Acoustic Scene Classification Using A Deeper Training Method for Convolution Neural Networks

Tan Doan*, Hung Nguyen†, Dat Thanh Ngo‡, Lam Pham§ and Ha Hoang Kha¶
Faculty of Electrical and Electronics Engineering, Ho Chi Minh City University of Technology, VNU-HCM, Vietnam
Email: *tandoan.hcmut@gmail.com, †nguyenhonghung96@gmail.com@gmail.com, ‡datt.ngo.hcmut@gmail.com,
§lamd.pham@hcmut.edu.vn, ¶hhkha@hcmut.edu.vn

***Abstract*—In this paper, we present a deep learning framework applied for acoustic scene classification (ASC) recognizing the environmental sounds. Since an audio scene related to a given location potentially contains numerous sound events, only few of these events supply helpful information on the scene, which makes the acoustic scene classification task become a very complex problem. To confront this challenge, we suggest a novel architecture consisting of two basic processes. The front-end process approaches a spectrogram feature, using Gammatone filters. Regarding the back-end classification, we propose a novel convolutional neural network (CNN) architecture that enforces the network deeply learning middle convolutional layers. Our experiments conducted over DCASE2016 task 1A dataset offer the highest classification accuracy of 84.4% as compared to 72.5% of DCASE2016 baseline.**

***Index Terms*—Acoustic scene classification, deep learning, convolutional neural network, Gammatone spectrogram.**

I. Introduction

Acoustic scene classification (ASC), aiming to categorize the types of locations where a sound was recorded, represents one of the main tasks of a recently appearing research field named "machine hearing" [1]. By exploiting information extracted from the soundscape, ASC is explored in various applications such as context-aware services [2], audio archive management [3], robotic navigation systems [4], and intelligent wearable devices [5]. The most challenge of this task is that a recording related to a given location can contain various sound events. A well-learned model, therefore, should not only focus on performing either background or foreground sounds. Additionally, concerned issues may come from datasets that show different class numbers, recording conditions, biased recording time, making it the most challenging task in the sound recognition area [6]. Hence, recent studies have dedicated to propose various methods for ASC task, and deep learning approach has recently proven effectively [7].

For the front-end extracted feature considered as one main step of an ASC model, mel-frequency cepstral coefficients (MFCC) has widely applied to the research of speech firstly explored in ASC [8]–[11]. Some did experiments on linear predictive coefficients (LPCs) to calculate a power spectrum of the signal [12]. To explore the statistics on MFCC vectors, i-vector which allows to compute statistic attributes has been widely applied [13]. However, acoustic scenes are less structured than speech signals explaining why the mentioned techniques have not shown efficiency. To address this problem, spectrogram features inspired from researches on image processing has recently employed in [14]. Regarding back-end learning models, conventional classifiers, which proved effectively on speech signals such as Gaussian mixture models (GMMs) [8], support vector machines (SVMs) [15], and hidden Markov models (HMMs) [16], were firstly exploited over the ASC task. However, deep learning techniques have recently become a trend for the ASC task [17] and have proved much more effectively [7]. Convolutional neural networks (CNNs) [18] are considered as the most effective classification for ASC tasks, which were early applied [19] and has be shown to be an effective approach. To enhance the classifier, various data augmentation techniques have been approached. Traditional methods are frequency shifting or timing extension mentioned in [20]. These techniques were also used by [21], proving pitch shifting more effective.

Inspired by the aforementioned techniques, this paper, therefore, invokes Gammatone filters [22] to transform audio segment into time-frequency shape before feeding into the back-end classification. We then introduce a baseline proposal based on CNN. Thus, motivated from the transfer learning technique proposed in [23], we propose a training process that forces the baseline learning the middle convolutional layer deeper. This work also applies a data augmentation, namely mixup that is useful to improve our model's performance. To evaluate the performance of our proposed method with different neural networks, we conducted extensive experiments over DCASE2016 task 1A dataset [24]. The experiments demonstrate that our architecture outperforms the conventional models such as GMM, SVM as well as the DCASE2016 baseline.

