HIERARCHICAL LEARNING FOR DNN-BASED ACOUSTIC SCENE CLASSIFICATION

Yong Xu, Qiang Huang, Wenwu Wang, Mark D. Plumbley

Centre for Vision, Speech and Signal Processing, University of Surrey, UK
{yx0001, q.huang, w.wang, m.plumbley}@surrey.ac.uk

ABSTRACT

In this paper, we present a deep neural network (DNN)-based acoustic scene classification framework. Two hierarchical learning methods are proposed to improve the DNN baseline performance by incorporating the hierarchical taxonomy information of environmental sounds. Firstly, the parameters of the DNN are initialized by the proposed hierarchical pre-training. Multi-level objective function is then adopted to add more constraint on the cross-entropy based loss function. A series of experiments were conducted on the Task1 of the Detection and Classification of Acoustic Scenes and Events (DCASE) 2016 challenge. The final DNN-based system achieved a 22.9% relative improvement on average scene classification error as compared with the Gaussian Mixture Model (GMM)-based benchmark system across four standard folds.

Index Terms— Acoustic scene classification, deep neural network, hierarchical pre-training, multi-level objective function

1. INTRODUCTION

In recent years, much research effort has been attracted for making sense of everyday or environmental sounds. It focuses on how to convert audio (non-speech and non-music) recordings into understandable and actionable information: specifically how to allow people to search, browse and interact with sounds. Some specific tasks were investigated in recent years, including acoustic scene classification (ASC) [1], sound event detection (SED) [2, 3] and domestic audio tagging. ASC aims to associate a semantic label to an audio segment that identifies the sound environment where it has been produced [1]. The goal of SED is to detect the sound events that are present within an audio signal, estimate their start and end times, and give a class label to each of the events. For audio tagging, there is no information about sound event onset or offset, only labels. This paper will focus on the ASC task.

The ASC problem was first proposed by Sawhney and Maes [4]. Recently, more related work was conducted during the IEEE AASP Challenge: Detection and Classification of Acoustic Scenes and Events [5, 6, 7]. Mel Frequency Cepstrum Coefficients (MFCCs) were used as the audio feature by most of the submitted systems. GMMs, Support Vector Machines (SVMs) or hidden Markov models (HMMs) were commonly used classifier [6, 8, 9]. Other methods, such as non-negative matrix factorization (NMF) approaches can also be used to extract an intermediate representation prior to classification [10].

Recently, deep learning methods have obtained great successes in speech, image and video fields [11, 12, 13, 14] since Hinton *et al* showed the insights in using a greedy layer-wise unsupervised learning procedure to train a deep model in 2006 [15]. Deep learning methods were also investigated for acoustic scene classification tasks in [16, 17, 18]. In [16], a series of experimental investigations on the DNN structure, including the number of hidden layers and input frame expansion, were presented. It also demonstrated that DNN can yield better results than GMM and SVM. Convolutional neural networks (CNNs) which are the variant of DNNs have been also adopted for environmental sound classification in [17].

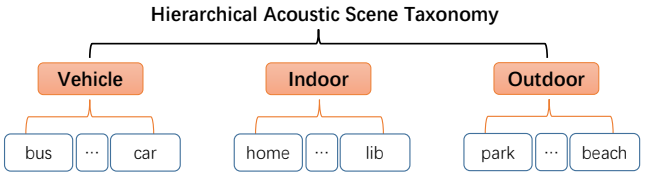

Figure 1: Example of a hierarchical acoustic scene taxonomy.

There is also a research about the taxonomy of the environmental sounds [19, 20]. The taxonomy of environmental sounds indicates that hierarchical categories information exists in sound classes. For example, environmental sounds can be coarsely classified into *indoor*, *outdoor* and *vehicle* in Fig. 1, and these are the high-level scene classes. Meanwhile, corresponding branches denote the low-level scene classes. In this paper, we propose a hierarchical learning method incorporating the acoustic scene taxonomy information for ASC in a DNN-based framework. Two approaches are presented to utilize the acoustic scene taxonomy information. Firstly, a high-level DNN is discriminatively trained to predict three high-level classes, namely *vehicle*, *indoor* and *outdoor*. Then the trained DNN is used to initialize the low-level DNN except for the top classification layer to learn the more difficult low-level scene classes, namely *bus*, *home*, *park*, etc. This learning process is named as **hierarchical pre-training**, which follows the common "easiest thing first hardest second" learning experience of human [21]. Hierarchical pre-training is a supervised process which is different from the common Restricted Boltzmann machine (RBM) based unsupervised pre-training [15]. The second idea is based on a proposed **multi-level objective function**, which means the DNN not only predicts the target low-level scene classes, but also predicts the three high-level scene classes as the auxiliary task. It is actually a multi-task learning [22] which has been demonstrated to be effective in recent DNN-based speech enhancement [23] and speech recognition [24].

