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

Unsupervised Filterbank Learning Using Convolutional Restricted Boltzmann Machine for Environmental Sound Classification

Hardik B. Sailor, Dharmesh M. Agrawal, and Hemant A. Patil

Speech Research Lab, Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar, India

{sailor_hardik, dm_agrawal, hemant_patil}@daiict.ac.in

Abstract

In this paper, we propose to use Convolutional Restricted Boltzmann Machine (ConvRBM) to learn filterbank from the raw audio signals. ConvRBM is a generative model trained in an unsupervised way to model the audio signals of arbitrary lengths. ConvRBM is trained using annealed dropout technique and parameters are optimized using Adam optimization. The subband filters of ConvRBM learned from the ESC-50 database resemble Fourier basis in the mid-frequency range while some of the low-frequency subband filters resemble Gammatone basis. The auditory-like filterbank scale is nonlinear w.r.t. the center frequencies of the subband filters and follows the standard auditory scales. We have used our proposed model as a front-end for the Environmental Sound Classification (ESC) task with supervised Convolutional Neural Network (CNN) as a back-end. Using CNN classifier, the ConvRBM filterbank (ConvRBM-BANK) and its score-level fusion with the Mel filterbank energies (FBEs) gave an absolute improvement of 10.65 %, and 18.70 % in the classification accuracy, respectively, over FBEs alone on the ESC-50 database. This shows that the proposed ConvRBM filterbank also contains highly complementary information over the Mel filterbank, which is helpful in the ESC task.

Index Terms: Unsupervised Filterbank Learning, ConvRBM, Sound Classification, CNN.

1. Introduction

Environmental sound classification is a growing research problem in the multimedia applications. The environmental sounds are a very diverse group of everyday audio events that cannot be described as only speech or music [1]. The environmental sounds are important for understanding the content of the multimedia. Therefore, the environmental sound classification (ESC) technology development is better for characterizing the essential role of environmental sounds in many multimedia applications, such as, audio scenes classification [2], audio surveillance system [3], hearing aids [4], smart room monitoring [5] and video content highlight generation [6], etc. In the previous approaches, the ESC is typically conducted based on handcrafted features [7], such as, log-Mel features [8], matrix factorization [9, 10], dictionary learning [3], and wavelet-based features [11]. The cepstral-based features, such as MFCC [12], GTCC [5], and TEO-based GTCC [13] are also used in the ESC task. Recently, deep learning-based classifiers are used for the ESC task [8, 14]. In particular, Convolutional Neural Network (CNN) has been observed to work better for this problem [8,14]. Since CNN classifier is useful for capturing the energy modulations across time and frequency-axis of audio spectrograms, it is well suited as classifier for ESC task [14].

The representation of a sound based on human auditory processing is of significant interest in developing features for the ESC task. The auditory models are based on mathematical modeling of auditory processing or psychophysical and physiological experiments. Mel filterbank is the state-of-the-art auditory-based features for the ESC task. Such handcrafted features rely on the simplified auditory models [15]. There are many approaches that are based on data-driven learning and/or optimization of parameters of auditory models. Data-driven learning or representation learning can be supervised (i.e., with label information) or unsupervised (where no such labels are available for each class). Recently, representation learning has gained a significant interest for feature learning in various signal processing areas including audio processing [16]. For the ESC task, various unsupervised representation learning architectures have been proposed, such as, dictionary learning [17]. In [18], authors have proposed end-to-end ESC system that can extract the features from the raw audio signals jointly with CNN. The work of [19], utilize a lot of unlabeled videos to learn the sound representation from the raw audio signals, which is trained by transferring a knowledge from vision into sound using CNN. Recently, we proposed a Convolutional Restricted Boltzmann Machine (ConvRBM) for filterbank learning [20], [21] directly from the raw speech signals of arbitrary lengths. The ConvRBM filterbank performs better in automatic speech recognition (ASR) task compared to Mel filterbank features [20], [21].

In this paper, we propose to exploit ConvRBM as a front-end for filterbank learning from the raw audio signals. Compared to our earlier works in [20], [21] and [22], here we have used Adam optimization [23] along with an annealed dropout technique [24]. Invariant representation is learned from the raw audio using ConvRBM and higher-level invariance is achieved using supervised CNN as a classifier. Experiments on the ESC-50 dataset shows that the proposed ConvRBM filterbank perform better than Mel filterbank. Score-level fusion of both filterbanks shows that ConvRBM filterbank contains significant complementary information.