II. System Description

A. Our Baseline Proposal

In this section, we introduce our baseline proposal described as Fig. 1. The proposed baseline utilizes Gammatone spectrogram in [22] for the front-end feature extraction. By splitting the entire spectrogram into patches with the time and frequency resolution of 128×128, these patches are then fed into the back-end classifier. Regarding the back-end process, the proposed model consists of four convolution blocks and three fully connected layers as detailed in Table I. The first convolution block, denoted as **C01**, uses batchnorm layers between the input and the output of the convolutional layer

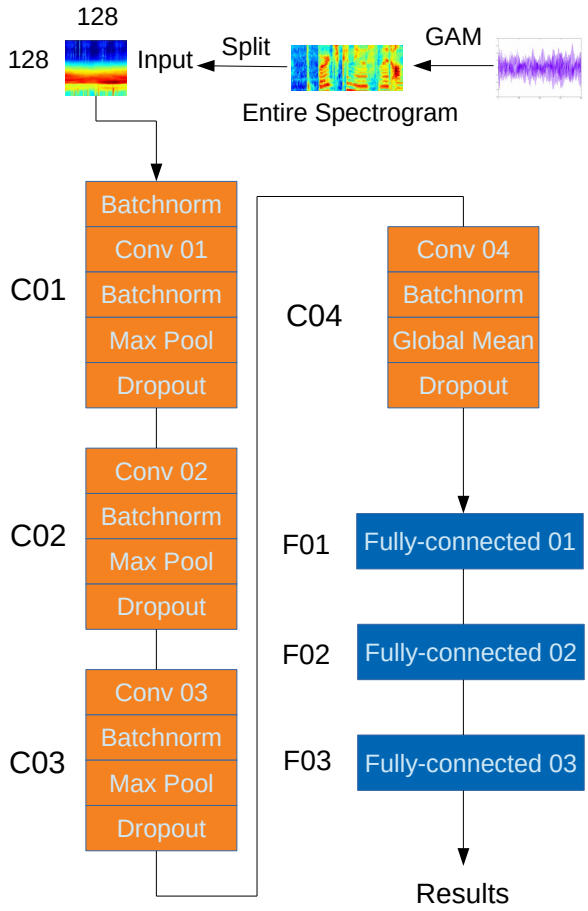

Fig. 1. Our baseline proposal.

to speed up the training process and to avoid the Internal Covariate Shift phenomenon [25]. After subsampling the obtained feature maps with a max-pooling layer, the dropout layer is employed for the purpose of preventing over-fitting. The second and the third blocks, denoted as **C02** and **C03**, have a similar structure to **C01**, a part from no batchnorm layer before convolutional layer. At the convolution block **C04**, instead of using a max-pooling layer, a global-mean pooling layer is applied to enhance the accuracy since all the spatial regions contribute to the output while the max-pooling layer considers the maximum value of local regions. The next three fully connected layers, denoted as **F01**, **F02**, and **F03**, have the role of classification. At the final layer, $softmax$ function, minimizing the cross-entropy as equation below, is applied to tune parameters θ,

$$E(\theta) = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(\hat{y}_i(\theta)) + \frac{\lambda}{2}||\theta||_2^2 \qquad (1)$$

where $E(\theta)$ is the loss function with all parameters θ of the proposed model, N is the total number of items of training data, the sum is over all training inputs, the constant λ is set to 0.0001 since we want the regularization effect to be small, y_i and $\hat{y}$ are expected and predicted results, respectively.