The rest of the paper is organized as follows. In Section 2 we describe the DNN-based acoustic scene classification baseline system. Hierarchical pre-training and multi-level objective function are presented in Section 3. In Section 4, we present a series of experiments to assess the system performance. Finally we summarize our findings in Section 5.

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) of the UK under the grant EP/N014111/1.

2. DNN-BASED ACOUSTIC SCENE CLASSIFICATION

DNN is a non-linear multi-layer model with powerful capability to extract robust feature related to a specific classification [13] or regression [12] task. ASC is a typical classification problem where a specific scene label should be assigned to an audio segment.

2.1. DNN baseline

A basic DNN consists of a number of different layers stacked together in a deep architecture: an input layer, several hidden layers and an output layer. More precisely, when the goal is to classify an audio feature $\mathbf{x}$ among N acoustic scene classes, a DNN estimates the posteriors p_j, $j \in \{1, ..., J\}$, of each class given the input feature $\mathbf{x}$. The input $\mathbf{x}$ which is fed into DNN represents the contextual audio feature, such as 11 consecutive frames centered at the current frame. Such contextual information was shown to improve the prediction performance in DNN-based speech enhancement or speech recognition [12, 13]. The activation functions used in each hidden unit of the hidden layers are non-linear sigmoid or Rectified Linear Units (ReLUs) [25] function. The ReLU, which is adopted in this work, has several advantages over the sigmoid: faster computation and more efficient gradient propagation and it is defined below:

$$f(y) = \max(0, y) \tag{1}$$

where y is the output of the hidden unit before activated by ReLU. The output is computed via the softmax nonlinearity to force the target label to have the maximum posterior while competing with other non-targets. The objective is to minimize the cross entropy between the predictions of DNN $\mathbf{p} = [p_1, ..., p_J]^T$ and the target probabilities $\mathbf{d} = [d_1, ..., d_J]^T$. The loss function is defined as following:

$$L = -\sum_{j=1}^{J} d_j \log(p_j) \tag{2}$$

The classical back-propagation (BP) algorithm [13] can be used to update the weights and bias of DNN based on the calculated error.

2.2. Dropout for the over-fitting problem

Deep learning architectures have a natural tendency to over-fitting especially when there is a little training data. Dropout is a simple but effective way to alleviate this problem [25]. In each training iteration, the feature value of every input unit and the activation of every hidden unit are randomly removed with a predefined probability (e.g., ρ). These random perturbations effectively prevent the DNN from learning spurious dependencies. At the decoding stage, the DNN discounts all of the weights involved in the dropout training by $(1 - \rho)$, regarded as a model averaging process [26].

For the acoustic scene classification task, the testing audio segment could be totally different from the used training audio segments due to the presence of background noise. Thus Dropout should be adopted to improve its robustness to generalize to variants of testing segments.

2.3. Decision maker based on average confidence

ASC aims to assign a single semantic label to an audio segment. Majority voting is often used to make a global decision across all of the single audio frames in this segment [1]. Here we proposed to use a more precise decision making scheme:

$$\hat{c} = \max_j (\frac{1}{T} \sum_{t=1}^{T} p_{t,j}) \tag{3}$$

where T is the total number of frames belonging to the current testing audio segment, $\hat{c}$ denotes the predicted global scene label based on the average confidence across the whole frames, and $p_{t,j}$ represents the estimated DNN posterior at the t-th frame for class j.

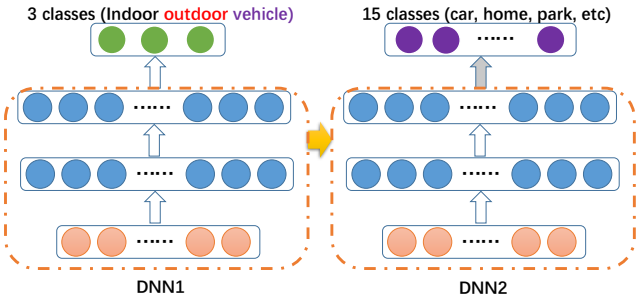

Figure 2: Proposed hierarchical pre-training.