2. ConvRBM for filterbank learning

ConvRBM is a probabilistic model with two layers, namely, a visible layer and a hidden layer. The input to the visible layer (denoted as $\mathbf{x}$) is an audio signal of length *n*-samples. Hidden layer (denoted as $\mathbf{h}$) consists of *K*-groups (i.e., the number of filters) with the filter length *m*-samples in each. Weights (also called as subband filters) are shared between visible and hidden units amongst all the locations in each group [21]. Denoting b_k as the hidden bias for the k^{th} group, the convolutional response for the k^{th} group is given as [21]:

$$\mathbf{I}_k = (\mathbf{x} * \tilde{\mathbf{W}}^k) + b_k, \quad (1)$$

http://dx.doi.org/10.21437/Interspeech.2017-831

where $\mathbf{x} = [x_1, x_2, ..., x_n]$ are samples of the audio signal, $\mathbf{W}^k = [w_1^k, w_2^k, ..., w_m^k]$ is a weight vector (i.e., k^{th} subband filter) and $\tilde{\mathbf{W}}$ denote a *flipped* array [21].

Dropout is a stochastic regularization technique that prevents a network from the overfitting by preventing co-adaptation of weights in the network. The term dropout refers to randomly dropping out neurons (i.e., assigning the zero value) in a network with probability p. In ConvRBM training, a dropout is applied before sampling the hidden units in both positive and negative phase of the contrastive divergence (CD) learning [25]. Applying a dropout to ConvRBM can be thought of as multiplying each unit in a k^{th} group with a binary mask (called as the *dropout mask*). The dropout mask for the k^{th} group is defined as random variables drawn from Bernoulli distribution, i.e., $\mathbf{m}_k = Bernoulli(p)$. With a noisy rectifier linear units (NReLU), the sampling equations for hidden and visible units ($\mathbf{x}_{recon}$ to reconstruct the audio signal) are given as:

$$\begin{aligned} \mathbf{h}^k &\sim max(0, \mathbf{m}_k \odot \mathbf{I}_k + N(0, \sigma(\mathbf{m}_k \odot \mathbf{I}_k))), \\ \mathbf{x}_{recon} &\sim \mathcal{N}\left(\sum_{k=1}^{K}(\mathbf{h}^k * \mathbf{W}^k) + c, 1\right), \end{aligned} \tag{2}$$

where c is a visible bias that is also shared, $N(0, \sigma(\mathbf{m}_k \odot \mathbf{I}_k))$ is a Gaussian noise with mean zero and sigmoid of $\sigma(\mathbf{m}_k \odot \mathbf{I}_k)$ as a variance. Here, $\odot$ indicates an elementwise multiplication. The block diagram of our proposed ConvRBM architecture is shown in Figure 1. In this paper, we have explored an annealed dropout (AD) training of ConvRBM that was proposed for supervised deep networks in [24]. In an annealed dropout, the dropout probability of the units in the network is gradually decreased over the training period. We have used the following annealing dropout schedule as suggested in [24]:

$$p[t] = max\left(0, 1 - \frac{t}{N}\right) p[0], \qquad t \in [0, N], \tag{3}$$

where $p[0]$ is the initial dropout rate at training iteration, $t = 0$. The dropout rate is decayed from $p[0]$ to a small value or zero for $t = N$ iterations. After N iterations, $p[t]$ is kept constant as 0 (i.e., no dropout). While calculating the relationship between the hidden and visible units, a deterministic ReLU (i.e., $max(0, \mathbf{I}_k)$) is used as an activation function [21]. Training of ConvRBM is based on single-step CD and parameters are updated using an Adam optimization technique [23]. It was shown that Adam optimization perform better than stochastic gradient-based methods due to use of first and second order moments of the gradient and bias correction terms [23]. After ConvRBM is trained, the feature extraction stages are as shown in Figure 2. Pooling is applied to reduce the representation of ConvRBM filter responses in the temporal-domain (i.e., from $K \times n$ samples to $K \times F$ frames) that is equivalent to the short-time averaging in the spectral features. Pooling parameters (i.e., window length and window shift) are similar as windowing parameters of Mel filterbank. We have experimented with both average and max-pooling. During the feature extraction stage, we used the 'same' length convolution and deterministic ReLU nonlinearity $max(0, \mathbf{I}_k)$ as an activation function. Logarithmic nonlinearity compresses the dynamic range of features.

