

An Acoustic Segment Model Based Segment Unit Selection Approach to Acoustic Scene Classification with Partial Utterances

Hu Hu¹, Sabato Marco Siniscalchi^{2,1}, Yannan Wang³, Xue Bai⁴, Jun Du⁴, Chin-Hui Lee¹

¹School of Electrical and Computer Engineering, Georgia Institute of Technology, USA

²Computer Engineering School, University of Enna, Italy

³Tencent Media Lab, Tencent Corporation, Shenzhen, Guangdong, China

⁴University of Science and Technology of China, HeFei, China

huhu@gatech.edu, marco.sinsalchi@unikore.it, yannanwang@tencent.com,
byxue@mail.ustc.edu.cn, jundu@ustc.edu.cn, chl@ece.gatech.edu

Abstract

In this paper, we propose a sub-utterance unit selection framework to remove acoustic segments in audio recordings that carry little information for acoustic scene classification (ASC). Our approach is built upon a universal set of acoustic segment units covering the overall acoustic scene space. First, those units are modeled with acoustic segment models (ASMs) used to tokenize acoustic scene utterances into sequences of acoustic segment units. Next, paralleling the idea of stop words in information retrieval, stop ASMs are automatically detected. Finally, acoustic segments associated with the stop ASMs are blocked, because of their low indexing power in retrieval of most acoustic scenes. In contrast to building scene models with whole utterances, the ASM-removed sub-utterances, i.e., acoustic utterances without stop acoustic segments, are then used as inputs to the AlexNet-L back-end for final classification. On the DCASE 2018 dataset, scene classification accuracy increases from 68%, with whole utterances, to 72.1%, with segment selection. This represents a competitive accuracy without any data augmentation, and/or ensemble strategy. Moreover, our approach compares favourably to AlexNet-L with attention.

Index Terms: acoustic scene classification, acoustic segment models, stop words detection, convolutional neural network

1. Introduction

The aim of the acoustic scene classification (ASC) refers to the task of identifying real-life sounds into environment classes, such as metro station, street traffic, or public square. An acoustic scene sound contains much information and rich content, which makes accurate scene prediction difficult. ASC has been an attracting research field for decades, and the IEEE Detection and Classification of Acoustic Scenes and Events (DCASE) challenge [1, 2, 3] provides the benchmark data and a competitive platform to promote sound scene research and analyses. In recent years, we have witnessed that the deep neural networks (DNNs) have gradually dominated the design of top ASC systems, and the main ingredient of their success is the application of deep convolutional neural networks (CNNs) [4, 5, 6, 7]. Furthermore, with the use of advanced deep learning techniques, such as attention mechanism [8, 9, 10] and deep network based data augmentation [11, 12, 13], a further boost in ASC system performances can be obtained.

In this study, we leverage upon acoustic segment models (ASMs) as an indicator of the indexing power of the input audio segment units with respect to the acoustic scenes being classified. A set of ASM models is employed to carry out acoustic

segment selection in the front-end. An initial ASM sequence for each given audio recording is obtained by unsupervised segmentation and clustering. Next, we use Gaussian mixture model (GMM)- or deep neural network (DNN)- hidden Markov model (GMM/DNN-HMM) [14, 15] to model the ASM sequences in a semi-supervised manner. Thus, each audio recording is segmented (tokenized) into a sequence of acoustic segment units, each having its relative ASM unit index. In this work, the terms acoustic segment units, acoustic segments, sub-utterance units, and tokens, are used interchangeably. A similar strategy to detect the stop words in information retrieval [16] is adapted, and a set of stop ASMs is identified using the training data. Stop ASMs represent meaningless ASM units, which carrying very low indexing power in retrieving most acoustic scenes. Just like word 'the', 'an' and 'or' in document retrieval problems, stop ASMs are therefore not useful to identify the target scene class. Those stop ASMs are used at a front-end level to block all of audio segments consisting of sequences of acoustic frames, belonging to those stop ASM models in both the training and evaluation stages. In doing so, noisy acoustic segments are eliminated in building models at the training stage, and not sent to the back-end acoustic scene classifier during the classification stages. The proposed approach is evaluated on DCASE 2018 Task 1a data set. Our experiments demonstrate that our solution improves an AlexNet like system, dubbed AlexNet-L, boosting the classification accuracy from 68.0% to 72.1%. The latter is a competitive result since neither data augmentation nor system combination are used. Furthermore, our segment-selection scheme with ASMs compares favourably with a recently proposed CNN classifiers using an ASM-based attention mechanism [17].

