HIGH-RESOLUTION ATTENTION NETWORK WITH ACOUSTIC SEGMENT MODEL FOR ACOUSTIC SCENE CLASSIFICATION

Xue Bai[1], *Jun Du*[1], *Jia Pan*[1], *Heng-shun Zhou*[1], *Yan-Hui Tu*[1],*Chin-Hui Lee*[2]

[1] University of Science and Technology of China, HeFei, China
[2] Georgia Institute of Technology, Atlanta, Georgia, USA

ABSTRACT

The spectral information of acoustic scenes is diverse and complex, which poses challenges for acoustic scene tasks. To improve the classification performance, a variety of convolutional neural networks (CNNs) are proposed to extract richer semantic information of scene utterances. However, the different regions of the features extracted from CNN-based encoder have different importance. In this paper, we propose a novel strategy for acoustic scene classification, namely high-resolution attention network with acoustic segment model (HRAN-ASM). In this approach, we utilize fully CNN to obtain high-level semantic information and then adopt two-stage attention strategy to select the relevant acoustic scene segments. Besides, the acoustic segment model (ASM) proposed in our recent work provides embedding vectors for this attention mechanism. The performance is evaluated on DCASE 2018 Task 1a, showing 70.5% good classification accuracy under single system and no data expansion, which is superior to CNN-based self-attention mechanism and highly competitive.

Index Terms— Acoustic scene classification, attention mechanism, acoustic segment model, fully convolutional neural network

1. INTRODUCTION

The goal of Acoustic Scene Classification (ASC) task is to classify the audio to specific scenes, like metro, airport, etc. Acoustic scene recordings contain a large amount of information and rich content. Therefore, the development of automatic identification system for ASC has broad prospects, and its analysis has great potential in a variety of applications such as intelligent sensing devices, audio-based multimedia search, security monitoring and so on. Great progress has been made by several important challenges for ASC, such as Detection and Classification of Acoustic Scenes and Events (DCASE) [1] [2]. However, the complexity of the acoustic scene utterances and the sparsity of effective frames bring difficulties to the ASC task. In this study, we focus on extracting and locating the critical acoustic segments to distinguish different scenes.

ASC has been an active research field for decades. Many traditional models have been investigated for ASC such as Gaussian mixture model (GMM) [3]. [4] investigated high-resolution modeling units of deep neural networks (DNNs) from concrete to abstract for acoustic scene classification based on GMM and ergodic hidden Markov model (HMM). [5] proposed a feature learning approach via decomposing time-frequency (TF) representations with GMM and archetypal analysis (AA). Recently, deep learning techniques have been applied to ASC, such as convolutional neural networks (CNNs) [6], recurrent neural networks (RNNs) [7], and convolutional recurrent neural networks (CRNNs) [8]. To further improve ASC performance, generative adversarial networks (GANs) [9] have been widely investigated. [10] proposed to use GAN-based method for generating additional training database. [11] adopted a calibration transformation to improve the performance of their binaural i-vector system for ASC. Even though the previous methods have improved performance a lot, there are still a lot of basic problems worth exploring. For example, many scenes are quite confusing between each other and have high similarity in time. Even for human beings, it is quite challenging to classify among them. Moreover, CNN-based approaches are hard to capture the correlation of fragments in different scenes. In this paper, we combine acoustic segment model (ASM) and attention mechanism framework to address these challenges in ASC.

Recently, we propose a novel acoustic scene modeling framework [12] in which each audio recording as an acoustic utterance is decoded into a sequence of acoustic events, some of them are highly discriminative, e.g., with high indexing power [13], to access certain acoustic scenes. Although the ASM model is able to analyze the acoustic scene information in detail. Due to the limited modeling ability of ASM by using GMM-HMM, in this study we consider to combine with more powerful model like attention network to further improve the ASC performance. Specifically, the fully convolutional neural network (FCN) is first adopted to extract high-level semantics representations. Then, a two-stage high-resolution attention mechanism is proposed to locate more discriminative acoustic segments. In the first-stage attention, a set of embedding vectors initialized by ASM are designed to achieve a high-resolution attention for the second stage. Compared with the conventional one-stage self-attention mechanism [14], our ASM initialized embedding vectors provide our system a higher-resolution representation and accordingly result in better performance under the premise of no data expansion.