3. Analysis of filterbank

3.1. Analysis of subband filters

For analysis of the subband filters, we first sort it according to the center frequencies (CFs) of the subband filters as done

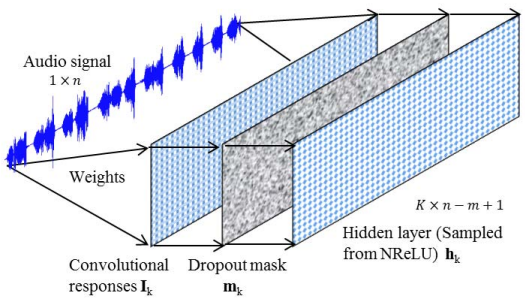

Figure 1: *Block diagram of the proposed ConvRBM with dropout mask. After [21], [22].*

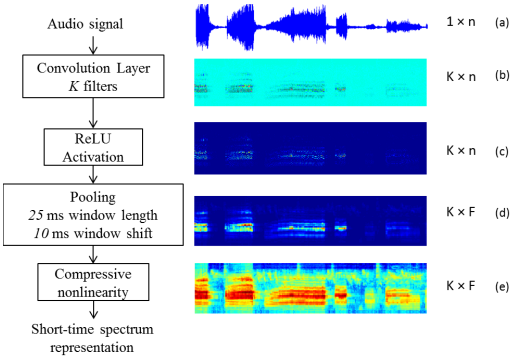

Figure 2: *Feature extraction using trained ConvRBM. (a) speech signal, (b) and (c) responses from the convolutional layer and ReLU nonlinearity, respectively, (d) representation after pooling, (e) logarithmic compression. After [21].*

in [21]. Weights of ConvRBM were initialized randomly and there is no constraint on filter shapes; interestingly, the model was able to learn meaningful representation from the audio signals. Weights of the model are called as impulse responses of the subband filters and its corresponding frequency responses are shown in Figure 3. It can be seen that many of the subband filters are Fourier-like basis functions that represent harmonic sounds such as animal vocalizations. Lower-frequency subband filters are gammatone-like basis functions. From Figure 3 (b), we can see that most of the subband filters are highly localized in frequency-domain. The frequency responses of the higher frequency subband filters are not localized that represents noisy-like sound classes such as rain, airplane, and thunderstorm. Similar insights have been discussed in [26–28] where the filterbanks were learned using an efficient coding principle. The work of [26, 27] analyze the filterbank on separate database of animal vocalizations and environmental sounds. Here, the ESC-50 database is a mixture of both of these categories. The subband filters are also different than we obtained when ConvRBM is trained on speech signals [20], [21]. This shows that transferring knowledge from the speech signals to the ESC tasks can also be helpful as done in [29].

3.2. Analysis of filterbank scale

In order to compare learned filterbank with the standard auditory filterbanks, we have shown a CF *vs.* subband filter index plot in Figure 4. We have also compared two ConvRBMs, the one that is trained using SGD, without dropout and the other

that is trained using AD and Adam optimization. Both ConvRBM filterbanks have a nonlinear relationship between CF and filter ordering similar to as other auditory filterbanks. This represents the placement of subband filters on the basilar membrane in the cochlea of the human ear. However, ConvRBM trained with AD and Adam optimization uses more number of subband filters in the frequency range 1.5-8 kHz (similar as observed in [30] when using Adam optimization). Since the ESC-50 dataset contains harmonic, transients and noise-like sound classes, the frequency-scale learned is also similar when ConvRBM is trained with the speech signals [21]. However, the subband filters are different compared to the speech [21].

