



# Deep Segment Model for Acoustic Scene Classification

Yajian Wang<sup>1</sup>, Jun Du<sup>1,\*</sup>, Hang Chen<sup>1</sup>, Qing Wang<sup>1</sup>, Chin-Hui Lee<sup>2</sup>

<sup>1</sup>University of Science and Technology of China, HeFei, China

<sup>2</sup>School of Electrical and Computer Engineering, Georgia Institute of Technology, GA, USA

yajian@mail.ustc.edu.cn, sjundu@ustc.edu.cn

## Abstract

In most state-of-the-art acoustic scene classification (ASC) techniques, convolutional neural networks (CNNs) are adopted due to their extraordinary ability in learning local deep features. However, the CNN-based approach is unable to effectively describe the structure of sound events in an audio clip, which is a key element in distinguishing acoustic scenes with similar characteristics, whereas the acoustic segment model (ASM) based approach shows its superiority. To take full advantage of these two types of approaches, we proposed a novel deep segment model (DSM) for ASC. DSM employs a fully convolutional neural network (FCNN) as a deep feature extractor and then guides the ASM to better capture semantic information among sound events. Specifically, the FCNN-based encoder is trained with the multi-task of classifying both three coarse-grained acoustic scenes and ten fine-grained acoustic scenes to extract multi-level acoustic features. Moreover, an entropy-based decision fusion strategy is designed to further utilize the complementarity of FCNN-based and DSM-based systems. The final system achieves an accuracy of 80.4% in the DCASE2021 Task1b audio dataset, yielding a relative error rate reduction of about 15% over the FCNN-based system.

**Index Terms:** acoustic scene classification, convolutional neural network, acoustic segment model, entropy-based decision fusion strategy

## 1. Introduction

The task of acoustic scene classification (ASC) aims to identify real-life audios into specific scene classes, such as street pedestrians, shopping malls, etc. A real-life scene audio consists of a series of sound events. The characteristics of sound scenes are complex and diverse which proposes challenges to ASC systems. The organizers of the Detection and Classification of Acoustic Scenes and Events (DCASE) challenge have promoted this field by releasing public datasets as well as a competitive platform [1, 2]. ASC is an active research field and has drawn much attention for years. Many traditional techniques are applied in ASC such as Gaussian mixture models (GMM) [3, 4, 5] and hidden Markov models (HMM) [6]. Recently, in most state-of-the-art (SOTA) ASC techniques, convolutional neural networks (CNNs) are adopted due to their extraordinary ability to extract deep features [7, 8, 9, 10, 11]. Besides, many advanced techniques emerge and further improve the ASC accuracy, such as transfer learning [12, 13], attention mechanism [14, 15, 16] and generative adversarial networks (GAN) based data augmentation [17].

Despite that the CNN has advantages in learning deep features with higher performance, it still causes confusion in some acoustic scenes such as street traffic and street pedestrian. By

analyzing these acoustic scenes, we find that these acoustic scenes contain similar sound events. For example, outdoor footsteps, a sound event with a high occurrence frequency, will exist in street traffic, street pedestrians and parks. In this situation, our brain tends to detect more discriminated acoustic clues to make a judgment, such as birdsong in the park or car engine sound in street traffic. Inspired by this phenomenon, we infer that higher accuracy will be gained in classifying confusing scenes by capturing prominent acoustic events in the audio clip. The fully convolutional neural network (FCNN) [18] relies on acoustic events and events duration to classify the input scene audios. Thus, the FCNN will be affected by long-term interference factors and ignore short-term saliency information, resulting in misclassification.

Acoustic segment models (ASMs) are a set of self-organized acoustic units that cover the entire range of acoustic characteristics and were originally proposed to represent basic units and acoustic lexicons [19]. The ASM framework is applied in a variety of fields such as spoken language recognition [20], emotion recognition [21] and music retrieval [22]. [23] employed the ASM in acoustic scene classification. Due to the limited modeling ability, the ASM approach performs poorly compared with CNN models. There have been several studies on the ensemble of ASM and CNN [14]. However, these methods focus on low-level features such as Mel-frequency cepstral coefficients (MFCC) and log Mel-filterbank (LMFB). Our work focuses on discovering semantic information over high-level acoustic features which can better represent the sound events.