2. THE PROPOSED ARCHITECTURE

In this paper, we propose a novel hybrid approach based on attention mechanism and acoustic segment model (ASM) [12] for acoustic scene classification. CNN-based models have been widely utilized to encode complicated scene utterances into high-level semantics representations. The key point of our method is to make generic CNN-based model be aware of more critical information of the input features by attention mechanism. Besides, the embedding of each scene generated from ASM is regarded as prior information and the key basis for our approach. With the help of the embedding vectors of different scenes and the direct guidance of ASC objective function, the attention mechanism is able to provide useful and discriminative information. As shown in Fig. 1, the overall framework of our proposed HRAN-ASM consists of three parts, namely CNN-based encoder, ASM sequences generation and attention module. The details are introduced as follows.

978-1-5090-6631-5/20/$31.00 ©2020 IEEE

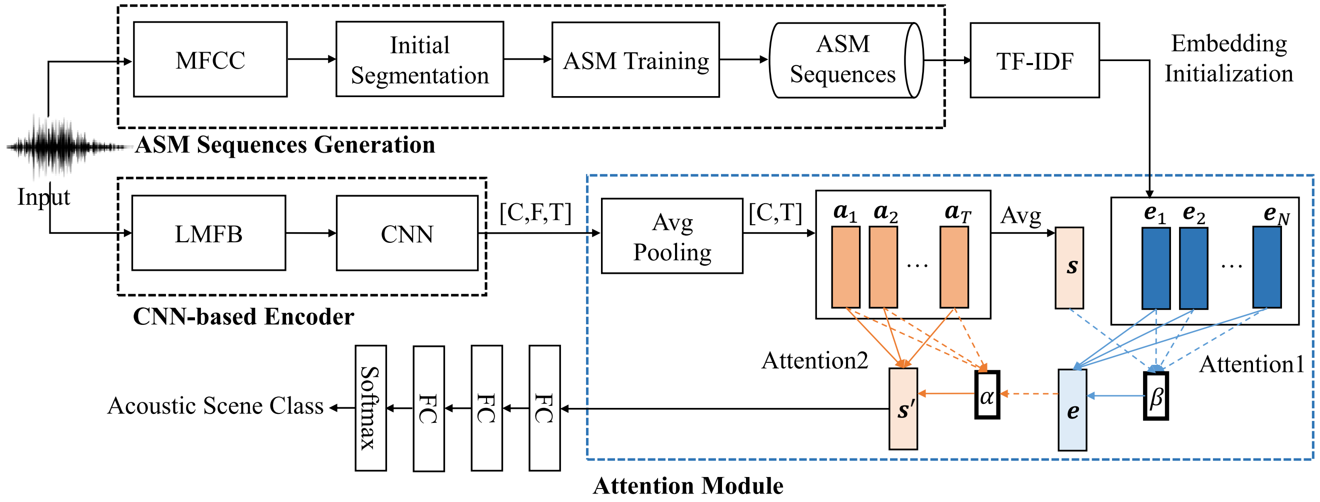


Fig. 1. Overall framework of our proposed HRAN-ASM approach. Attention1 indicates the first attention and attention2 indicates the second attention.