TABLE I
THE PROPOSED CNN BASELINE

Notation	Layer	Output Shape	Kernel Size/Drop
C01	BatchNorm	128x128x1	
	Conv 01	128x128x32	3x3
	BatchNorm	128x128x32	
	Max pooling	64x64x32	2x2
	Dropout	64x64x32	0.1
C02	Conv 02	64x64x64	3x3
	BatchNorm	64x64x64	
	Max pooling	32x32x64	2x2
	Dropout	32x32x64	0.1
C03	Conv 03	32x32x128	3x3
	BatchNorm	32x32x128	
	Max pooling	16x16x128	2x2
	Dropout	16x16x128	0.2
C04	Conv 04	16x16x256	3x3
	BatchNorm	16x16x256	
	Global mean pooling	256	
	Dropout	256	0.2
F01	Fully-connected	512	
F02	Fully-connected	1024	
F03	Fully-connected	15	

B. Deeper Training Method

Inspired by the FreezeOut method suggested by Andrew Brock et al. [26], the training process only trains the hidden layers for a set of portion of the training run, freezes them out one-by-one and excluding them from the backward pass. We then apply this training method on the baseline proposal detailed in Fig. 2.

Our proposed deeper training process can be separated into five sub-training processes namely process **A, B, C, D,** and **E**. First, the training process **A** aims at deeply learning the layer **C01** of the baseline. By extracting the global mean of this layer and adding more fully-connected layers known as **F11, F12, F13** and **F14** detailed as Table II, we have another loss function that focuses on learning the layer **C01**. Both loss functions use (1) and the score is obtained from the original loss function of the baseline proposal. In training process B, we target the layer **C02**. We, therefore, extract global mean add fully-connected layers to learn this layer while the trainable parameters of layer **C01**, transferring from the training process **A**, are remained. Similar to the previous training processes, the **C** and **D** deeply learn layers **C03** and **C04** respectively. Eventually, global mean of the final convolution layer **C04** is extracted and goes through a deep neural network as presented in Table III.

C. Data Augmentation

By increasing data variation, data augmentation has shown itself effective at improving performance in ASC tasks. In this

TABLE II
FULLY-CONNECTED LAYERS TO LEARN MIDDLE LAYERS

Notation	Layer	Output Shape
F11	Fully-connected	256
F12	Fully-connected	512
F13	Fully-connected	512
F14	Fully-connected	15

