INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

A Hybrid Approach to Acoustic Scene Classification Based on Universal Acoustic Models

Xue Bai[1], *Jun Du*[1], *Zi-Rui Wang*[1], *Chin-Hui Lee*[2]

[1]National Engineering Laboratory of Speech and Language Information Processing,
University of Science and Technologoy of China
[2]Georgia Institute of Technology

byxue@mail.ustc.edu.cn, jundu@ustc.edu.cn, cs211@mail.ustc.edu.cn,
chl@ece.gatech.edu

Abstract

For the acoustic scenes classification, the main challenge is distinguishing similar acoustic segments between different scenes. To solve this problem, many deep learning based approaches have been proposed without considering the relevance of different acoustic scenes. In this paper, we propose a novel acoustic segment model (ASM) for acoustic scene classification. ASM aims at giving finer segmentation and covering all acoustic scenes through searching for the underlying phoneme like acoustic units. Furthermore, acoustic segments are modeled by Hidden Markov Models (HMMs) and each audio is decoded into ASM sequences without prior linguistic knowledge. Similar to the term vector of a text document, these ASM sequences are converted into co-occurrence statistics feature vectors and SVM/DNN is used as classifier back-end. Validated on the DCASE 2018 task, the proposed approach can achieve a competitive performance with single model and no data augment. By using visualization analysis, we excavate the potential similar units hidden in auditory sense.

Index Terms: acoustic scene classification, acoustic segment models, hidden Markov models, latent semantic analysis

1. Introduction

Acoustic scene classification (ASC) is a task to identify sounds in realistic soundscapes. Acoustic scenes carry much useful information and its analysis has huge potentiality in several applications such as context-aware devices [1], audio based multimedia search [2] etc. Therefore, a significant amount of research on ASC has been investigated in recent years. However, there are several difficulties to develop practical ASC system. First, not all information in a piece of audio has a high degree of discrimination. Second, the sound events in recorded audio overlap (simultaneously occurring) and the boundary between them is often blurry. Last but not the least, samples in different scene categories may have commonalities, e.g. the similar speech components in most scenes.

Recently, many new techniques have emerged and been widely used for ASC, including traditional classifier methods such as Gaussian mixture models (GMM) [3, 4], hidden Markov models (HMM) [4] and deep learning based approaches: deep neural networks (DNN) [5], convolutional neural networks (CNN) [6], recurrent neural networks (RNN) [7], and convolutional recurrent neural networks (CRNN) [8]. While researchers have explored many different approaches for speech/audio processing, most state-of-the-art results in ASC task were obtained by CNN based methods. In DCASE2017 challenge [9], generative adversarial nets (GAN) scored first place by training data augmentation [10]. Sakashita [11] extractd mel-spectrogram from various channels, adaptively divided the spectrogram into multiple ways and learned 9 neural networks for ASC to obtain good score in DCASE2018 challenge [12]. Even though the previous methods have improved perfomance a lot, there still exist a lot of basic problems worth exploring. For example, the purpose of CNN is simply to learn the feature mapping relationship between input features and labels, which makes such kind of CNN-based approaches fail to capture the correlation of segments in different scenes.

To address the challenge of confusion in ASC, this paper presents acoustic segment model (ASM) framework. ASM was first proposed to characterize fundamental units and acoustic lexicons [13] for automatic speech recognition. Recently, ASM has also been applied to spoken language recognition [14], music retrieval [15] and emotion recognition [16] which achieves quite good performance. Just as language governs the syntax of phonemes and words, there are also fundamental units (acoustic events) and internal relevance in different acoustic scenes. Therefore, we make the assumption that the sound characteristics of all scene audios can be covered by a universal set of automatically derived acoustic units with no direct link to phonetic definitions.