2.1. CNN-based Encoder

CNN has been constantly used as a feature extractor to get abstract and effective information. And it has been proved that CNN-based system can obtain a better accuracy compared with the traditional systems on the ASC task [15]. Inspired by [16], we convert VGGNet-16 [17] into a fully convolutional network (FCN) by simply removing its fully connected layers. Then, we use the VGGNet-16 based FCN as our CNN-based encoder. All the convolution layers are followed by a batch normalization (BN) [18] and a ReLU activation function. In addition, five max pooling layers are added to remove noise and extract robust features. Log Mel-filterbank (LMFB) [19] is as our input feature with the size of $c \times f \times t$, where c is 3 and the 3 channels have the same LMFB. We assume that the output of CNN-based encoder is a 3-dimensional array of size $C \times F \times T$, where F and T are the sizes regarding to the frequency and time domains and C is the number of channels. In order to focus on specific time regions of high-level representations, the attention module is expected to deal with the output of CNN-based encoder.

2.2. ASM Sequences Generation

The ASM sequences generation is one of the most important parts of our method, since it provides additional information for attention mechanism. Originally, the acoustic segment model method was introduced for ASR [20] and recently has been applied to ASC tasks in our previous work [12]. In acoustic segment approach, an acoustic scene is represented as a sequence of acoustic alphabets, which is characterized by a common set of fundamental acoustic units used to span the acoustic space of all possible acoustic events. And these fundamental acoustic units also called ASM units. In this work, the acoustic segment model is used to generate representative embedding for each scene.

Here we introduce how to generate the embedding of different scenes briefly. In general, ASM sequences generation consists of two stages: initial segmentation and ASM training. In the initialization phase, we use GMM-HMM-based method to refine the segment boundaries and the segment labels by the hidden states. Then the hidden states serve as the standard corpus to represent all latent semantic acoustic scene events. In ASM training stage, each ASM unit is modeled by a GMM-HMM and every scene utterance is decoded into a sequence of ASM units. An example of how these ASM sequences look like is “$S3$ $S17$ $S10$ $S8$...”, where “$S3$” represents the first “event” in this sample, “$S17$” is the second, etc.

In order to yield the embedding of each scene category, we first linearly stitch together the ASM sequences belonging to the same category. After that, term frequency (TF) and inverse document frequency (IDF) (**TF-IDF**) [21] are used to obtain the ASM unit counts in each scene. The former TF is the frequency of occurrence of individual ASM unit in each scene and the latter IDF reflects the frequency of the ASM unit in all scenes. Moreover, the dimension of these embeddings is equal to the total number of useful features based on unigram and bigram counts. If there are J ASM units in a scene corpus and assume all unigrams and bigrams exist, then the dimension of the embedding is $K = J \times (J + 1)$. The TF of ASM unit m in the n-th scene is given by

$$TF_{m,n} = \frac{c_{m,n}}{\sum_{k=1}^{K} c_{k,n}} \tag{1}$$

where $c_{m,n}$ is the count of m in the n-th scene. The IDF is given by

$$IDF_m = \log \frac{L+1}{L(m)+1} \tag{2}$$

where L is the number of all scene types and $L(m)$ is the total number of times that ASM unit m appears in all scenes. So each element in the embedding $\boldsymbol{e}_n$ is given by

$$e_{m,n} = TF_{m,n} \times IDF_m \tag{3}$$

Finally, the embedding of N kinds of scene is given by

$$E = \{\boldsymbol{e}_1, \boldsymbol{e}_2, ..., \boldsymbol{e}_N\}, \boldsymbol{e}_n \in \mathbb{R}^K \tag{4}$$

2.3. Attention Module

In this subsection, we propose two-stage attention mechanism to extract discriminative information from the high-level representation. In fact, CNN-based self-attention [14] can also calculate the weight of different time frames. In HRAN-ASM approach we introduce the effective information obtained by the ASM model to help us obtain more accurate discriminative fragments of acoustic scene utterances. Specifically, CNN is designed in speech to diminish the frequency domain variation [22] and reduce the distinguished information in frequency dimension at the output of CNN-based encoder. Therefore, we apply an average pooling layer for frequency dimension reduction and lay emphasis on exploring the impact of different time frames on the classification network. As shown in the attention module, the high-level representation with the size of $C \times F \times T$ is converted to $C \times T$, in which each element is a C-dimensional vector represented as $\boldsymbol{a}_i$. For the feasibility of the algorithm, we set C to be K, which is the dimension of the embedding. And we describe the output as $\boldsymbol{s}$ shown in Eq. (6) through a simple expression.