In this study, we propose a novel hybrid approach called deep segment model (DSM) where fully convolutional neural networks (FCNNs) are adopted to obtain high-level features, and ASMs are used to capture semantic information among sound events based on high-level acoustic features. Specifically, the FCNN-based encoder is trained with the multi-task of classifying both three coarse-grained acoustic scenes and ten fine-grained acoustic scenes to extract multi-level acoustic features. Next, we rely on the ASM to perform semantic modeling on these high-level features. Unsupervised hierarchical K-Means clustering is used to discrete the audio features into acoustic segment units and obtain the initial ASM sequence for each audio. Then the Gaussian mixture model-hidden Markov model (GMM-HMM) is employed to model ASM units and each audio is decoded into the final ASM sequences. We transfer the ASM sequences into embeddings with TF-IDF and map the embedding into ten acoustic scenes by a simple DNN classifier. However, detailed local information such as the duration and the confidence level of sound event could be possibly lost during acoustic segment modeling. Thus, an entropy-based decision fusion strategy is designed to further utilize the complementarity of FCNN and DSM systems. The final result achieves an accuracy of 80.4% in the DCASE2021 Task1b audio dataset.

\*corresponding author

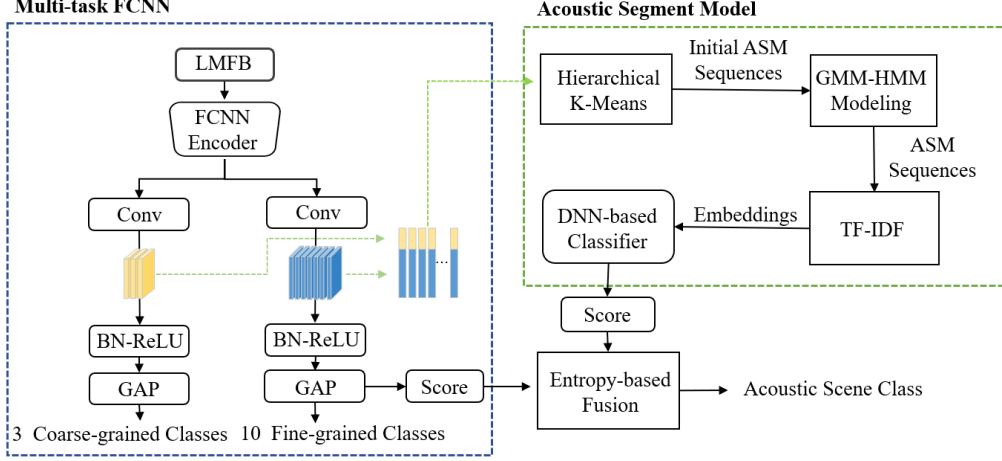


Figure 1: The overall framework of our proposed approach

## 2. The Proposed Framework

The system flowchart of the proposed DSM framework is shown in Figure 1. It mainly consists of three parts: multi-task FCNN, acoustic segment model and entropy based fusion module. In the following subsections, we will discuss these three parts in detail.

### 2.1. Multi-task FCNN

The FCNN model is adopted as our baseline, which has been applied in our previous work [18] and achieves the SOTA results in DCASE2020 task1a data set [24]. In this study, we extend it to multi-task FCNN by classifying both coarse-grained scene classes and fine-grained scene classes. To extract acoustic features of sound events at different levels, we cluster 10 fine-grained acoustic scenes into 3 coarse-grained categories, i.e. in-door, out-door, and transportation, according to our prior knowledge. The clustered 3 categories are also adopted as additional higher-level classification target to train a multi-task FCNN. Acoustic scenes in the coarse-grained class have similar acoustic characteristics and therefore are more difficult to classify. Instead of training two completely independent classifiers, we adopt multi-task learning, where the two tasks share the basic encoder, while preserving several task-specific decoding layers. Parameter sharing alleviates the overfitting issue of models while reducing learned parameters. The multi-task cross-entropy loss  $L_{\text{MTL}}$  is calculated as:

$$L_{\text{MTL}} = -\frac{1}{N_{\text{B}}} \sum_{n=1}^{N_{\text{B}}} \log p(y_n^c | \mathbf{X}_n) - \frac{1}{N_{\text{B}}} \sum_{n=1}^{N_{\text{B}}} \log p(y_n^f | \mathbf{X}_n) \quad (1)$$

where  $\mathbf{X}_n$  is the input feature vector sequence of  $n$ -th utterance,  $y_n^c$  and  $y_n^f$  are the ground-truth labels of  $\mathbf{X}_n$  for coarse-grained and fine-grained classifiers,  $p(y_n^c | \mathbf{x}_n)$  and  $p(y_n^f | \mathbf{x}_n)$  are the predicted probability posterior of coarse-grained and fine-grained classifiers,  $N_{\text{B}}$  denotes the minibatch size.

The basic FCNN encoder consists of 8 convolution layers as the shared hidden layers. Followed by each convolution layer, a batch normalization (BN) and a ReLU activation function are applied. To eliminate overfitting and extract robust features, max-pooling layers and dropout layers are used. In particular, we reduce the downsampling of the temporal dimension to pre-

serve the information in the time dimension. The output of the basic encoder will be parallelly fed to decoders of both 3-class branch and 10-class branch. A global average pooling (GAP) layer and soft-max are used to get the final utterance level prediction results. As shown in Figure 1, deep features are extracted after the convolution layer in decoders of both 3-class and 10-class respectively. We convert the embedding from the size of  $C \times F \times T$  to  $T \times CF$ , where  $F$  and  $T$  are the sizes regarding the frequency and time domains and  $C$  represents the number of channels. The representations of two levels are concatenated as the input of the acoustic segment model.

### 2.2. Acoustic Segment Model

The core idea of ASM is to express the scene audio as a sequence of basic acoustic units named ASM sequences [19]. We assume that the acoustic sound features can be encompassed by a set of acoustic units as the phoneme units.

#### 2.2.1. ASM Sequence Generation

There are two main stages to generate the ASM sequences. First, initialize the ASM sequences by converting the raw continuous acoustic features into a discrete sequence of ASM units. Second, apply GMM-HMM to refine the segment boundaries and labels of initial sequences. Hierarchical K-Means [25] algorithm is used to tokenize the acoustic features and generate initial ASM sequences. We could modify the hierarchy levels and the number of clusters  $d_i$  in each level  $i$ . The number of clusters in total is  $\prod_i d_i$ . We set 2 hierarchy levels with  $d_1 = 20$ ,  $d_2 = 2$ , which yields  $20 \times 2 = 40$  clusters in total. Each cluster denotes an ASM unit. Thus, a list of feature vectors could be transformed into a sequence of ASM units according to the clustering results of each vector, which is the initial ASM sequence. Hierarchical K-Means is applied on all feature vectors of training data to find the centroids. The whole process of initializing ASM sequences is unsupervised due to no prior knowledge being used. The initial ASM sequences could roughly represent the variation of sound events according to changes in feature vectors but cannot well reflect the temporal correlation of sound events. Thus, we use the GMM and left-to-right HMM to model each ASM unit to refine the ASM sequences. In particular, we adopt 1-state HMM to model

each ASM unit because initial ASM sequences already label each frame of features which is more accurate than initial force alignment. Moreover, 1-state HMM could reduce training parameters. Due to the input deep features already being highly aggregated, the capability of GMM is sufficient for modeling and could prevent overfitting compared to other deep models.