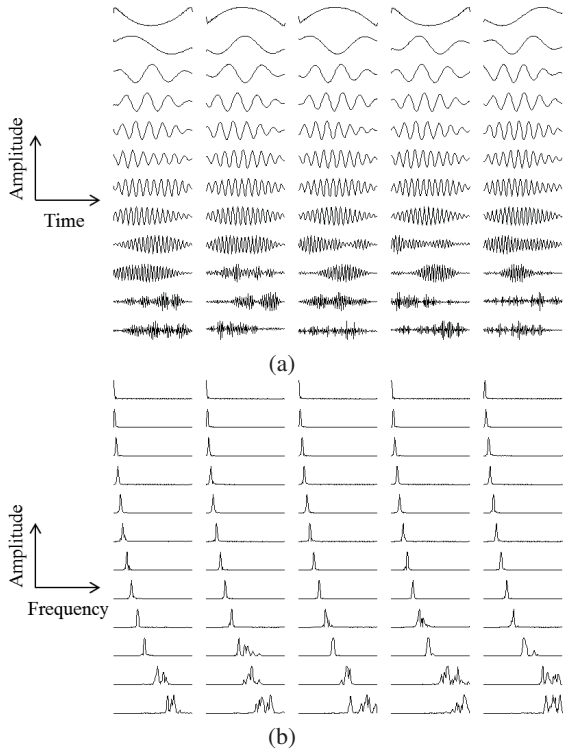

Figure 3: *Examples of the subband filters trained on the ESC-50 database: (a) subband filters in the time-domain (i.e., impulse responses), (b) subband filters in the frequency-domain.*

4. Experimental Setup

4.1. Dataset

In this paper, we have used the publicly available database ESC-50 [1] for the ESC task. The ESC-50 dataset consists of 2000 short (5-seconds) environmental recordings. These recordings are divided into 50 equally balanced classes. These 50 classes are divided into 5 major groups, namely, animals, natural soundscapes and water sounds, human non-speech sounds, interior/domestic sounds and exterior/urban noises. The files are prearranged in 5-folds for comparable cross-validation. Due to this reason, the results of the experiments can be directly compared to the baseline results and with the previous approaches.

4.2. Training of ConvRBM and Feature Extraction

We have trained ConvRBM with an annealed dropout using $p = 0.3$ and $p = 0.5$ that decayed to zero (i.e., $p = 0$) during training. The learning rate was chosen to be 0.001 and decayed according to the learning rate schedule as suggested in [23]. The

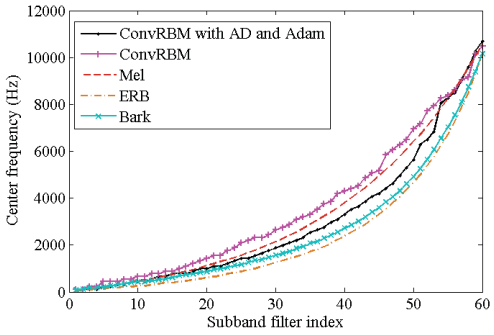

Figure 4: *Comparison of the filterbank learned using ConvRBM with the standard auditory filterbanks on the ESC-50 dataset. Here AD represents annealed dropout.*

moment parameters of Adam optimization chosen to be β_1=0.5 and β_2=0.999. We have trained the model with 60 number of subband filters (i.e., K) with different convolution window lengths (i.e., m=132, 176, 220 samples). The delta features were also appended resulting in two channels (60-dimensional each) for CNN classifier. The notation of ConvRBM filterbank is ConvRBM-BANK as used in [21].

4.3. CNN Classifier

We have used the CNN classifier with the architecture as proposed in [8] for the ESC task. However, we have not used data augmentation technique. Since the objective of this paper is to compare the performance of the front-end feature representation, we have not used the augmentation to analyze that how these features perform in all the classes. Before, feature extraction for CNN classifier, we first pre-processed the audio signal. All the audio files were downsampled to 22.05 kHz. To extract features, the audio files were divided into frames by using 25 ms Hamming window with 50 % overlap. Then, we applied silence removal algorithm. For silence removal, we first check for more than three consecutive silence frames (approximately 50 ms duration). If silence is present in more than three frames, then we remove the silence frames else we keep those frames. Simple energy thresholding algorithm was used to remove the silence regions. Mel Filterbank Energies (FBEs) are used as the baseline features. We have also used an auditory inspired Gammatone filterbank. The short segments of 41 frames were used as the input to the CNN. The segments were extracted with 50 % overlap from the audio files.