$$A = \{\boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_T\} \tag{5}$$

$$\boldsymbol{s} = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{a}_t \tag{6}$$

2.3.1. The First Attention of HRAN-ASM Approach

The purpose of the first attention mechanism is to explore the intrinsic connection between the current utterance and different scenes. The ASM model is able to provide N embedding vectors for each sence and these embedding vectors is regard as guiding information for the first attention mechanism. We use the following formulas as the first attention:

$$\beta_i = \frac{\exp(\boldsymbol{e}_i^\top \cdot \boldsymbol{s})}{\sum_{n=1}^{N} \exp(\boldsymbol{e}_n^\top \cdot \boldsymbol{s})}, i \in (1, N) \tag{7}$$

$$\boldsymbol{e} = \sum_{i=1}^{N} \beta_i \boldsymbol{e}_i \tag{8}$$

In Eq. (7), β_i is the attention value scoring the similarity between $\boldsymbol{e}_i$ and $\boldsymbol{s}$ through inner product and softmax function, where $\boldsymbol{e}_i$ is the scene embedding from ASM model and $\boldsymbol{s}$ is the vector representation of the current utterance. Moreover, the Eq. (8) calculates weighted average of $\boldsymbol{e}_i$, which provides addition information for finding discriminative frames.

2.3.2. The Second Attention of HRAN-ASM Approach

The second attention uses the embedding, which is calculated from the first attention, to focus on effective time regions of the output. For the second attention, we use the following formulas:

$$\alpha_j = \frac{\exp(\boldsymbol{e}^\top \cdot \boldsymbol{a}_j)}{\sum_{t=1}^{T} \exp(\boldsymbol{e}^\top \cdot \boldsymbol{a}_t)}, j \in (1, T) \tag{9}$$

$$\boldsymbol{s}' = \sum_{j=1}^{T} \alpha_j \boldsymbol{a}_j \tag{10}$$

The second attention mechanism is designed to extract the elements that are important to the scene of the utterance and aggregate those element arrays to form a scene vector. We measure the importance weight of the $\boldsymbol{a}_j$ by the inner product between $\boldsymbol{a}_j$ and the embedding $\boldsymbol{e}$ calculated from the first attention. After that, the normalized importance weight α_j is calculated through the softmax function. Finally, the summary of current scene $\boldsymbol{s}'$ is computed by Eq. (10).

For multi-classification tasks, we usually use cross-entropy as network loss function. In the training process, to ensure orthogonality between the scene embedding, we add the cosine distance as a penalty to the cross entropy loss. Accordingly, the loss L is defined by

$$L = L_{\text{ce}} + \gamma L_{\cos} \tag{11}$$

$$L_{\cos} = \sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} d_{\cos}(\boldsymbol{e_i}, \boldsymbol{e_j}) \tag{12}$$

$$d_{\cos}(\boldsymbol{e_i}, \boldsymbol{e_j}) = \frac{\boldsymbol{e_i} \cdot \boldsymbol{e_j}}{\|\boldsymbol{e_i}\| \|\boldsymbol{e_j}\|} \tag{13}$$

L_{ce} is cross-entropy loss, $L_{\cos}$ is cosine loss and $d_{\cos}$ is the cosine distance between two embedding vectors. γ is a weighting factor.

3. EXPERIMENTS AND ANALYSIS

3.1. Dataset and Feature Extraction

The experiments were conducted on DCASE2018 Task 1a, which is widely used as a benchmark for acoustic scene classification. The audio recordings with 48 kHz sampling rate in 10 different scenes were recorded by electret binaural microphone. The length of each audio recording is 10 seconds. For this study, we convert the binaural recordings into mono recordings. The input of CNN-based encoder is 128-dimensional log Mel-filterbank (LMFB) [19] features. For the ASM sequences generation, the 60-dimensional mel-frequency cepstral coefficients (MFCC) with Δ and $\Delta\Delta$ are utilized as input features. Both LMFB and MFCC adopt a 40-ms observation window with a 20-ms overlap. According to the official requirement, the development dataset is divided into training and test subsets. The training subset and test subset contain 6122 and 2518 segments, respectively.