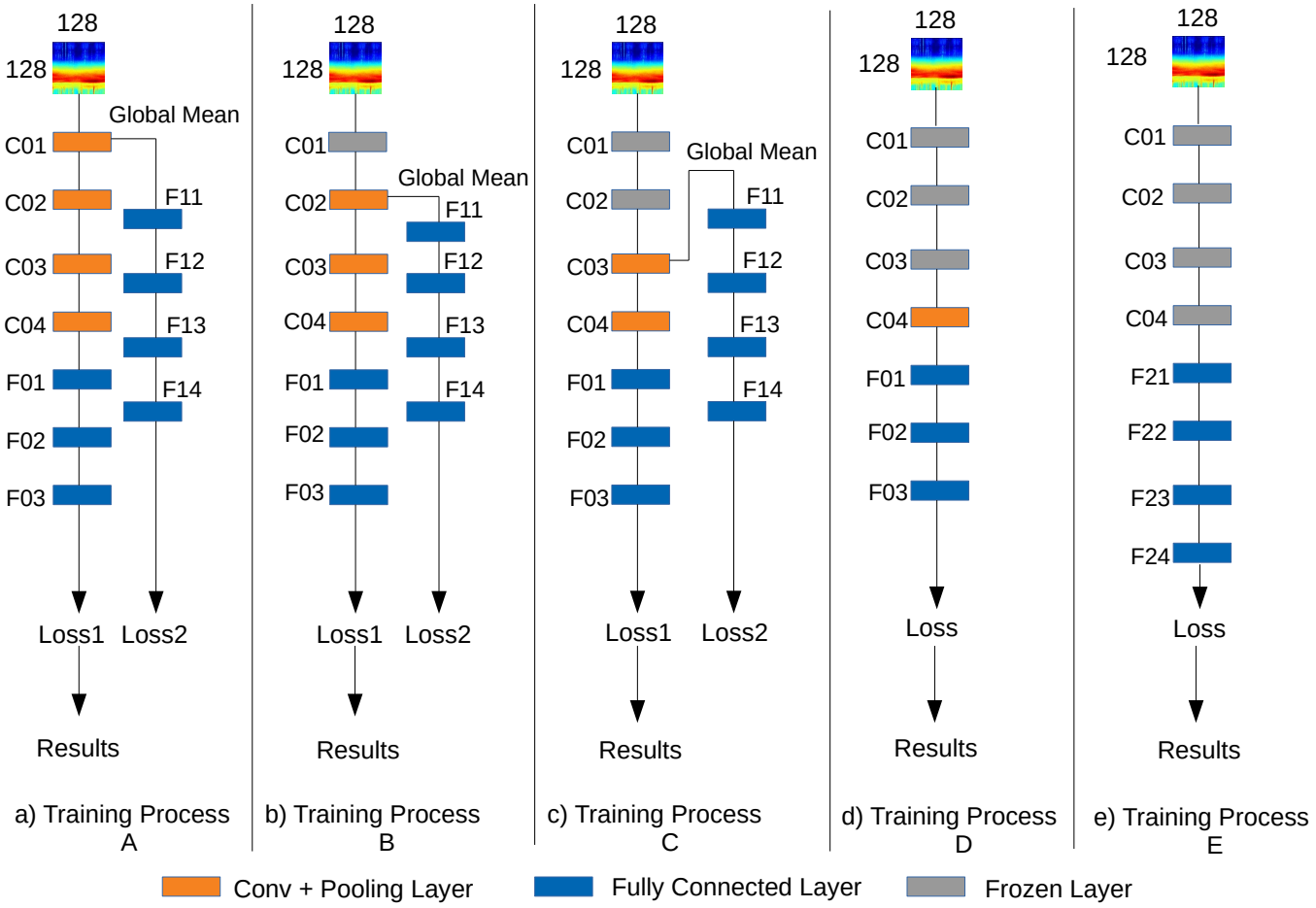

Fig. 2. Our proposed deeper training method.

TABLE III
DEEP NEURAL NETWORK

Notation	Layer	Output Shape
F21	Fully-connected	512
F22	Fully-connected	1024
F23	Fully-connected	1024
F24	Fully-connected	15

case we apply the mixup technique to improve the between-class training. Let X_1, X_2 and y_1, y_2 be the original inputs fed into a learning model and expected one-hot labels from two classes, respectively. From this we generate new mixup data, as follows:

$$X_{mp1} = X_1 * \lambda + X_2 * (1 - \lambda) \quad (2)$$

$$X_{mp2} = X_1 * (1 - \lambda) + X_2 * \lambda \quad (3)$$

$$y_{mp} = y_1 * \lambda + y_2 * (1 - \lambda) \quad (4)$$

$$y_{mp2} = y_1 * (1 - \lambda) + y_2 * \lambda \quad (5)$$

with $\lambda \in U(0, 1)$ is a random mixing coefficient.

We feed both original data and generated mixup data into learning models to the double batch size from 100 to 200, and considerably extend the training time of model. In this work, we apply this technique to both our baseline proposal and our proposed deeper training process (namely training process **A, B, C, D, E**).

III. EXPERIMENT RESULTS

A. Dataset

This paper exploits the TUT Urban Acoustic Scenes 2016 dataset, DCASE2016 [24]. As regards the dataset, the audio signals are recorded in six large European cities, in different locations for each scene class. For each recording location, there are 5-6 minutes of audio. The original recordings are split into segments with a length of 30 seconds that are provided in individual files and the sampling frequency is at 44100 Hz. The dataset includes 15 scenes which are **Bus, Cafe, Car, City Center, Forest path, Grocery Store, Home, Lakeside Beach, Library, Metro Station, Office, Residential Area, Train, Tram, Urban Park**. In this work, we use development set to train the model and test over the evaluation set.