### 2.2.2. TF-IDF and Classifier

After translating the acoustic scene audios into ASM sequences, we adopt text vectorization techniques to capture semantic information in ASM sequences. The term frequency (TF) and inverse document frequency (IDF) [26] are employed to describe the indexing power of basic units [27]. We use the ASM n-grams to describe the constraints of acoustic segments. In addition to a single significant event, distinction clues may exist in the contextual acoustic segment. Thus, apart from uni-gram, bi-gram which is the group of two ASM units is also used as basic term. Here,  $N$  signifies the total number of basic terms. TF represents the occurrence frequency of the basic terms in the text and is calculated by

$$TF_{i,j} = \frac{o_{i,j}}{\sum_{n=1}^N o_{n,j}} \quad (2)$$

where  $o_{i,j}$  is the count of  $i$  in the ASM sequence  $j$ . IDF calculate the frequency of an ASM unit in all text to measure how much information the ASM unit provides to reflect the importance of each ASM. IDF is calculated by

$$IDF_i = \log \frac{M+1}{M_j+1} \quad (3)$$

where  $M$  is the number of training scene ASM sequences and  $M_j$  is the total number of times that ASM unit  $i$  appears in the training scene transcripts. The final vector  $v_n$  is given by

$$v_{i,j} = TF_{i,j} \times IDF_i \quad (4)$$

Finally, we use a simple classifier of fully connected neural networks to classify the vectors into acoustic scenes.

### 2.3. Entropy-based Decision Fusion Strategy

ASM provides a holistic view of sound events and describes the significance of acoustic events in an audio clip as well as the contextual relevance. FCNN, on the other hand, extracts the specific local characteristics of acoustic events. DSM system combines the two models to take advantage of their respective strengths. However, there are still some problems worth exploring. First, detailed local information such as the duration of a sound event could be possibly lost during acoustic segment modeling, which is preserved by the FCNN-based model. Second, acoustic segment modeling is sensitive to some anomalies fragment of several frames in sound clips which leads to misclassification. Based on this observation, we propose an entropy-based ensemble strategy to further take advantage of the complementarity of the two models. In information theory, the entropy of a random variable is the average level of information or uncertainty inherent in the variable's possible outcomes [28]. Here, information entropy is applied to denote the uncertainty of the posterior probability of the model output. If we use  $p(i|\mathbf{X})$  to denote the probability of input feature sequence  $\mathbf{X}$  being classified as  $i$ -th scene class, the entropy  $H$  can be calculated as:

$$H = -\sum_{i=1}^N p(i|\mathbf{X}) \log p(i|\mathbf{X}) \quad (5)$$

where  $N$  denotes the number of acoustic scene classes. We assume that the higher the information entropy, the higher the uncertainty of the classification results. Thus, we pay more attention to the model with lower information entropy. We design the weight  $w$  as

$$w = \left( \frac{H_{\text{FCNN}}}{H_{\text{FCNN}} + H_{\text{DSM}}} \right)^\gamma \quad (6)$$

where  $H_{\text{FCNN}}$  and  $H_{\text{DSM}}$  present the entropy of outputs of FCNN-based model and DSM system, respectively.  $\gamma$  is a hyperparameter. The final score is calculated by

$$\mathbf{P}_{\text{Fusion}} = w\mathbf{P}_{\text{DSM}} + (1-w)\mathbf{P}_{\text{FCNN}} \quad (7)$$

where  $\mathbf{P}_{\text{DSM}}$ ,  $\mathbf{P}_{\text{FCNN}}$  represent output vector of the DSM and the FCNN, respectively.

## 3. Experiments and Analysis

### 3.1. Experimental Setup

The proposed system is evaluated on the DCASE 2021 Task1b audio development data set [29], which consists of 34h 10-second two-channel audio clips recorded in 10 different acoustic scenes. The audio data is recorded with a 48 kHz sampling rate and 24-bit resolution. The development data set is divided into training and test sets, each having 8648 and 3645 utterances, according to the official recommendation. Log Mel-filterbank (LMFB) features are employed as inputs of the system. To generate LMFB, the short-time Fourier transform (STFT) with 2048 FFT points is applied to both left and right channels of audios, utilizing a window size of 2048 samples and a hop length of 1024 samples. Deltas and delta-deltas of LMFB are added to the input feature and the final size of the feature tensor is  $6 \times 128 \times 461$ . The FCNN model is built with Pytorch. ADAM [30] optimizer is used to train the model with the learning rate of 0.001. High-level features are extracted with the size of  $10 \times 4 \times 115$  in a 10-class decoder and the size of  $3 \times 4 \times 115$  in a 3-class decoder. We convert the two 3-dimension features to the size of  $115 \times 40$  and the size of  $115 \times 12$  vectors and concatenate them in the frequency dimension. The final size of input features vectors for acoustic scene modeling is  $115 \times 52$ .