2. Related Work and Our Contributions

Several approaches based on feature learning have been proposed for the ASC task. For example, low-level features [18, 19, 20], which are directly extracted from the input signal at the front-end level, are thoroughly investigated to boost ASC performance. With deep models, mid-level features [21, 22, 23, 24] are instead induced from a DNN hidden layer, which takes into account the overall information embedded in training set. In addition to these low-level or mid-level features, the use of raw waveform to feed end-to-end systems has also been investigated in [25, 26]. However, to the best of the authors' knowledge, all of those feature learning works on ASC employ the whole input audio recording at the input layer to obtain high-dimensional feature vectors, and there aren't investigations concerned with

segment selection at a front-end level.

From a human listening perspective, sound recognition is often guided by detecting prominent acoustic events and/or audio cues useful to identify particular acoustic scenes [27]. For example, human listeners may leverage upon a car horn sound to determine that it is from a street traffic scene, or a loud plane engine sound to determine it is from an airport. Those sounds generated from car horns and plane engines, have stronger indexing power than other sounds for classifying these two acoustic scenes. Hence we argue that we are bound to get better ASC accuracy if we can block acoustic segments with little indexing power. Our idea could be related to an attention mechanism [8, 9, 10, 17], which uses an ad-hoc internal connectionist block and a huge amount of data to weight hidden internal representations accordingly to its saliency to the target outputs. However, an attention mechanism requires extra amount of parameters, and the performance highly depends on the model tuning. Our approach introduces a well-known approach from the information retrieval field [16] to detect meaningless sound events, and the experimental evidence confirms our claim.

In order to find the semantic salience of sound events, we use ASMs. ASMs are a set of self-organized sound units that are intended to cover the overall acoustic characteristics using available training data [28]. The ASM framework has recently been adopted in many audio sequence classification tasks, such as language identification [29], speaker identification [30], emotion recognition [31], music genre classification [32] and the acoustic scene classification [33]. As for ASC, it makes the assumption that the acoustic characteristics of all scenes can be covered by a universal set of acoustic units. Thus, input audio recordings can be transformed into ASM sequences, which are in turn processed by latent semantic analysis (LSA) [34] to obtain feature vectors with semantic information. Finally, a CNN based ASC system with an attention mechanism using ASM units is proposed in [17]. Different from the conventional ASM framework, in this work, ASM sequences are not used for a follow-up feature extraction process. In the experimental section, we demonstrated that our front-end solution outperforms that with the attention mechanism in [17].

3. Acoustic Segment Modeling

Like the phoneme representation for the speech utterance, we assume that the sound characteristics of acoustic scenes can also be covered by a universal set of acoustic units. The ASM approach aims to build a tokenizer to transfer the scene audio into a sequence of ASMs, i.e., the acoustic units specified in an acoustic inventory. The ASM sequence is generated in two main steps: (i) an unsupervised approach is used to seed the initial ASMs, with each acoustic unit having a fixed length (acoustic segment), (ii) either a GMM-HMM or DNN-HMM system is built on top of the initial ASMs and then used to generate the ASM sequence for a given audio recording.