B. Baseline comparison

The obtained average accuracy over the evaluation set reported by the proposed baseline method and by the DCASE2016 baseline [27] is displayed in Table IV. Regarding

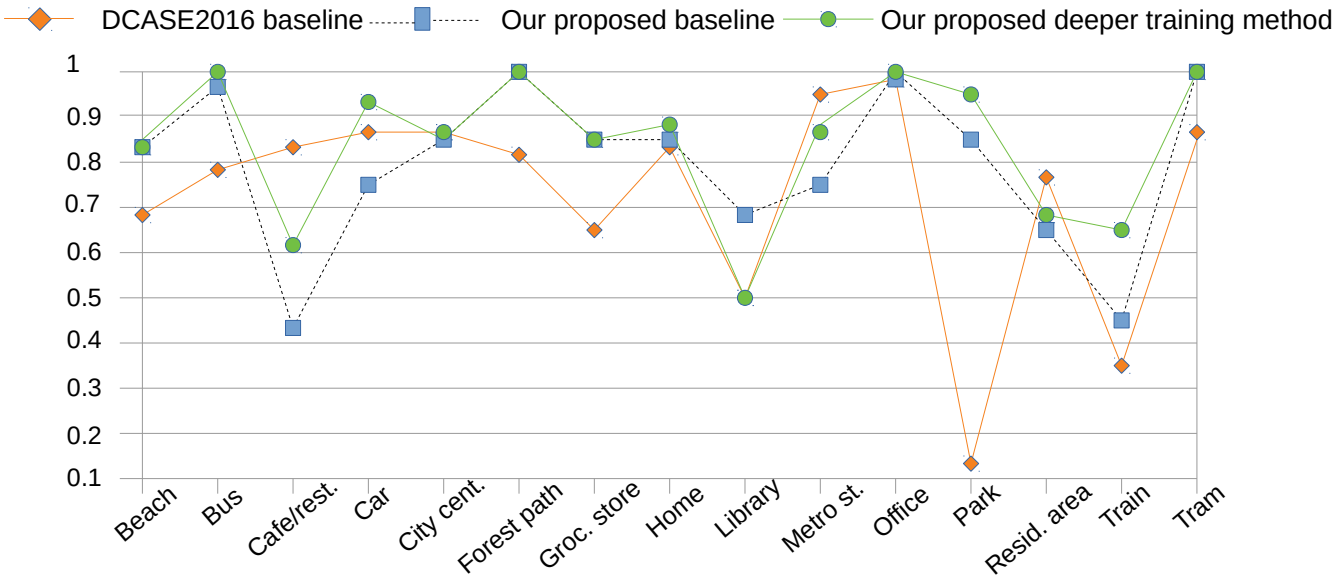

Fig. 3. Performance comparison among DCASE2016 baseline [27], our baseline proposal and deeper training method.

results over the evaluation set, the classification accuracy on our baseline proposal improves the accuracy by 6% compared to DCASE2016 baseline. Specifically, while the accuracy acquired from our proposed baseline method over **park** class is significantly higher than from DCASE2016 baseline, the results over **Cafe/restaurant** of our baseline is much lower.

C. Experiment Results After Deeper Training Process

Using mentioned deeper training method above, the overall results are improved by almost 5% compared to the baseline proposal, 11% compared to the DCASE2016 baseline as shown in Fig. 3. As regards every class, our class accuracy outperforms the baseline proposal, and our method enhances almost the classification accuracy with the exception of the **Library**.

Next, by looking at the confusion matrix in Fig. 4, we are able to know which classes are mostly misclassified. These

TABLE IV
PERFORMANCE COMPARISON WITH THE DCASE2016 BASELINE

Class	DCASE2016	Our Baseline Proposal
Beach	69.3	84.6
Bus	79.6	100
Cafe/rest.	83.2	61.5
Car	87.2	92.3
City Cent.	85.5	84.6
Forest path	81.0	100
Groc. store	65.0	84.6
Home	82.1	88.5
Library	50.4	50
Metro st.	94.7	88.5
Office	98.6	100
Park	13.9	96.2
Resid. area	77.7	69.2
Train	33.6	65.4
Tram	85.4	100
Overall	**72.5**	**79.7**