Figure 5 shows the details of each layer in the CNN architecture that we have used in ESC task. The network was implemented using Keras [31] with theano back-end on NVIDIA Titan-X GPU. A mini-batch implementation with 200 batch size was used to train the network. Network parameters were similar as used in [8]. The learning rate of 0.002, L^2 regularization with the coefficient 0.001 and network was trained for 300 epochs. At the testing time, the class of the test audio files were decided using the probability prediction scheme [8]. We have also done score-level fusion of different feature sets as used in [32].

5. Experimental Results

To evaluate the performance of the proposed learned filterbank with different tuning parameters, 5-fold cross-validation was performed on the ESC-50 database as shown in Table 1. Experiments were conducted with different filter lengths in Con-

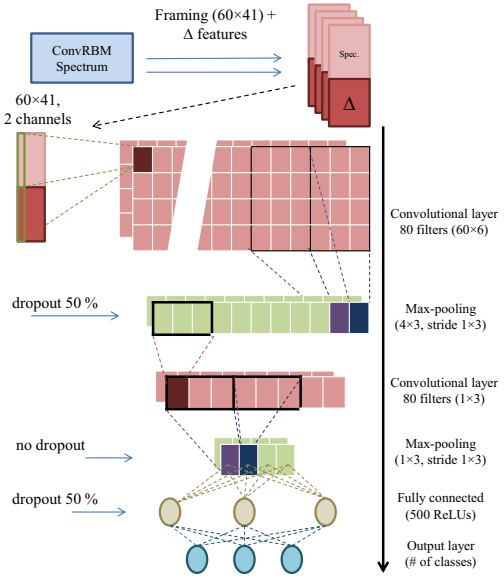

Figure 5: *CNN architecture for ESC task. After [8], [13].*

vRBM training. We observed that 132 samples (i.e., 6 ms) filter length gave better performance than other filter lengths. In all the cases, Adam optimization performed better than stochastic gradient descent (SGD) in the ConvRBM training. From Table 1, it can be seen that max-pooling with dropout significantly, works better than average-pooling in ConvRBM for the ESC task. This observation is different than what we observed in the speech recognition task, where average-pooling in ConvRBM-BANK performed well [20]. We also perform the experiments with different dropout probabilities (p) for filterbank learning. The dropout with probability 0.5 performed better than 0.3 with same configurations of ConvRBM. Hence, we have selected ConvRBM with 132 filter length, 0.5 dropout probability, Adam optimization, and max-pooling for rest of the experiments.

Table 1: *% Classification accuracy using ConvRBM-BANK features with different tunning parameters. Here, m is ConvRBM filter length and p is the annealed dropout probability.*

m	Optimizer	p	Pooling	Accuracy (%).
132	SGD	-	average	59.85
132	SGD	-	max	76.95
132	ADAM	-	average	66.55
132	ADAM	-	max	76.15
132	ADAM	0.3	average	67.45
132	ADAM	0.3	max	78.15
132	ADAM	0.5	max	**78.45**
176	ADAM	0.3	average	57.40
176	ADAM	0.3	max	74.90
176	ADAM	0.5	max	75.30
220	ADAM	0.5	max	73.25

We compare the performance of ConvRBM-BANK with FBEs and Gammatone Spectral Coefficients (GTSC). The overall results of the proposed method and baseline feature sets are summarized in Table 2 with CNN classifier. ConvRBM-BANK perform significantly better than FBEs with an absolute improvement of 10.65 % in classification accuracy. Gammatone filterbank is inspired from the auditory physiology [33] whereas we have learned an auditory-like filterbank from the raw audio signals with randomly initialized weights. Interestingly, it gives a comparable classification accuracy with GTSC (79.10 % *vs.* 78.45 %). The score-level fusion of ConvRBM-BANK with FBEs and GTSC improves the performance. However, the score-level fusion of ConvRBM-BANK (78.45 %) and FBEs (67.80 %) achieved the best accuracy of 86.50 % in this paper. This shows that the proposed ConvRBM-BANK contains highly complementary information over the Mel filterbank, which is helpful in the ESC task.