3.1. Initial ASM Sequence Generation

ASM initialization is a critical factor for the success of the ASM framework. The initial ASM sequence generation is performed at a feature level, i.e. log-mel filter bank (LMFB) energies, or mel frequency cepstral coefficients (MFCCs). First, a given input audio recording is divided into a sequence of fixed-length segments. In our experiments, an audio recording is split into 50 acoustic segments, each having 20 LMFB or MFCC frames. The arithmetic mean of these 20 frames is used to generate a

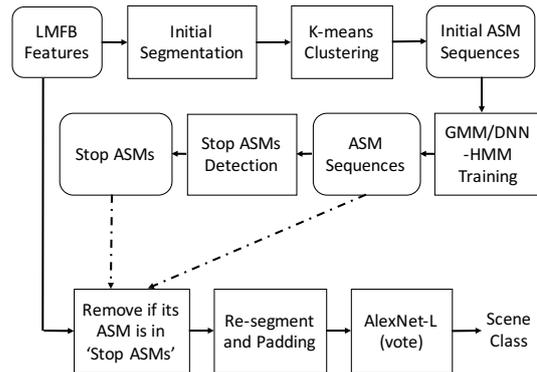


Figure 1: The framework of proposed ASM-guided segment selection approach for ASC.

single feature vector representing the whole segment. Next, all generated feature vectors for the training material are used with the K-means clustering algorithm to find a set of centroids. Audio segments are grouped into a small number of acoustic classes (each class is an ASM), to represent the whole acoustic space scattered by the training data. We do not leverage any prior knowledge when building our acoustic inventory; therefore, the set of ASMs and its corresponding model arises in an unsupervised manner. Finally, the centroids can be used to map a given audio recording into an initial ASM sequence. Each initial ASM sequence has a fixed number of 50 ASM units, since we have split the input recording into 50 fixed-length segments.

3.2. GMM/DNN-HMM Training

The initial ASMs can provide a rough segmentation of the audio scene, where segments have fixed-length in terms of the number of audio frames, which does not adhere to real scenarios. A finer segmentation result can be obtained leveraging the HMM framework, in which the state probability density functions (pdfs) can be obtained with a GMM or DNN. We therefore deployed both GMM-HMM and DNN-HMM tokenizers as follows:

1. Seed a GMM-HMM for each ASM unit based on the initial ASM segmentation.
2. Use the model obtained in Step 1 and perform Viterbi decoding on all the training utterances.
3. Train the model with the new transcriptions generated in Step 2.
4. Repeat Step 2 and Step 3 until convergence.
5. Build a DNN-HMM using the GMM-HMM and new ASM segmentation.

4. Front-end Segment Selection via ASM

For acoustic scenes, differently from spoken utterances, only few real meaningful segments characterize the whole scene. For example, a car horn sound is an import sign to determine a street traffic scene, but many other segments in that acoustic scene do not carry any key information to make the correct classification. However, it's not easy to detect meaningful segments for most of the acoustic scenes. Therefore, more and more ASC systems simply use a CNN end-to-end approach to learn the mapping from the input audio recordings to the output scene class.

Nonetheless, CNNs are good at extracting local features but not for overall segment selection. ASMs are used in our work to find useful acoustic segments, which have high indexing power with respect to the target acoustic scene. If useless segments can be removed in the front-end processing stage, that will be beneficial to the back-end classifier, as proven in the experimental section.

The proposed ASM-based front-end segment selection framework for sub-utterance acoustic scene classification is shown in Figure 1. The dashed lines indicate where ASM sequences and stop ASMs are used. Stop ASMs, and ASM sequences are generated in the segment selection block to remove consecutive feature frames not useful for final scene classification. The stop rules and segment selection steps are described in detail in the following section.

4.1. Stop ASMs Detection

Stop ASMs, as the name reveals, takes inspiration from stop words in information retrieval [16]. Given the inventory of ASMs, stop ASMs are a subset of the original ASMs that does not carry much information for retrieving the target acoustic scenes. Compared with other ASMs in the inventory, stop ASMs have either a lower indexing power, or no indexing power at all when it comes to make the final classification decision. We use D to denote the total number of ASMs in the inventory, and N to denote the total number of utterances in the training set. In this study, we consider four different methods to detect the stop ASMs [35, 36, 37]:

- Mean Probability (MP): it is the average probability of each unit M_j in ASM sequences of the training set. MP considers the frequency of each ASM and is calculated by

$$MP(M_j) = \frac{\sum_{i=1}^N P_{i,j}}{N}, \quad (1)$$

in which $P_{i,j}$ is the probability of the ASM unit M_j in utterance U_i , and calculated by dividing its frequency by the total number of the ASMs in U_i .