3.2. Experimental Configuration and Results

In this subsection, we show the experimental configuration and the results of the acoustic scene classification based on our method. First, the VGGNet-16 removing its fully connected layer is used as the CNN-based encoder and the number of output channels in the last convolutional layer is set to 405. And we initialize the encoder with random initialization parameters. Second, we implement acoustic segment model to generate 20 ASM units and transcribe each utterance to the ASM sequences. **TF-IDF** is performed on the entire training subset to obtain the 405-dimensional embedding vectors for each scene. More detailed setups of ASM can refer to [12]. Finally, two-stage attention mechanism combined with embedding vectors is utilized to get salient frames of each scene and improve recognition performance. The CNN model is trained with stochastic gradient descent (SGD) [23] based backpropagation (BP) algorithm. The initial learning rate is set to 0.005 and 60 epoch are conducted. The coefficient γ is set to 1.

Table 1. The performance of different approaches on test set.

System	VGGNet-16[24]	ASM[12]	Self-Attention
Accuracy	67.4%	66.1%	68.9%

Table 1 shows the performance comparison of different approaches on test set. “VGGNet-16”, “ASM[12]” and “Self−Attention” denote the VGGNet-16-based classifier without attention mechanism, our previously proposed ASM approach [12] and VGGNet-16 with self-attention approach [14]. First, we can find that the self-attention approach can improve the ASC performance comparing to “VGGNet-16”, which demonstrates that the attention mechanism can be aware of more critical information for ASC task. Second, “ASM” can achieve the comparable performance to “VGGNet-16”, although these are two very different approaches. This result may imply that there are complementary between the two methods.

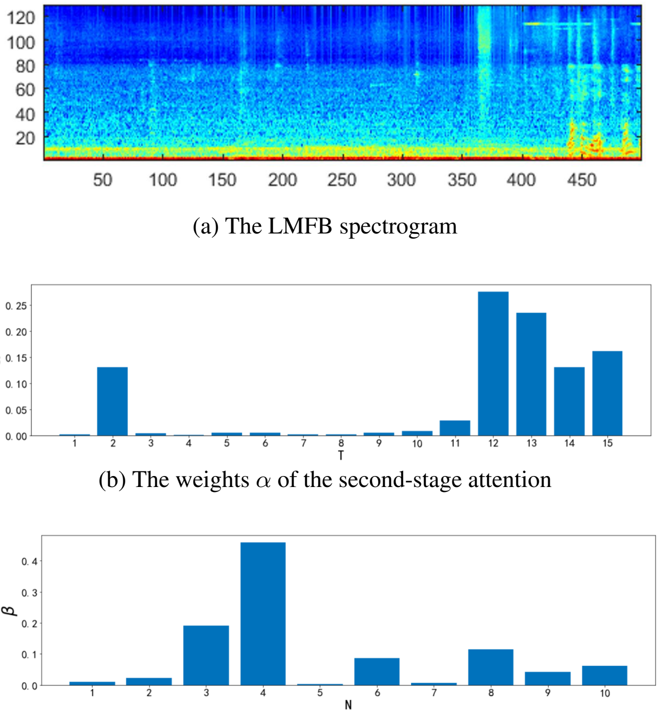
Table 2. The performance comparison of our HRAN approach with different initialization methods for the embedding vectors on test set.

