDA	83.67	77.96	**83.17**	**87.23**	**83.01**
Sys. 8	149 × 149 Mel-scaled filterbank SIF	CNN-SV-PLDA	**87.45**	76.40	82.09	81.08	81.76

Table 2: *Performance evaluation of the proposed CNN, GMM-SV, CNN-SV and fusion systems with different spectrogram image features on the DCASE 2016 ASC development and evaluation dataset.*

ID	System	%Accuracy (Dev)	%Accuracy (Eval)
Ref [6]	149 × 149 Log Spectrogram with CNN [6]	79.49	83.30
Ref [19]	60 × 60 Log Mel-spectrogram with CNN [19]	79.00	86.20
Sys. 3	149 × 149 Log-scaled filterbank SIF-CNN	80.62	85.38
Sys. 4	149 × 149 Mel-scaled filterbank SIF-CNN	79.90	84.87
Sys. 5	60D MFCC-GMM-SV-PLDA	81.14	83.85
Sys. 7	149 × 149 Log-scaled filterbank SIF CNN-SV-PLDA	83.01	87.95
Sys. 8	149 × 149 Mel-scaled filterbank SIF CNN-SV-PLDA	81.76	87.18
Fusion-1	Linear score fusion of Systems (5+7)	**86.23**	86.67
Fusion-2	Linear score fusion of Systems (7+8)	84.07	**88.46**

trained CNN system and fed it as an input to the PLDA back-end. In this system, training is performed in two stages. In the first stage, the CNN model is trained on the SIF/SAI features on the training dataset as described in Sec. 3. Once the model is trained, all the training and test data is evaluated using the CNN model and the flattening layer activations are combined to form a high-dimensional super vector similar to Sec. 4 (e.g. 4096 dimensions in case of log-scaled SIF). Next, the extracted training CNN-SV features are post-processed by LDA, WCCN and length normalization according to Sec. 4.3. The processed CNN-SV features are used to train the PLDA model as described in Sec. 4.4.

6. Experimental Evaluation

System development experiments are performed using the four folds of the DCASE 2016 ASC development data. The development and evaluation results are reported in Tables 1 and 2, respectively. In Table 1, Sys. 1–4 compare different SIF and SAI features. Here, we observe that with a fixed 149×149 sized SIF, the performance difference between logarithmic scaled and Mel-scaled SIF is negligible with about 80% average accuracy, whereas the linear scaled SIF shows inferior performance at 73.3% mean accuracy. In contrast, SAI features perform significantly worse with a mean accuracy of 60.31%. This may have been due to the effect of down-sampling and input image size reduction done to improve the computational efficiency. Further examination using this feature with higher resolution images may be necessary to faithfully compare SAI performance with other SIF images. Our GMM-SV system (Sys. 5 of Table 1) shows slightly improved mean accuracy compared to the CNN systems.

Finally, the proposed CNN-SV configurations, i.e., Sys. 6–8 of Table 1, consistently outperform their CNN counterparts. For example, the Log-SIF-CNN system (Sys. 3) shows an accuracy of 80.62% whereas the Log-SIF-CNN-SV with PLDA (Sys. 7) reaches an accuracy of 83.01%. In Table 2, we observe that this performance gain is also consistently achieved in the evaluation dataset, achieving an accuracy of 87.95% with the Log-SIF-CNN-SV configuration (Sys. 7). Thus the results indicate the effectiveness of using the PLDA back-end with the proposed CNN-SV features. Results reported in [6] and [19] on the DCASE development and evaluation set are also shown in Table 2 for comparison.

To determine if the systems contain complimentary information, we performed simple linear score fusion of a few selected systems. The fusion process consists of mean and range normalization (divide by the maximum value) of scores obtained from the 15 classes in each test segment. The normalized scores are then averaged across different systems to obtain the fused scores. As noted in Table 2, fusion of Log-SIF CNN-SV (Sys. 7) and Mel-SIF CNN-SV (Sys. 8) provides the best performance on the evaluation data reaching an accuracy of 88.46%, while fusion of MFCC-GMM-SV-PLDA (Sys. 5) and Log-SIF CNN-SV (Sys. 7) provides the highest accuracy of 86.23% in the development data. The latter represents a 7% relative improvement over the baseline CNN system (Sys. 3).