- Inverse Document Frequency (IDF): it measures how much information the ASM provides to reflect the importance of each ASM. IDF is calculated by

$$IDF(M_j) = \log \frac{N+1}{N_j+1}, \quad (2)$$

in which N_j is total number of times the ASM unit M_j appears in the training utterances.

- Variance of Probability (VP): it considers the variance of each ASM unit. VP is calculated by

$$VP(M_j) = \frac{\sum_{i=1}^N (P_{i,j} - MP(M_j))^2}{N}. \quad (3)$$

- Statistical Values (SATs): this metric considers both the mean and the variance. If an ASM unit has high SAT values, it implies that M_j occurs frequently and uniformly in all the training utterances, and M_j is very likely to be a stop ASM. SAT is calculated by

$$SAT(M_j) = \frac{MP(M_j)}{\sqrt{VP(M_j)^{1/2}}}. \quad (4)$$

In our experiments, we select the top P ASMs for functioning as stop ASMs. The segments whose ASMs are in stop ASMs dominate in most utterances.

4.2. Data Processing with Stop ASMs

The front-end processing is performed with stop ASMs and ASM sequences. As shown in Figure 1, for a given input with an ASM sequence, the corresponding LMFB feature frames will be removed if their ASMs are stop ASMs. The remaining feature fragments will be re-segmented and padded into a group of fixed-length acoustic segments. In our experiments, if a fragment is more than 20 frames, we will divide it into segments with the length of 20 frames, otherwise, we will perform zero-padding to make it having 20 frames. Hence, each acoustic scene would eventually have a different number of segments, which depends on the number of frames that have been blocked at the front-end level. After re-segmenting and padding process, the new generated fixed-length segments are fed into the back-end classifier. Each segment is assigned to a scene class by AlexNet-L back-end classifier, and the final scene class is obtained via majority voting among all segments.

5. Experiments and Result Analysis

5.1. Experimental Setup

The proposed approach is evaluated on the DCASE 2018 Task1a development data set [3]. It contains 24 hours of acoustic scene audio recorded with the same device at a 48kHz sampling rate in 10 different acoustic scenes. Following the official recommendation, the development data set is divided into training and test sets containing 6122 and 2518 utterances, respectively. For each 10-second binaural signal, STFT with 2048 points is applied separately on the left and right channels, with a window length of 25ms and an overlap length of 10ms. Mel filter-banks with 128 bins are applied to obtain the log-mel filter bank (LMFB) features. Our ASC baseline system is based on the AlexNet [38] model. Nonetheless, different from the original AlexNet, we reduce the parameter size due to internal resource constrains. The baseline is denoted as AlexNet-L in the rest of this work. It has five convolutional layers with a kernel size of 4×4 , and two fully connected layers with hidden dimension of 1024. Each layer consists of convolution, batch normalization, ReLU activation function and max pooling. AlexNet-L is trained with a stochastic gradient descent (SGD) algorithm with a cosine based learning rate scheduler. Each input utterance is segmented to 20 frames (0.2 seconds per segment). After AlexNet-L classification, the final scene class of the input waveform is voted by a majority using the classification result on each segments.

In the acoustic segment modeling stage, the initial segment length is set to 20 frames, which is the same with the segment length in the baseline AlexNet-L. Hence each utterance is divided into 50 segments. The size D of the ASM inventory is set to 64. According to our experiments, D is robust to the parameter setting. GMM-HMMs are powerful for modeling sequential data, and we here to refine the initial tokenization phase. A left-to-right HMM topology is used in each 6-state GMM-HMM. In DNN-HMM, the DNN has six hidden layers, each having 2048 neurons. The output layer estimates the state probability density function (pdf) of the 64 HMMs. During the stop ASMs detection, top-3 ASMs function as stop ASMs for each metric discussed in Section 4.1. After removing segments whose ASMs are in the stop ASMs set, the remaining acoustic fragments are re-segmented and padded to obtain 20 acoustic frames per segment. As done with AlexNet-L, majority voting is used to decide the final scene class.