System	Self-Attention	HRAN-Orth	HRAN-ASM
Accuracy	68.9%	68.3%	70.5%

Table 2 shows performance comparison of our HRAN approach with different initialization methods for the embedding vectors on test set. “HRAN-Orth” and “HRAN-ASM” denote the proposed HRAN approach with random orthogonal initialization (Orth) [25] and ASM initialization for embedding vectors, respectively. First, we can find that both “HRAN-Orth” and ”HRAN-ASM” outperform “VGGNet-16” in Table 1, which indicates that the proposed attention mechanism also can be aware of more critical information for ASC task. Second, “HRAN-ASM” can further improve the performance comparing to “HRAN-Orth”, e.g., the accuracy from 68.3% to 70.5%. It demonstrates that the first attention mechanism based on ASM initialization is better than that based on orthogonal initialization. Considering that different scenes might include similar segments, the embedding vectors in HRAN-Orth are not sufficient to characterize each scene comparing to that in HRAN-ASM. Hence, the embedding initialized by ASM leads to much better performance than orthogonal initialization. Finally, our proposed HRAN-ASM approach which utilized the embedding vectors obtained by ASM and VGGNet-16 simultaneously achieves the best performance, which demonstrates the strong complementarity between the embedding vectors obtained by ASM and VGGNet-16.

3.3. Analysis and Visualization

To illustrate the attention mechanism effect in HRAN-ASM approach, we visualize the weights of the first-stage attention and the second-stage attention respectively. Fig. 2 (a) shows the LMFB spectrogram of one example from the Bus scene. Fig. 2 (b) shows the corresponding weights α of the second-stage attention over the time axis. In our experiments, T is 15. For the first-stage attention weights α displayed in Fig. 2 (c), the blue bars represent the weights of N ASM embedding vectors, where the N is 10. From Fig. 2 (b), the LMFB features at different time indices are assigned with different weights, which describes the importance contribution at different time indices to current acoustic scene. In the last three seconds of this utterance the female conductor calls out and subsequently the bus is arriving, so it’s easy to judge that the category is “bus”. Obviously our HRAN-ASM approach is able to explicitly show larger weights to those critical segments and give better classification results accordingly. Compared with the conventional self-attention approach, our approach can generate different weights (β in Fig. 2 (c)) to the set of embedding vectors initialized by ASM while only a global embedding vector is adopted in the self-attention case.



(a) The LMFB spectrogram

(b) The weights α of the second-stage attention

(c) The weights β of the first-stage attention

Fig. 2. Visualization of attention for one example in Bus scene.

4. CONCLUSIONS

In order to consider more critical information of the input features, we propose a novel approach based on acoustic segment model and two-stage attention mechanism to solve scene confusion problems. The acoustic segment model is used to generate representative embedding for each scene as a guided information, and a two-stage attention mechanism combined with embedding vectors is utilized to get salient frames of each scene and improve recognition rate. The experiments verify that our approach achieves highly competitive performance under single system and no data expansion. Besides, we demonstrate that our hybrid approach is able to explicitly show the correlation of segments in different scenes by the two-stage attention.

5. ACKNOWLEDGEMENT

This work was supported in part by the National Key R&D Program of China under contract No. 2017YFB1002202, the National Natural Science Foundation of China under Grant Nos. 61671422 and U1613211, the Key Science and Technology Project of Anhui Province under Grant No. 17030901005. This work was also funded by Huawei Noah’s Ark Lab.

6. REFERENCES

[1] Dan Stowell, Dimitrios Giannoulis, Emmanouil Benetos, Mathieu Lagrange, and Mark D Plumbley, “Detection and classification of acoustic scenes and events,” *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1733–1746, 2015.

[2] Houwei Zhu, Chunxia Ren, Jun Wang, S Li, L Wang, and L Yang, “Dcase 2019 challenge task1 technical report,” Tech. Rep., DCASE2019 Challenge, Tech. Rep, 2019.

[3] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen, “Tut database for acoustic scene classification and sound event detection,” in *2016 24th European Signal Processing Conference (EUSIPCO)*. IEEE, 2016, pp. 1128–1132.