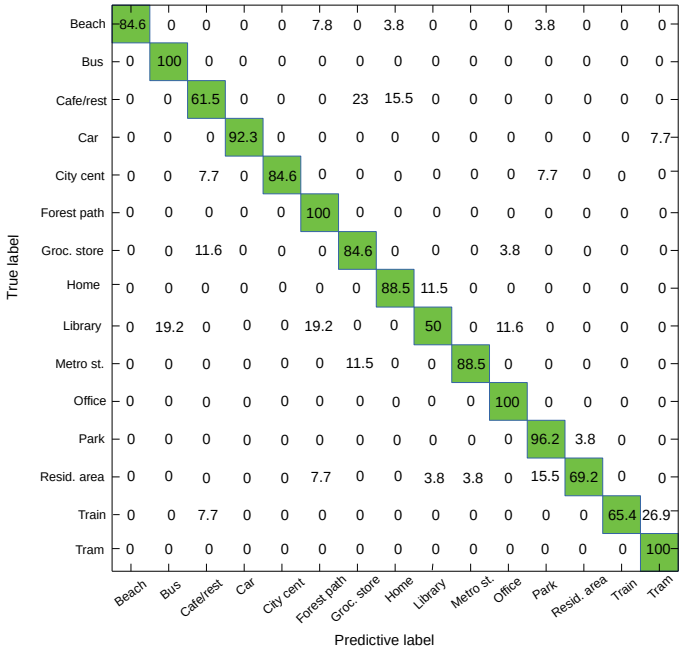

Fig. 4. Confusion matrix for the proposed method evaluated on the evaluation set.

results prove that our model depends more on the background noise than on acoustic event occurrences.

Finally, the overall result of our method is compared with the results of DCASE2016 challenge [27], which is reported in Table V (noting that only single classification models are mentioned since plenty of methods show ensemble approach). The number in Table V reveals that our best result is very competitive to the top results over the single classification and the CNN approach shows strong classification.

TABLE V
PERFORMANCE COMPARISON WITH TOP-TEN DCASE2016 ON THE EVA SET - DCASE2016

System	Classifier	Accuracy
Bae et al. [28]	CNN-RNN	84.1
Lee et al. [29]	CNN	84.6
Takahashi et al. [30]	DNN-GMM	85.6
Kumar et al. [31]	SVM	85.9
Valenti et al. [7]	CNN	86.2
Bisot et al. [32]	NMF	87.7
Our method	**CNN-DNN**	**84.4**

IV. CONCLUSIONS

This paper has presented a novel deep learning framework for the classification of acoustic scenes. Our approach is developed by using on the front-end Gammatone spectrogram and the back-end CNN classification. To deal with implicit challenges in the ASC task, we investigated whether Gammatone spectrogram features could be effective to compare with other spectrogram features as CQT or log-Mel, and whether applying the deeper learning method could improve classification accuracy, allied with the mixup technique. For future research, we plan further investigation on different classifier fusions, as well as explore a combination of bag-of-features front-end processing, since it likely enables us to obtain better performance.

REFERENCES

[1] R. F Lyon, "Machine hearing: An emerging field," *IEEE Sig. Proc. Mag.*, vol. 27, pp. 131–5, 9, 01 2010.

[2] B. N. Schilit, N. Adams, R. Want *et al.*, *Context-aware computing applications*. Xerox Corporation, Palo Alto Research Center, 1994.

[3] S. Chu, S. Narayanan, C.-C. J. Kuo, and M. J. Mataric, "Where am i? scene recognition for mobile robots using audio features," in *2006 IEEE International conference on multimedia and expo*. IEEE, 2006, pp. 885–888.

[4] T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen, "Context-dependent sound event detection," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2013, no. 1, p. 1, 2013.

[5] Y. Xu, W. J. Li, and K. K. C. Lee, *Intelligent wearable interfaces*. Wiley Online Library, 2008.

[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*. IEEE, 2013, pp. 1–4.

[7] M. Valenti, A. Diment, G. Parascandolo, S. Squartini, and T. Virtanen, "Dcase 2016 acoustic scene classification using convolutional neural networks," in *Proc. Workshop Detection Classif. Acoust. Scenes Events*, 2016, pp. 95–99.