5.2. Stop ASMs Detection Results

The stop ASMs detection result is shown in Table 1. Different detection criteria lead to a different set of stop ASMs. We selected the top three ASMs according to highest MP, lowest IDF, lowest VP and highest SAT, as our stop ASMs. From Table 1, we can notice that some ASMs, such as M_8 , M_{14} , appear independently of the metrics. The latter implies that some acoustic segments do satisfy our assumption, that is, there are acoustic segments having low or no indexing power for scene classification. Stop ASMs found by GMM-HMM and DNN-HMM are similar. When using the SAT criterion, the same set of stop ASMs is found independently of the tokenizer. However, although different metrics can select the same stop ASMs, the technique to obtain ASMs would lead to different segment boundaries, which eventually affects final classification results.

Table 1: Stop ASMs detection results with different metrics. M_i indicates i -th ASM in the ASM inventory. Initial ASMs imply that initial ASM sequences are used in stop ASMs detection. GMM-HMM and DNN-HMM refer to sequences obtained with the those models.

Metric	Initial ASMs	GMM-HMM	DNN-HMM
MP	M_8, M_{35}, M_{43}	M_{21}, M_{35}, M_{43}	M_8, M_{35}, M_{43}
IDF	M_8, M_{10}, M_{14}	M_8, M_{10}, M_{51}	M_8, M_{10}, M_{14}
VP	M_4, M_{39}, M_{53}	M_{39}, M_{42}, M_{56}	M_{39}, M_{55}, M_{56}
SAT	M_8, M_{14}, M_{52}	M_8, M_{14}, M_{52}	M_8, M_{14}, M_{52}

5.3. ASC Experimental Results

The proposed ASC system is shown in Figure 1, and related experimental results are given in Table 2. For a comprehensive evaluation, the high-resolution attention network with ASM (HRAN-ASM) system proposed in [17] is also implemented, and its classification results are reported. In particular, we have adopted the two attention modules with ASM embedding into our AlexNet-L baseline model. The first two rows in Table 2 are official baseline [3] and our AlexNet-L baseline. The AlexNet-L attains a classification accuracy of 68.0%, which is improved to 69.5% by with the HRAN-ASM in the third row. Although our experiments are conducted with the same data sets adopted in [17], experimental results are slightly different because of different speech features and models used in our work.

Table 2 lists experimental results obtained with the proposed approach in the last three rows when SAT is used as the metric to extract the stop ASMs. We can use our acoustic segment blocking approach with three different ASM tokenizers, namely (i) the initial unsupervised ASM (initial ASM), (ii) GMM-HMM, and (iii) DNN-HMM. In case (i), stop ASMs are first detected using the tokenization obtained with initial ASM. Since the segment length of initial ASM sequences and the input sequences to AlexNet-L are the same, we can directly block the whole segment of the input sequence if its corresponding ASM token is in the set of stop ASMs. Thus, the blocking operation is based on the segment level, in which the re-segmenting and padding is not needed on processed segments. Although initial ASMs are simple models, AlexNet-L accuracy can be boosted from 68.0% to 70.1% (compare first and third rows). In cases (ii) and (iii) with either GMM-HMM or DNN-HMM tokenizers, we can obtain more precise ASM sequences, which in turn improve the final ASC accuracy as shown in the last two rows of Table 2. In details, GMM-HMM boosts AlexNet-L classification up to 71.6%. DNN-HMM can deliver more accurate alignments and boundaries for each ASM segment, which

Table 2: Evaluation results on DCASE 2018 Task1a data set.