3. PROPOSED HIERARCHAL LEARNING FOR ASC

In this section, two novel methods: hierarchical pre-training and multi-level objective function incorporating the scene taxonomy information for DNN-based ASC are presented.

3.1. Hierarchical pre-training

Pre-training is crucial to avoid the algorithm getting stuck in a local optimum for training a deep model especially when the training data is not sufficient. The two most notable pre-training methods are the RBMs [15] based and stacked auto-encoders [27] based greedy layer-wise algorithms. They are both unsupervised while the proposed hierarchical pre-training is supervised. In the acoustic scene taxonomy research [20], the acoustic scenes are naturally categorized into hierarchical classes. Fig. 2 shows how the proposed DNN-based method incorporates the hierarchical taxonomy information. The hierarchical pre-training consists of two steps. Firstly, the DNN1 was trained to predict the three high-level acoustic scene classes, namely indoor, outdoor and vehicle. DNN2 was then trained to estimate the posterior of the 15 target low-level acoustic scene classes with the initialized weights from DNN1. Note that the classification layer of DNN2 was initialized with random weights because this top layer is different from DNN1. It is easier for DNN to learn the three coarsely classified high-level classes than the 15 target classes. However, the DNN2 can be better fine-tuned based on DNN1. It follows the common sense of human learning process: easiest things first hardiest second. The experience of learning easier things could benefit the learning for harder things.

3.2. Multi-level objective function

Multi-task learning [22] is successfully adopted in DNN-based speech enhancement [23] and DNN-based speech recognition [24]. The auxiliary target was demonstrated beneficial for the primary target. Inspired by this, multi-level objective function is proposed to

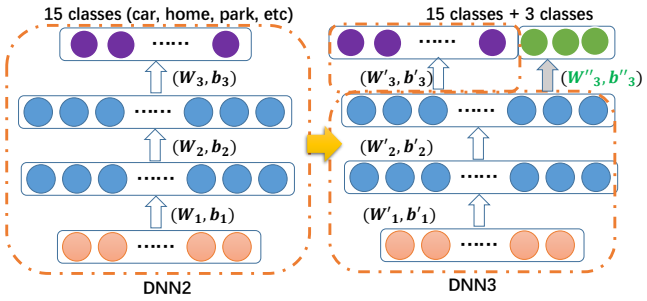

Figure 3: Proposed multi-level objective function based on the well trained DNN2. **W** and **b** denote the weights and bias, respectively.

incorporate the hierarchical acoustic scene taxonomy information into the integrated objective function.

Fig. 3 shows the proposed multi-level objective function based on the well trained DNN2. The main difference between DNN2 and DNN3 is that an additional softmax layer is designed to describe the three high-level classes (indoor, outdoor and vehicle). **W** and **b** denote the weights and bias, respectively. $\mathbf{W}'$ and $\mathbf{b}'$ of DNN3 were both initialized by **W** and **b** of DNN2. The additional softmax layer $(\mathbf{W}'', \mathbf{b}'')$ was randomly initialized. With this modification, the cross entropy based loss function should be changed to contain two parts as follows:

$$L_{1:N} = -\alpha \sum_{t=1}^{N} \sum_{j=1}^{J} d_{t,j} \log(p_{t,j}) - (1-\alpha) \sum_{t=1}^{N} \sum_{k=1}^{K} d_{t,k} \log(p_{t,k}) \tag{4}$$

where N is the mini-batch size, $d_{t,j}$ denotes the target probability at the t-th frame for the j-th low-level scene class, $d_{t,k}$ denotes the DNN predicted posterior at the t-th frame for the k-th high-level scene class. α is the weighting factor to tune the error contribution from the above two parts. J and K represent the 15 low-level classes and the three high-level classes, respectively.

Hence, the proposed multi-level objective function is another idea to utilize the hierarchical scene taxonomy information besides the proposed pre-training in Sec. 3.1.