[4] Xiao Bao, Tian Gao, Jun Du, and Li-Rong Dai, “An investigation of high-resolution modeling units of deep neural networks for acoustic scene classification,” in *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2017, pp. 3028–3035.

[5] Vinayak Abrol, Pulkit Sharma, and Anshul Thakur, “Gmm-aa system for acoustic scene classification,” Tech. Rep., DCASE2017 Challenge, Tech. Rep, 2017.

[6] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen, “Acoustic scene classification: An overview of dcase 2017 challenge entries,” in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2018, pp. 411–415.

[7] Jeroen Zegers et al., “Cnn-lstm models for multi-speaker source separation using bayesian hyper parameter optimization,” *Proceedings Interspeech 2019*, 2019.

[8] Hugo Jallet, Emre Cakır, and Tuomas Virtanen, “Acoustic scene classification using convolutional recurrent neural networks,” *the Detection and Classification of Acoustic Scenes and Events (DCASE)*, pp. 1–5, 2017.

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, “Generative adversarial nets,” in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[10] Seongkyu Mun, Sangwook Park, David K Han, and Hanseok Ko, “Generative adversarial network based acoustic scene training set augmentation and selection using svm hyperplane,” *Proc. DCASE*, pp. 93–97, 2017.

[11] Hamid Eghbal-Zadeh, Bernhard Lehner, Matthias Dorfer, and Gerhard Widmer, “Cp-jku submissions for dcase-2016: a hybrid approach using binaural i-vectors and deep convolutional neural networks,” *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2016.

[12] Xue Bai, Jun Du, Zi-Rui Wang, and Chin-Hui Lee, “A hybrid approach to acoustic scene classification based on universal acoustic models,” *Proc. Interspeech 2019*, pp. 3619–3623, 2019.

[13] Bin Ma, Haizhou Li, and Chin-Hui Lee, “An acoustic segment modeling approach to automatic language identification,” in *Ninth European Conference on Speech Communication and Technology*, 2005.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, “Attention is all you need,” in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[15] Yoonchang Han and Kyogu Lee, “Convolutional neural network with multiple-width frequency-delta data augmentation for acoustic scene classification,” *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events*, 2016.

[16] Yuanyuan Zhang, Jun Du, Zirui Wang, Jianshu Zhang, and Yanhui Tu, “Attention based fully convolutional network for speech emotion recognition,” in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2018, pp. 1771–1775.

[17] Karen Simonyan and Andrew Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.

[18] Sergey Ioffe and Christian Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” *arXiv preprint arXiv:1502.03167*, 2015.

[19] Vivek Tyagi and Christian Wellekens, “On desensitizing the mel-cepstrum to spurious spectral components for robust speech recognition,” in *Proceedings.(ICASSP’05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005*. IEEE, 2005, vol. 1, pp. I–529.

[20] Chin-Hui Lee, Frank K Soong, and Bing-Hwang Juang, “A segment model based approach to speech recognition,” in *ICASSP-88., International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 1988, pp. 501–541.

[21] Juan Ramos et al., “Using tf-idf to determine word relevance in document queries,” in *Proceedings of the first instructional conference on machine learning*. Piscataway, NJ, 2003, vol. 242, pp. 133–142.

[22] Tara N Sainath, Oriol Vinyals, Andrew Senior, and Haşim Sak, “Convolutional, long short-term memory, fully connected deep neural networks,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4580–4584.

[23] Léon Bottou, “Large-scale machine learning with stochastic gradient descent,” in *Proceedings of COMPSTAT’2010*, pp. 177–186. Springer, 2010.

[24] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen, “A multi-device dataset for urban acoustic scene classification,” *arXiv preprint arXiv:1807.09840*, 2018.

[25] Mikko Lehtokangas, Jukka Saarinen, Kimmo Kaski, and Pentti Huuhtanen, “Initializing weights of a multilayer perceptron network by using the orthogonal least squares algorithm,” *Neural Computation*, vol. 7, no. 5, pp. 982–999, 1995.