Table 2: *% Classification accuracy of ESC-50 dataset with different feature sets and its score-level fusion. The $\oplus$ sign and α indicate score-level fusion and fusion factor, respectively.*

feature sets	α	Accuracy (%)
FBEs	-	67.80
GTSC	-	79.10
ConvRBM-BANK	-	78.45
FBEs $\oplus$ ConvRBM-BANK	0.5	**86.50**
GTSC $\oplus$ ConvRBM-BANK	0.5	83.00

Our proposed work is also compared with the other studies in literature in Table 3. ConvRBM-BANK performs significantly better than CNN with FBEs [8], [18]. In [18], filterbank is learned from the raw audio signal using CNN as an end-to-end system. The EnvNET [18] performs better when combining with log Mel CNN. However, our proposed ConvRBM-BANK outperform EnvNET [18] even without the system combination. this shows the significance of unsupervised generative training using ConvRBM.

Table 3: *Comparison of classification accuracy of ESC-50 dataset in the literature. The $\otimes$ sign indicated system combination before soft-max.*

feature sets	Accuracy (%)
ConvRBM-BANK (proposed)	**78.45**
FBEs $\oplus$ ConvRBM-BANK (proposed)	**86.50**
Piczak FBEs-CNN [8]	64.50
Human [1]	81.30
EnvNET [18]	64.00
logmel-CNN [18]	66.5
logmel-CNN $\otimes$ EnvNet [18]	71.00

6. Summary and Conclusions

In this study, we propose the unsupervised filterbank learning using ConvRBM for ESC task. The learned frequency scale is nonlinear similar as other standard auditory scales. We achieved an absolute improvement of 18.70 % in the classification accuracy over the state-of-the-art FBEs features with a score-level fusion of ConvRBM-BANK and FBEs using CNN classifier. Furthermore, we analyzed the filters learned with our system, and showed that our features are capable of capturing invariant representation from the raw audio using generative model. Score-level fusion of ConvRBM-BANK with FBEs. shows that both filterbank contains complementary information. Our future works includes using Unsupervised Deep Auditory Model (UDAM) using a stack of ConvRBM [34].

7. Acknowledgements

Authors would like to thank Dept. of Electronics and Information Technology (DeitY), Govt. of India, for sponsored projects and authorities of DA-IICT for providing infrastructure. They also thank NVIDIA for providing hardware grant of Titan-X GPU for research purposes.

8. References

[1] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proc. of the* 23^{rd} *Int. Conf. on Multimedia*, Brisbane, Australia, 2015, pp. 1015–1018.

[2] D. P. Ellis, "Prediction-driven computational auditory scene analysis," Ph.D. dissertation, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology (MIT), USA, 1996.

[3] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, and M. Vento, "Audio surveillance of roads: a system for detecting anomalous sounds," *IEEE Trans. on Intell. Transp. Syst.*, vol. 17, no. 1, pp. 279–288, 2016.

[4] 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 Trans. on Audio, Speech, and Lang. Process.*, vol. 15, no. 8, pp. 2249–2256, 2007.

[5] M. Vacher, J.-F. Serignat, and S. Chaillol, "Sound classification in a smart room environment: an approach using GMM and HMM methods," in *The* 4^{th} *IEEE Conference on Speech Technology and Human-Computer Dialogue (SpeD), Publishing House of the Romanian Academy (Bucharest)*, vol. 1, 2007, pp. 135–146.

[6] L. Ballan, A. Bazzica, M. Bertini, A. Del Bimbo, and G. Serra, "Deep networks for audio event classification in soccer videos," in *Int. Conf. on Multimedia and Expo (ICME)*. New York, USA: IEEE, 2009, pp. 474–477.

[7] S. Chachada and C.-C. J. Kuo, "Environmental sound recognition: A survey," *APSIPA Transactions on Signal and Information Processing*, vol. 3, no. 14, pp. 1–15, 2014.

[8] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in 25^{th} *Int. Workshop on Machine Learning for Signal Processing (MLSP)*, Boston, MA, USA, 2015, pp. 1–6.

[9] V. Bisot, R. Serizel, S. Essid, and G. Richard, "Acoustic scene classification with matrix factorization for unsupervised feature learning," in *Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, 2016, pp. 6445–6449.