4. EXPERIMENTAL SETUP AND RESULTS

The proposed methods were evaluated on the task1 of DCASE 2016 challenge. Task1 is about acoustic scene classification aiming to classify a test recording into one of the predefined classes that characterizes the environment where it was recorded. The dataset consists of recordings from various acoustic scenes, all having distinct recording locations. For each recording location, 3-5 minutes long audio recording was captured. The original recordings were then split into 30-second segments for the challenge. There are 15 [1] acoustic scenes for this task. Three high-level scene classes are also indicated.

[1] 15 scene classes in DCASE2016 task1, C1: Lakeside beach (outdoor); C2: Bus, traveling by bus in the city (vehicle); C3: Cafe / Restaurant, small cafe/restaurant (indoor); C4: Car, driving or traveling as a passenger (vehicle); C5: City center (outdoor); C6: Forest path (outdoor); C7: Grocery store, medium size grocery store (indoor); C8: Home (indoor); C9: Library (indoor); C10: Metro station (indoor); C11: Office, multiple persons, typical work day (indoor); C12: Urban park (outdoor); C13: Residential area (outdoor); C14: Train (traveling, vehicle); C15: Tram (traveling, vehicle).

For all of the acoustic scenes, each of the recordings was captured in a different location: different streets, different parks and different homes. Recordings were made using a Soundman OKM II Klassik/studio A3, electret binaural microphone and a Roland Edirol R-09 wave recorder using 44.1 kHz sampling rate and 24 bit resolution. The recordings are down-sampled into 16 kHz in this paper. The microphones are specifically made to look like headphones, being worn in the ears. As an effect of this, the recorded audio is very similar to the sound that reaches the human auditory system of the person wearing the equipment.

The TUT Acoustic scenes 2016 dataset consists of two subsets: a development dataset and an evaluation dataset. In this paper, only the development dataset is used for evaluation because the labels of the evaluation dataset have not been released. The development dataset contains 1170 segments in total with 30 seconds length for each. A cross-validation setup with four folds is provided for the development dataset. The scoring of acoustic scene classification will be based on classification accuracy. Each segment is considered as an independent test sample. Confusion matrix among various acoustic scene classes would also be presented.

The official baseline system is based on the MFCC acoustic features and GMM classifier. The system learns one acoustic model per acoustic scene class, and performs the classification with maximum likelihood classification scheme. The length of each frame is 40 ms with 50% hop size. The acoustic features include 20-dimension MFCC static coefficients (0th coefficient included), delta coefficients and acceleration coefficients.

For the DNN method, 11 frames of Mel-filter bank features with 40 channels were used as the input. Two hidden layers with 500 ReLU hidden units for each layer were adopted for DNN. The learning rate was 0.005. The momentum was set to 0.9. Weight cost was not used. The dropout value for the input layer was 0.1 while 0.3 for hidden layers. α in Eq. 4 was 0.6. NVIDIA-Tesla-M2090 GPU was used to train the DNN models. The output unit number for DNN1, DNN2 and DNN3 were 3, 15 and 18, respectively.

4.1. Evaluations for the proposed methods

System	Fold 1	Fold 2	Fold 3	Fold 4	Average
DNN1	93%	90%	89%	91%	90%

Table 1: Frame-wise accuracy (%) for three high-level scene classes on different cross-validation (CV) folds using DNN1. All of the related CV audio segments were excluded from the training samples.

As shown in Fig. 2, DNN1 should be trained as the pre-trained model for DNN2. Table 1 gives the frame-wise accuracy (%) for the three high-level scene classes on four cross-validation (CV) folds using DNN1. All of the related CV audio segments were excluded from the training sampels. An average of frame-level 90% accuracy can be obtained for the classification of three high-level acoustic scene classes, namely indoor, outdoor and vehicle. Therefore, DNN can easily deal with this learning. It would offer a good starting optimization point for the fine-tuning of DNN2 with the initialized weights from DNN1.

Then the DNN2 was trained to predict the 15 target acoustic scene classes based on the well trained DNN1. Table 2 presented the overall comparison of acoustic scene accuracy (%) on different CV folds among the DCASE2016 official GMM baseline, the DNN

Systems	Fold 1 ACC (%)	Fold 2 ACC (%)	Fold 3 ACC (%)	Fold 4 ACC (%)	Average ACC (%)
GMM-baseline	72.50	66.80	70.10	75.70	71.28
DNN-baseline (+dropout)	79.62	67.24	75.84	78.08	75.19
DNN2 (+hierarchical pre-training)	80.69	71.72	77.52	78.77	77.17
DNN3 (++multi-level objective func)	**81.38**	**72.41**	**77.85**	**79.79**	**77.86**

Table 2: The overall comparison of acoustic scene accuracy (%) on different cross-validation (CV) folds among the DCASE2016 official GMM baseline, the DNN baseline improved by dropout, the DNN2 with the hierarchical pre-training based on DNN baseline, and the DNN3 optimized by the proposed multi-level objective function based on DNN2. All of the related CV audio segments were excluded from the training samples.