7. Conclusions

In this paper, we have analyzed the performance of a CNN based acoustic scene classification system using different auditory and spectrogram image features. We have also implemented an MFCC based GMM-SuperVector system using a PLDA classifier. Observing the advantage of a GMM-SV-PLDA system, we have utilized a higher dimensional CNN-SuperVector by concatenating the output of the final layer of the CNN activations and used them as features in a PLDA classifier. The evaluation results in DCASE 2016 acoustic scene classification dataset demonstrate the effectiveness of the CNN-SV approach compared to individual CNN and GMM-SV systems. Further improvement in system accuracy has been attained by score fusion of multiple systems.

8. References

[1] M. L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Proc. IEEE ICASSP*, May 2013, pp. 7398–7402.

[2] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-r. Mohamed, N. Jaitly, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep Neural Networks for Acoustic Modeling in Speech Recognition," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, 2012.

[3] A. L.-C. Wang, "An Industrial Strength Audio Search Algorithm." *Proc. ISMIR*, pp. 7–13, Oct. 2003.

[4] E. Benetos and S. Dixon, "Multiple-instrument polyphonic music transcription using a temporally constrained shift-invariant model." *J. Acoust. Soc. Am.*, vol. 133, no. 2009, pp. 1727–41, 2013.

[5] M. Crocco, M. Cristani, A. Trucco, and V. Murino, "Audio Surveillance: a Systematic Review," *ACM Comput. Surv.*, vol. 48, no. 4, pp. 1–44, 2016.

[6] 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," in *Workshop on DCASE 2016*, Budapest, Hungary, Sep. 2016.

[7] G. Takahashi, T. Yamada, S. Makino, and N. Ono, "Acoustic Scene Classification Using Deep Neural Network and Frame-Concatenated Acoustic Feature," in *DCASE2016 Chall.*, 2016.

[8] M. Akbacak and J. H. Hansen, "Environmental sniffing: noise knowledge estimation for robust speech systems," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 15, no. 2, pp. 465–477, 2007.

[9] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 28, no. 4, pp. 357–366, Aug 1980.

[10] R. G. Bachu, S. Kopparthi, B. Adapa, and B. D. Barkana, "Separation of Voiced and Unvoiced using Zero crossing rate and Energy of the Speech Signal," *Adv. Tech. Comput. Sci. Softw. Eng.*, pp. 279–282, 2010.

[11] S. Chu, S. Narayanan, and C. C. J. Kuo, "Environmental sound recognition with timeFrequency audio features," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 17, no. 6, pp. 1142–1158, 2009.

[12] T. C. Walters, "Auditory- Based Communication," Ph.D. dissertation, University of Cambridge, 2011.

[13] J. Dennis, H. D. Tran, and H. Li, "Spectrogram image feature for sound event classification in mismatched conditions," *IEEE Signal Process. Lett.*, 2011.

[14] J. Dennis, H. D. Tran, and E. S. Chng, "Image feature representation of the subband power distribution for robust sound event classification," *IEEE Trans. Audio, Speech Lang. Process.*, 2013.

[15] G. Guo and S. Z. Li, "Content-based audio classification and retrieval by support vector machines." *IEEE Trans. Neural Netw.*, vol. 14, no. 1, pp. 209–215, 2003.

[16] C. Clavel and T. Ehrette, "Events Detection for an Audio-Based Surveillance System," in *WIT Trans. Inf. Commun. Technol.*, 2008.

[17] X. Zhou, X. Zhuang, M. Liu, H. Tang, M. Hasegawa-Johnson, and T. Huang, "HMM-based acoustic event detection with adaboost feature selection," in *Lect. Notes Comput. Sci.*, 2008.

[18] E. Marchi, D. Tonelli, X. Xu, F. Ringeval, J. Deng, and B. Schuller, "The UP system for the 2016 DCASE challenge using deep recurrent neural network and multiscale kernel subspace learning," 2016.