Model	Accuracy %
Official baseline [3]	59.7
AlexNet-L baseline	68.0
+ HRAN-ASM [17]	69.5
+ initial ASM (SAT)	70.1
+ ASM-GMM-HMM (SAT)	71.6
+ ASM-DNN-HMM (SAT)	72.1

leads to a final scene classification accuracy of 72.1%, which representing a 4.1% absolute improvement when compared to AlexNet-L.

The above results allow us to conclude that: (a) the ASM-based segment selection approach can significantly improve ASC system, attaining a final classification accuracy of 72.1%, which is a competitive performance given that data augmentation and ensemble methods have not been used in our work, and (b) the proposed solution outperforms AlexNet-L with the attention mechanism initialised with ASM, which the later is reported to compare favourably against self-attention in [17]. The latter demonstrates that our front-end segment selection approach outperforms a more standard attention scheme.

5.4. Evaluation with Different Detection Metrics

The effect of different stop ASMs metrics is shown in Table 3. The initial ASM sequences are used to evaluate different metrics. From Table 3, we can see that different stop ASMs detection metrics result in different classification outcomes. Using MP with initial ASM sequences does not lead to any improvement over AlexNet-L. SAT shows the best performance among all metrics. Those results makes sense since SAT considers both the mean and the variance of the distribution of each ASM unit.

Table 3: Results with different stop ASMs detection metrics using initial ASMs.

Model	Accuracy %
AlexNet-L baseline	68.0
+ initial ASM (MP)	68.0
+ initial ASM (IDF)	68.7
+ initial ASM (VP)	69.3
+ initial ASM (SAT)	70.1

6. Summary

In this paper, instead of using whole utterances for scene modeling, we propose an ASM based front-end segment selection approach to acoustic scene classification. The overall framework is based on two modules: (i) acoustic segment modeling and selection, and (ii) CNN based classification. ASMs are first generated in an unsupervised manner and refined with GMM/DNN-HMM models. Then stop ASMs detection is performed using ASM sequences for training data. ASM sequences and stop ASMs are used for segment selection before CNN classifier. It implies that segments with low or no indexing power are removed. The proposed approach is evaluated on DCASE 2018 Task1a, and experimental evidences demonstrate the viability of sub-utterance ASC. A classification accuracy of 72.1% is obtained, which is highly competitive for single system and no data expansion.