Next, all transformed features are transcribed by 40 ASM units using hierarchical K-Means. Then, the 1-state left-to-right GMM-HMM is used to refine the ASM sequences. Each state has 60 Gaussian mixtures models. Finally, we adopt the fully connected layer with one hidden layer, followed by a dropout and ReLU layer to map the embedding to 10 acoustic scenes. The neurons of the hidden layer are set as 128 and the dropout rate is 0.2. SGD [31] is used to train the classifier with the learning rate of 0.03. In the entropy-based fusion stage, the hyperparameter  $\gamma$  is set as 1.5.

### 3.2. ASC Experimental Results

Table 1: Results on DCASE 2021 Task1b audio data set

System	Accuracy
FCNN	76.7%
ASM[23]	62.9%
DSM	79.0%
DSM-AF	80.0%
DSM-EBF	80.4%

Table 2: Results of DSM with different level of features

Model	Accuracy
FCNN-3class	95.3%
FCNN-10class	76.7%
2stage-FCNN[18]	77.5%
DSM (3class)	47.8%
DSM(10class)	78.4%
DSM(3&10class)	79.0%

Table 1 shows the ASC accuracy of different systems. The FCNN baseline model achieves 76.7% classification accuracy. The ASM approach performs poorly due to the limited modeling ability of the GMM model. The last three rows show the results obtained by our proposed approach. DSM system achieves the accuracy of 79%, which proves the validity of ASM over high-level characteristics. Two fusion strategies are applied with the output scores of FCNN and DSM separately. "DSM-AF" presents the average score fusion strategy and "DSM-EBF" denotes the proposed entropy-based fusion strategy over DSM and FCNN approaches. The results show that the accuracy boosts up to 80.4% after applying entropy-based fusion, which proves the strong complementarity between FCNN and ASM. And our entropy-based fusion strategy has proven advantages over the straight average score fusion method.

### 3.3. Results with Features of Different Levels

In this subsection, we discuss the reflection caused by different levels of embeddings, derived from the 3-class and 10-class decoders, on the performance DSM system. As shown in table 2, the first two rows show the accuracy of multi-task FCNN to classify 3 broad class scenes and 10 class scenes. The 3-class task is relatively easy and achieves an ASC accuracy of 95.3%. Then we apply acoustic scene modeling with 3-class features, 10-class features, and the combination of the two features, respectively. The last three rows in Figure 2 show the result. The 10-class features have a more accurate representation of the acoustic audio than the 3-class features, so it has a better classification performance. However, the 3-class representations alone yield poor results. The classification accuracy is further improved by the combination of the two features, which proves the complementarity between different levels of features. In addition, the two-stage approach proposed in [18] is also deployed for a comprehensive assessment, in which score fusion was used to combine a broad 3-class classifier and a particular 10-class classifier and achieve the ASC accuracy of 77.5%. From the results, we can see that our fusion method of 3-class and 10-class classifiers has better classification performance.

### 3.4. Analysis and Visualization

To explore the ASM effect in the DSM system, we extract the feature tensors with the size of  $N \times F \times T$  before the last global average pooling layer of FCNN, where  $F$ ,  $T$  and  $N$  denote the frequency, the time and the number of acoustic scenes, respectively. Then we calculate the average value in dimension  $F$  to obtain the frame-level probability belonging to each scene. In Figure 2, we visualize the 10-class frame-level posterior probability of an audio referring to the bus scene. This audio clip is misclassified as tram by the FCNN approach but correctly classified by the DSM system. This 10-second audio consists of the sound of talk, wind, and bus engine. Figure 2 shows that although the audio clip exists with short feature fragments that have strong directivity to the bus scene, the result is affected by