[19] M. Valenti, A. Diment, G. Parascandolo, S. Squartini, and T. Virtanen, "DCASE 2016 acoustic scene classification using convolutional neural networks," in *Workshop on DCASE 2016*, Budapest, Hungary, Sep. 2016.

[20] D. Zhang, D. Gatica-Perez, S. Bengio, and I. McCowan, "Semi-supervised adapted hmms for unusual event detection," in *Proc. IEEE CVPR*, vol. 1, 2005, pp. 611–618.

[21] T. Heittola, A. Mesaros, T. Virtanen, and A. Eronen, "Sound event detection in multisource environments using source separation," *Work. Mach. List. Multisource Environ.*, pp. pp. 36–40, 2011.

[22] V. Bisot, R. Serizel, S. Essid, and G. Richard, "Acoustic scene classification with matrix factorization for unsupervised feature learning," in *Proc. IEEE ICASSP*, 2016, pp. 6445–6449.

[23] W. Campbell, D. Sturim, D. Reynolds, and A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," in *Proc. IEEE ICASSP*, May 2006, pp. 97–100.

[24] S. Ioffe, "Probabilistic linear discriminant analysis," in *Computer Vision–ECCV 2006*. Springer, 2006, pp. 531–542.

[25] M. K. Nandwana and T. Hasan, "Towards smart-cars that can listen: Abnormal acoustic event detection on the road," in *Interspeech 2016*, Sep 2016, pp. 2968–2971.

[26] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front end factor analysis for speaker verification," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[27] P. Matejka, O. Glembek, F. Castaldo, M. Alam, O. Plchot, P. Kenny, L. Burget, and J. Cernocky, "Full-covariance UBM and heavy-tailed PLDA in i-vector speaker verification," in *Proc. IEEE ICASSP*, Prague, Czech Republic, May 2011, pp. 4828 – 4831.

[28] T. Hasan, S. O. Sadjadi, G. Liu, N. Shokouhi, H. Bořil, and J. H. Hansen, "CRSS Systems for 2012 NIST Speaker Recognition Evaluation," in *Proc. IEEE ICASSP*, Vancouver, Canada, May. 2013.

[29] "Detection and Classification of Acoustic Scenes and Events 2016," http://www.cs.tut.fi/sgn/arg/dcase2016/.

[30] S. Böck, F. Korzeniowski, J. Schlüter, F. Krebs, and G. Widmer, "madmom: A new Python Audio and Music Signal Processing Library," in *Proc. ACM Mult. Conf.*, Amsterdam, The Netherlands, 10 2016, pp. 1174–1178.

[31] R. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand, "Complex sounds and auditory images," *Auditory physiology and perception*, vol. 83, pp. 429–446, 1992.

[32] H. Zhang, I. McLoughlin, and Y. Song, "Robust sound event recognition using convolutional neural networks," in *Proc. IEEE ICASSP*, 2015, pp. 559–563.

[33] S. Bleeck, T. Ives, and R. D. Patterson, "Aim-mat: The auditory image model in matlab," *Acta Acustica United with Acustica*, vol. 90, no. 4, pp. 781–787, 2004.

[34] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks." in *Aistats*, vol. 15, no. 106, 2011, p. 275.

[35] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," *arXiv:1502.03167*, pp. 1–11, 2015.

[36] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[37] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Process.*, vol. 10, no. 1–3, pp. 19 – 41, 2000.

[38] T. Hasan and J. H. L. Hansen, "A study on universal background model training in speaker verification," *IEEE Trans. Audio, Speech, Lang. Process.*, pp. 1890–1899, Sep. 2011.

[39] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE Trans. Speech, Audio Process.*, vol. 2, no. 2, pp. 291–298, 1994.

[40] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Proc. Interspeech*, Florence, Italy, Oct. 2011, pp. 249 – 252.

[41] P. Kenny, "Bayesian speaker verification with heavy tailed priors," in *Proc. Odyssey*, Brno, Czech Republic, 2010.