[8] J.-J. Aucouturier, B. Defreville, and F. Pachet, "The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music," *The Journal of the Acoustical Society of America*, vol. 122, no. 2, pp. 881–891, 2007.

[9] L. Ma, B. Milner, and D. Smith, "Acoustic environment classification," *ACM TASLP*, vol. 3, no. 2, pp. 1–22, 2006.

[10] B. Cauchi, "Non-negative matrix factorisation applied to auditory scenes classification," *Master's thesis, Master ATIAM, Université Pierre et Marie Curie*, 2011.

[11] J. T. Geiger, B. Schuller, and G. Rigoll, "Large-scale audio feature extraction and svm for acoustic scene classification," in *2013 IEEE WASPAA*. IEEE, 2013, pp. 1–4.

[12] F. Itakura, "Line spectrum representation of linear predictor coefficients of speech signals," *The Journal of the Acoustical Society of America*, vol. 57, no. S1, pp. S35–S35, 1975.

[13] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Twelfth annual conference of the international speech communication association*, 2011.

[14] L. Hertel, H. Phan, and A. Mertins, "Classifying variable-length audio files with all-convolutional networks and masked global pooling," DCASE2016 Challenge, Tech. Rep., September 2016.

[15] A. Rakotomamonjy and G. Gasso, "Histogram of gradients of time–frequency representations for audio scene classification," *IEEE/ACM TASLP*, vol. 23, no. 1, pp. 142–153, 2015.

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

[17] O. Gencoglu, T. Virtanen, and H. Huttunen, "Recognition of acoustic events using deep neural networks," in *2014 22nd EUSIPCO*. IEEE, 2014, pp. 506–510.

[18] 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.

[19] 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," *IEEE AASP Challenge on DCASE*, 2016.

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

[21] R. Lu, Z. Duan, and C. Zhang, "Metric learning based data augmentation for environmental sound classification," in *2017 IEEE WASPAA*. IEEE, 2017, pp. 1–5.

[22] J. Dennis, Q. Yu, H. Tang, H. D. Tran, and H. Li, "Temporal coding of local spectrogram features for robust sound recognition," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 803–807.

[23] A. Diment and T. Virtanen, "Transfer learning of weakly labelled audio," in *2017 ieee waspaa*. IEEE, 2017, pp. 6–10.

[24] A. Mesaros, T. Heittola, and T. Virtanen, "Tut database for acoustic scene classification and sound event detection," in *2016 24th EUSIPCO*. IEEE, 2016, pp. 1128–1132.

[25] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[26] A. Brock, T. Lim, J. M. Ritchie, and N. Weston, "Freezeout: Accelerate training by progressively freezing layers," *arXiv preprint arXiv:1706.04983*, 2017.

[27] A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: Outcome of the dcase 2016 challenge," *IEEE/ACM TASLP*, vol. 26, no. 2, pp. 379–393, 2018.

[28] S. H. Bae, I. Choi, and N. S. Kim, "Acoustic scene classification using parallel combination of lstm and cnn," in *Proceedings of the DCASE2016*, 2016, pp. 11–15.

[29] Y. Han and K. Lee, "Convolutional neural network with multiple-width frequency-delta data augmentation for acoustic scene classification," *IEEE AASP Challenge on DCASE*, 2016.

[30] G. Takahashi, T. Yamada, S. Makino, and N. Ono, "Acoustic scene classification using deep neural network and frame-concatenated acoustic feature," *Detection and Classification of Acoustic Scenes and Events*, 2016.

[31] B. Elizalde, A. Kumar, A. Shah, R. Badlani, E. Vincent, B. Raj, and I. Lane, "Experiments on the dcase challenge 2016: Acoustic scene classification and sound event detection in real life recording," *arXiv preprint arXiv:1607.06706*, 2016.

[32] V. Bisot, R. Serizel, S. Essid, and G. Richard, "Supervised nonnegative matrix factorization for acoustic scene classification," *IEEE AASP Challenge on DCASE*, p. 27, 2016.