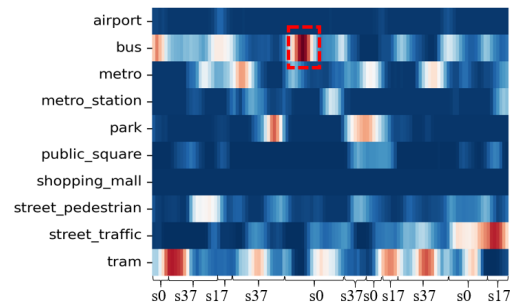


Figure 2: The frame-level posterior probability output by the FCNN classifier and the ASM sequence of a sample from bus scene. This sample is misclassified by FCNN as the tram scene but correctly classified by the DSM approach.

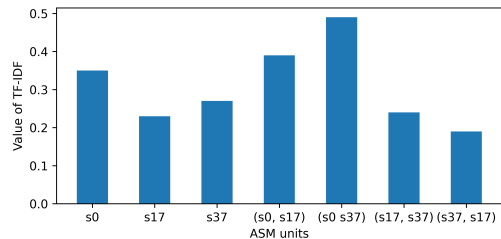


Figure 3: The TF-IDF value of several ASM units in the sample

some long-term interference factors that are similar to the tram scene. This phenomenon leads to the misclassification of the FCNN approach. Moreover, the decoded ASM sequence of this audio sample is shown with the explicit segment at the bottom of figure 2. The audio sample is transcribed by the dictionary of 40 ASM units from  $s_0$  to  $s_{39}$ . TF-IDF describes the relevance between the ASM unit and the audio clip [27]. Several uni-gram and bi-gram ASM units with high TF-IDF values in the ASM sequence of the bus sample are shown in Figure 3. Different ASM units have different contributions to the audio clip. 's0' and the bi-gram units of 's0' achieve higher values which indicates that the relevant segment features are paid more attention. In the red dotted box, the bus scene achieves the most important characteristics and is transcribed as the 's0' unit. Thus, DSM pays more attention to this part and correctly classifies the audio clip.

## 4. Conclusions

In this paper, we proposed a novel deep segment model (DSM) to take full advantages of FCNN and ASM approach. DSM employs a fully convolutional neural network (FCNN) as a deep feature extractor and then guides the ASM to better capture semantic information among sound events. Specifically, the FCNN-based encoder is trained with the multi-task of classifying both three coarse-grained acoustic scenes and ten fine-grained acoustic scenes to extract multi-level acoustic features. Moreover, an entropy-based decision fusion strategy is designed to further utilize the complementarity of FCNN-based and DSM-based systems. The final system achieves an accuracy of 80.4% in the DCASE2021 Task1b audio dataset.