[10] B. Ghoraani and S. Krishnan, "Time–frequency matrix feature extraction and classification of environmental audio signals," *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 19, no. 7, pp. 2197–2209, 2011.

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

[12] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proc. of the* 22^{nd} *Int. Conf. on Multimedia*, Orlando, Florida, 2014, pp. 1041–1044.

[13] D. M. Agrawal, H. B. Sailor, M. H. Soni, and H. A. Patil, "Novel TEO-based gammatone features for environmental sound classification," in *submitted in European Signal Processing Conf. (EUSIPCO)*, Kos island, Greece, August 28 – 2 September 2017.

[14] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Process. Lett.*, vol. 24, no. 3, pp. 279–283, March 2017.

[15] R. M. Stern and N. Morgan, "Features based on auditory physiology and perception," in *Techniques for Noise Robustness in Automatic Speech Recognition*. T. Virtanen, B. Raj, and R. Singh, (Eds.) John Wiley and Sons, Ltd, New York, NY, USA, 2012, pp. 193–227.

[16] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. on Pattern Anal. and Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[17] J. Salamon and J. P. Bello, "Unsupervised feature learning for urban sound classification," in *Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, 2015, pp. 171–175.

[18] Y. Tokozume and T. Harada, "Learning environmental sound with end-to-end convolutional neural network," in *IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, New Orleans, USA, 2017, pp. 2721–2725.

[19] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," in *Advances in Neural Information Processing Systems (NIPS)*, 2016, pp. 892–900.

[20] H. B. Sailor and H. A. Patil, "Filterbank learning using convolutional restricted Boltzmann machine for speech recognition," in *Int. Conf. on Acoust., Speech and Signal Process. (ICASSP) 2016*, Shanghai, China, March 2016, pp. 5895–5899.

[21] H. B. Sailor and H. A. Patil, "Novel unsupervised auditory filterbank learning using convolutional RBM for speech recognition," *IEEE/ACM Trans. on Audio, Speech and Lang. Process.*, vol. 24, no. 12, pp. 2341–2353, Dec. 2016.

[22] H. B. Sailor and H. A. Patil, "Auditory feature representation using convolutional restricted Boltzmann machine and Teager energy operator for speech recognition," *Journal of Acoustical Society of America Express Letters (JASA-EL)*, vol. 141, no. 6, pp. EL500–EL506, June. 2017.

[23] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR), San Diego, USA*, 2015, pp. 1–15.

[24] S. J. Rennie, V. Goel, and S. Thomas, "Annealed dropout training of deep networks," in *IEEE Spoken Language Technology Workshop (SLT), South Lake Tahoe, California and Nevada*, 2014, pp. 159–164.

[25] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[26] M. S. Lewicki, "Efficient coding of natural sounds," *Nature Neuroscience*, vol. 5, no. 4, pp. 356–363, 2002.

[27] E. C. Smith and M. S. Lewicki, "Efficient auditory coding," *Nature*, vol. 439, no. 7079, pp. 978–982, 2006.

[28] R. G. Erra and J. Gervain, "The efficient coding of speech: Cross-linguistic differences," *PLOS ONE*, vol. 11, no. 2, pp. 1–18, 2016.

[29] H. Lim, M. J. Kim, and H. Kim, "Cross-acoustic transfer learning for sound event classification," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China*, March,2016, pp. 2504–2508.

[30] H. Seki, K. Yamamoto, and S. Nakagawa, "A deep neural network integrated with filterbank learning for speech recognition," in *Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, USA, 2017, pp. 5480–5484.

[31] F. Chollet, "Keras," https://github.com/fchollet/keras { Last Accessed on 26^{th} February, 2017}.

[32] J. Li, W. Dai, F. Metze, S. Qu, and S. Das, "A comparision of deep learning methods for environmental sound detection," in *IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, New Orleans, USA, 2017, pp. 126–130.

[33] B. R. Glasberg and B. C. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, no. 1, pp. 103–138, 1990.

[34] H. B. Sailor and H. A. Patil, "Unsupervised deep auditory model using stack of convolutional RBMs for speech recognition," in *INTERSPEECH*, San Francisco, California, USA, September 2016, pp. 3379–3383.