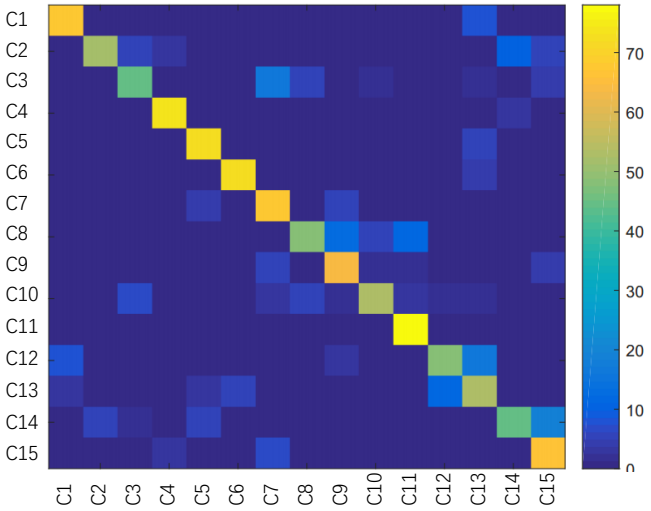

Figure 4: The confusion matrix among 15 acoustic scene classes by comparing the DNN predicted class with the target class on all of the four folds. $Cx, x \in \{1, ..., 15\}$ represent the indices of the 15 classes which were defined in footnote 1.

baseline improved by dropout, the DNN2 with the hierarchical pre-training based on DNN baseline, and the DNN3 optimized by the proposed multi-level objective function based on DNN2. All of the related CV audio segments were excluded from the training samples. The DNN baseline improved by dropout outperformed the provided GMM-MFCC baseline at all folds. The acoustic scene accuracy was increased from 71.28% to 75.19% on average. It also should be noted that the DNN is just slightly better than GMM on Fold 2 where the performance is the lowest. However, with the proposed hierarchical pre-training, its accuracy was significantly improved from 67.24% to 71.72% on Fold 2. Therefore, it demonstrates that the proposed hierarchical pre-training is important in challenging scene classification situations. DNN2 obtains an 8% relative improvement compared with the DNN baseline from 75.19% to 77.17%.

The DNN3 optimized by the proposed multi-level objective function gives further improvement. The final average acoustic scene accuracy was increased to 77.86%. It indicates that the additional constraint imposed in Eq. 4 can benefit the primary target. Finally, the proposed DNN system offers 22.9% and 10.8% relative improvements compared with the GMM-MFCC baseline and the DNN baseline, respectively. Note that the DNN baseline is a strong system since it is optimized by dropout training.

4.2. Further discussions

Fig. 4 presents the confusion matrix among 15 acoustic scene classes by comparing the DNN predicted class with the target class on all of the four folds. $Cx, x \in \{1, ..., 15\}$ represents the indices of the 15 classes which were defined in footnote 1. Observed from this confusion matrix, one phenomenon is that *park* ($C12$) easily gets confused by *residential area* ($C13$), and vice versa. It could be explained that similar acoustic events happened in both acoustic environments, like the *bird singing* and *car passing-by*. Another interesting case is that *grocery store* ($C7$) tends to be mis-recognized as *restaurant* ($C3$) due to the common human speech events. This might suggest that the presence of common human speech needs to be reduced in the audio segments before the acoustic scene classification is conducted. *Tram* ($C15$) also has the tendency to be incorrectly identified as *Train* ($C14$).

In this acoustic scene classification task, the characteristics of the test audio segment can be very different from the used training audio segments because of the randomly happening acoustic events. The adopted Dropout method can alleviate this problem. However, more robust feature learning methods should be developed to extract the specific acoustic characteristics of the certain acoustic environments.