## 5. Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant No. 62171427.

## 6. References

- [1] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," in *24th European Signal Processing Conference 2016 (EUSIPCO 2016)*, Budapest, Hungary, 2016.
- [2] —, "A multi-device dataset for urban acoustic scene classification," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*, November 2018, pp. 9–13.
- [3] 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.
- [4] 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 Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 379–393, 2017.
- [5] S. Ntalampiras, "A novel holistic modeling approach for generalized sound recognition," *IEEE Signal Processing Letters*, vol. 20, no. 2, pp. 185–188, 2013.
- [6] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, "Acoustic event detection in real life recordings," in *2010 18th European signal processing conference*. IEEE, 2010, pp. 1267–1271.
- [7] H. Chen, Z. Liu, Z. Liu, P. Zhang, and Y. Yan, "Integrating the data augmentation scheme with various classifiers for acoustic scene modeling," DCASE2019 Challenge, Tech. Rep., June 2019.
- [8] Y. Wu and T. Lee, "Enhancing sound texture in cnn-based acoustic scene classification," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 815–819.
- [9] S. S. R. Phayre, E. Benetos, and Y. Wang, "Subspectralnet – using sub-spectrogram based convolutional neural networks for acoustic scene classification," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 825–829.
- [10] L. D. Pham, I. V. McLoughlin, H. Phan, and R. Palaniappan, "A robust framework for acoustic scene classification," in *INTER-SPEECH*, 2019, pp. 3634–3638.
- [11] M. D. McDonnell and W. Gao, "Acoustic scene classification using deep residual networks with late fusion of separated high and low frequency paths," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 141–145.
- [12] S. Mun, S. Shon, W. Kim, and H. Ko, "Deep neural network bottleneck features for acoustic event recognition," in *Interspeech*, 2016, pp. 2954–2957.
- [13] S. Mun, S. Shon, W. Kim, D. K. Han, and H. Ko, "Deep neural network based learning and transferring mid-level audio features for acoustic scene classification," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 796–800.
- [14] X. Bai, J. Du, J. Pan, H.-s. Zhou, Y.-H. Tu, and C.-H. Lee, "High-resolution attention network with acoustic segment model for acoustic scene classification," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 656–660.
- [15] J. Wang and S. Li, "Self-attention mechanism based system for dcase2018 challenge task1 and task4," *Proc. DCASE Challenge*, pp. 1–5, 2018.
- [16] Z. Ren, Q. Kong, K. Qian, M. D. Plumbley *et al.*, "Attention-based convolutional neural networks for acoustic scene classification," in *DCASE 2018 Workshop Proceedings*. University of Surrey, 2018.
- [17] S. Mun, S. Park, D. Han, and H. Ko, "Generative adversarial network based acoustic scene training set augmentation and selection using SVM hyper-plane," DCASE2017 Challenge, Tech. Rep., September 2017.
- [18] H. Hu, C.-H. H. Yang, X. Xia, X. Bai, X. Tang, Y. Wang, S. Niu, L. Chai, J. Li, H. Zhu, F. Bao, Y. Zhao, S. M. Siniscalchi, Y. Wang, J. Du, and C.-H. Lee, "A two-stage approach to device-robust acoustic scene classification," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 845–849.
- [19] C.-H. Lee, F. K. Soong, and B.-H. Juang, "A segment model based approach to speech recognition," in *ICASSP-88., International Conference on Acoustics, Speech, and Signal Processing*. IEEE Computer Society, 1988, pp. 501–502.
- [20] H. Li, B. Ma, and C.-H. Lee, "A vector space modeling approach to spoken language identification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 271–284, 2006.
- [21] H.-y. Lee, T.-y. Hu, H. Jing, Y.-F. Chang, Y. Tsao, Y.-C. Kao, and T.-L. Pao, "Ensemble of machine learning and acoustic segment model techniques for speech emotion and autism spectrum disorders recognition," in *INTERSPEECH*, 2013, pp. 215–219.
- [22] M. Riley, E. Heinen, and J. Ghosh, "A text retrieval approach to content-based audio retrieval," in *Int. Symp. on Music Information Retrieval (ISMIR)*, 2008, pp. 295–300.
- [23] X. Bai, J. Du, Z.-R. Wang, and C.-H. Lee, "A hybrid approach to acoustic scene classification based on universal acoustic models," in *Interspeech*, 2019, pp. 3619–3623.
- [24] T. Heittola, A. Mesaros, and T. Virtanen, "Acoustic scene classification in dcase 2020 challenge: generalization across devices and low complexity solutions," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, 2020, pp. 56–60.
- [25] A. Moore, "K-means and hierarchical clustering," 2001.
- [26] D. Hull, "Improving text retrieval for the routing problem using latent semantic indexing," in *SIGIR '94*. Springer, 1994, pp. 282–291.
- [27] J. Ramos *et al.*, "Using tf-idf to determine word relevance in document queries," in *Proceedings of the first instructional conference on machine learning*, vol. 242, no. 1. Citeseer, 2003, pp. 29–48.
- [28] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.
- [29] S. Wang, A. Mesaros, T. Heittola, and T. Virtanen, "A curated dataset of urban scenes for audio-visual scene analysis," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, accepted. [Online]. Available: <https://arxiv.org/abs/2011.00030>
- [30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [31] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010*. Springer, 2010, pp. 177–186.