7. References

- [1] 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, 2018.
- [2] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "DCASE 2017 challenge setup: Tasks, datasets and baseline system," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, November 2017, pp. 85–92.
- [3] A. Mesaros, T. Heittola, and T. Virtanen, "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.
- [4] Y. Han and K. Lee, "Acoustic scene classification using convolutional neural network and multiple-width frequency-delta data augmentation," *arXiv preprint arXiv:1607.02383*, 2016.
- [5] D. Battaglino, L. Lepouloux, and N. Evans, "Acoustic scene classification using convolutional neural networks," DCASE2016 Challenge, Tech. Rep., September 2016.
- [6] K. Koutini, H. Eghbal-zadeh, M. Dorfer, and G. Widmer, "The receptive field as a regularizer in deep convolutional neural networks for acoustic scene classification," *CoRR*, vol. abs/1907.01803, 2019.
- [7] H. Hu, C.-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, "Device-robust acoustic scene classification based on two-stage categorization and data augmentation," DCASE2020 Challenge, Tech. Rep., June 2020.
- [8] J. Wang and S. Li, "Self-attention mechanism based system for dcase2018 challenge task1 and task4," in *Proc. DCASE Challenge*, 2018, pp. 1–5.
- [9] Z. Ren, Q. Kong, K. Qian, M. D. Plumbley, B. Schuller *et al.*, "Attention-based convolutional neural networks for acoustic scene classification," in *DCASE 2018 Workshop Proceedings*, 2018.
- [10] J. Guo, N. Xu, L.-J. Li, and A. Alwan, "Attention based cldnns for short-duration acoustic scene classification," in *INTERSPEECH*, 2017, pp. 469–473.
- [11] T. Nguyen and F. Pernkopf, "Acoustic scene classification with mismatched devices using cliquenets and mixup data augmentation," in *Interspeech*, 2019.
- [12] 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.
- [13] T. Zhang, K. Zhang, and J. Wu, "Data independent sequence augmentation method for acoustic scene classification," in *Interspeech*, 2018, pp. 3289–3293.
- [14] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.
- [15] 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.
- [16] W. J. Wilbur and K. Sirotkin, "The automatic identification of stop words," *Journal of information science*, vol. 18, pp. 45–55, 1992.
- [17] X. Bai, J. Du, J. Pan, Z. Heng-shun, Y.-H. Tu, and C.-H. Lee, "High-resolution attention network with acoustic segment model for acoustic scene classification," in *ICASSP*, 2020.
- [18] 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.
- [19] A. Rakotomamonjy and G. Gasso, "Histogram of gradients of time–frequency representations for audio scene classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 142–153, 2014.
- [20] V. Bisot, R. Serizel, S. Essid, and G. Richard, "Feature learning with matrix factorization applied to acoustic scene classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1216–1229, 2017.
- [21] 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 Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2016.
- [22] H. Zeinali, L. Burget, and J. Cernocky, "Convolutional neural networks and x-vector embedding for dcase2018 acoustic scene classification challenge," *arXiv preprint arXiv:1810.04273*, 2018.
- [23] P. Sharma, V. Abrol, and A. Thakur, "Ase: Acoustic scene embedding using deep archetypal analysis and gmm," in *Interspeech*, 2018, pp. 3299–3303.
- [24] J. Cramer, H.-H. Wu, J. Salamon, and J. P. Bello, "Look, listen, and learn more: Design choices for deep audio embeddings," in *ICASSP*. IEEE, 2019, pp. 3852–3856.
- [25] L. Pham, I. McLoughlin, H. Phan, and R. Palaniappan, "A robust framework for acoustic scene classification," *Proc. Interspeech 2019*, pp. 3634–3638, 2019.
- [26] J. Chen, J. Hao, K. Chen, D. Xie, S. Yang, and S. Pu, "An end-to-end audio classification system based on raw waveforms and mix-training strategy," *arXiv preprint arXiv:1911.09349*, 2019.
- [27] V. T. Peltonen, A. J. Eronen, M. P. Parviainen, and A. P. Klauri, "Recognition of everyday auditory scenes: potentials, latencies and cues," *PREPRINTS-AUDIO ENGINEERING SOCIETY*, 2001.
- [28] C.-H. Lee, F. K. Soong, and B.-H. Juang, "A segment model based approach to speech recognition," in *ICASSP*, 1988, pp. 501–541.
- [29] 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, pp. 271–284, 2006.
- [30] Y. Tsao, H. Sun, H. Li, and C.-H. Lee, "An acoustic segment model approach to incorporating temporal information into speaker modeling for text-independent speaker recognition," in *ICASSP*. IEEE, 2010, pp. 4422–4425.
- [31] 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.
- [32] 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.
- [33] X. Bai, J. Du, Z.-R. Wang, and C.-H. Lee, "A hybrid approach to acoustic scene classification based on universal acoustic models," *Proc. Interspeech 2019*, pp. 3619–3623, 2019.
- [34] T. K. Landauer, P. W. Foltz, and D. Laham, "An introduction to latent semantic analysis," *Discourse processes*, vol. 25, no. 2-3, pp. 259–284, 1998.
- [35] W. J. Wilbur and K. Sirotkin, "The automatic identification of stop words," *Journal of Information Science*, vol. 18, no. 1, pp. 45–55, 1992. [Online]. Available: <https://doi.org/10.1177/016555159201800106>
- [36] D. Na and C. Xu, "Automatically generation and evaluation of stop words list for chinese patents," *Telkommika*, vol. 13, no. 4, p. 1414, 2015.
- [37] R. T.-W. Lo, B. He, and I. Ounis, "Automatically building a stop-word list for an information retrieval system," in *Journal on Digital Information Management: Special Issue on the 5th Dutch-Belgian Information Retrieval Workshop (DIR)*, vol. 5, 2005, pp. 17–24.
- [38] 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.