5. CONCLUSIONS

In this paper, we have studied how to incorporate the taxonomy information into deep learning framework, and developed two DNN-based hierarchical learning methods for the acoustic scene classification task. The first novel method, called hierarchical pre-training which is a supervised learning process, can help the second DNN to get a better initialized weights based on the learning experience from the three high-level coarsely classified classes. It can achieve an 8% relative improvement compared with the DNN baseline improved by Dropout. The second proposed approach was the multi-level objective function which was inspired by the multi-task learning. It can help improve the prediction accuracy of the primary 15 target low-level classes by adding additional estimation of the three high-level classes in the DNN output, which was also regarded as imposing more constraint on the cross-entropy loss function. This idea can further improve the scene classification performance. Finally, the proposed DNN system has obtained 22.9% and 10.8% relative improvements over the GMM-MFCC baseline and the well trained DNN baseline, respectively.

In our future work, the hierarchical learning will also be investigated in more complicated network structure, such as the CNN and the long-short term memory (LSTM) model based framework, for the acoustic scene classification task.

6. REFERENCES

[1] D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley, "Acoustic scene classification: Classifying environments from the sounds they produce," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 16–34, 2015.

[2] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, "Acoustic event detection in real life recordings," in *18th European Signal Processing Conference*. IEEE, 2010, pp. 1267–1271.

[3] X. Zhuang, X. Zhou, M. A. Hasegawa-Johnson, and T. S. Huang, "Real-world acoustic event detection," *Pattern Recognition Letters*, vol. 31, no. 12, pp. 1543–1551, 2010.

[4] N. Sawhney and P. Maes, "Situational awareness from environmental sounds," *URL: http://web. media. mit. edu/˜ nitin/papers/Env_Snds/EnvSnds. html*, 1997.

[5] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange, and M. Plumbley, "Ieee aasp challenge: Detection and classification of acoustic scenes and events," Technical Report, Queen Mary University of London, Tech. Rep., 2013.

[6] 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)*. IEEE, 2013, pp. 1–4.

[7] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1733–1746, 2015.

[8] J. T. Geiger, B. Schuller, and G. Rigoll, "Large-scale audio feature extraction and svm for acoustic scene classification," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2013, pp. 1–4.

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

[10] V. Bisot, R. Serizel, S. Essid, *et al.*, "Acoustic scene classification with matrix factorization for unsupervised feature learning," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 6445–6449.

[11] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, 2014.

[12] ——, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2015.

[13] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[15] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[16] Y. Petetin, C. Laroche, and A. Mayoue, "Deep neural networks for audio scene recognition," in *2015 23rd European Signal Processing Conference (EUSIPCO)*. IEEE, 2015, pp. 125–129.

[17] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2015, pp. 1–6.

[18] M. Ravanelli, B. Elizalde, K. Ni, and G. Friedland, "Audio concept classification with hierarchical deep neural networks," in *2014 Proceedings of the 22nd European Signal Processing Conference (EUSIPCO)*. IEEE, 2014, pp. 606–610.

[19] M. Niessen, C. Cance, and D. Dubois, "Categories for soundscape: toward a hybrid classification," in *Inter-Noise and Noise-Con Congress and Conference Proceedings*, vol. 2010, no. 5. Institute of Noise Control Engineering, 2010, pp. 5816–5829.

[20] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proceedings of the ACM International Conference on Multimedia*. ACM, 2014, pp. 1041–1044.

[21] Y. J. Lee and K. Grauman, "Learning the easy things first: Self-paced visual category discovery," in *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2011, pp. 1721–1728.

[22] R. Camana, "Multitask learning: A knowledge-based source of inductive bias," in *Proceedings of the Tenth International Conference on Machine Learning*, 1993, pp. 41–48.

[23] Y. Xu, J. Du, Z. Huang, L.-R. Dai, and C.-H. Lee, "Multi-objective learning and mask-based post-processing for deep neural network based speech enhancement," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[24] Z. Huang, J. Li, S. M. Siniscalchi, I.-F. Chen, J. Wu, and C.-H. Lee, "Rapid adaptation for deep neural networks through multi-task learning," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[25] G. E. Dahl, T. N. Sainath, and G. E. Hinton, "Improving deep neural networks for lvcsr using rectified linear units and dropout," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 8609–8613.

[26] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.

[27] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *The Journal of Machine Learning Research*, vol. 11, pp. 3371–3408, 2010.