To find the universal acoustic units, much of the proposed approach is inspired by previous work conducted by Jeremy Reed for music genre classification [17]. A typical ASM process involves two stages, namely initial segmentation and iterative modeling. In the initialization phase, a novel method has been proposed. Unlike the typical initial segmentation based on maximum likelihood segmentation [13], we model each class using GMM-HMMs [18] and divide a continuous scene audio into variable-length segments defined by the hidden states. In other words, the hidden state in each topology corresponds to a GMM and similar frames are clustered onto the same GMM which is regarded as the initial acoustic model for an ASM unit. These initial acoustic models are then used for scene audios so as to generate initial label sequences. Specially, these ASM units are generated in a data-driven way without any prior linguistic knowledge. For iterative modeling, each ASM is often modeled by a GMM-HMM and then scene audios are decoded into a sequence of acoustic units. The transcription of each audio is similar to the term vector of a text document. Naturally, by using lattent semantic analysis (LSA) [19] , we convert the ASM sequences to co-occurrence statistics feature vectors, which are then fed the final classifiers. This approach provides an initial foundation for future improvements through the use of "scene language"-based rules.

The remainder of this paper is organized as follows. In Section 2, we discussed the method of the hybrid approach. In Section 3, experimental results and and analysis are discussed.

http://dx.doi.org/10.21437/Interspeech.2019-2171

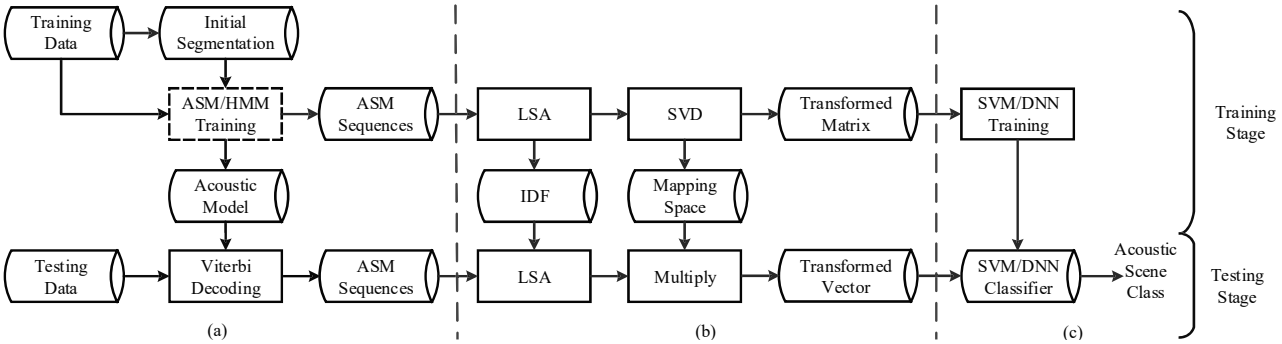

Figure 1: *Overall framework of ASC system with ASM acoustic units. (a)Universal acoustic modeling approach. (b)Latent semantic analysis. (c)Vector-based classifiers*

Finally, our conclusions are given in Section 4.

2. Method

The ASC framework based on universal acoustic models is illustrated in Figure 1. The purpose of the ASM method is to find a dictionary of ASM units and transcribes each acoustic scene recording into ASM sequences. Incidentally, the acoustic model generated during the iteration is used to transcribe the testing data. The final transcripts (ASM sequences) can be regarded as text symbols which are used to characterize each scene recording. And the validated text categorization methods commonly applied in the information retrieval community can be used to handle such symbol formats, such as LSA. While the transcribed sequence is converted to a vector through LSA and singular value decomposition (SVD) [20], the entire training data set is mapped to a matrix. The test scene audios are converted in a similar fashion to create the test vectors. In this study, we compare the support vector machine (SVM) and DNN for vector-based classifier design.

2.1. Acoustic Segment Modeling

The key idea of using acoustic segment models here is to represent an acoustic scene recording as a temporal cascade of basic sound units, just like that sentences are made up of phonemes and words. The ASM approach consists of two stages. First, the initial segmentation stage explores the boundaries between the changes of acoustic features. Instead of using the vector quantization (VQ) [13], we use a novel 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 transcribe each scene audio to a sequence of units. Second, the universal set of units called ASMs are represented and estimated by HMMs.

2.1.1. Initial Segmentation

The initial segmentation is a critical procedure to the success of ASM. To find an appropriate set of segments for each scene audio, many meaningful segmentation methods have been proposed. The GMM-HMMs are powerful for modeling sequential data and have been widely applied in automatic speech recognition. Therefore, we use the GMM-HMMs for initial segmentation and clustering of audio frames by the hidden chain.

First, each acoustic scene is modeled with a GMM-HMM model. As the similar sound events might occur in different time periods, we adopt a left-to-right HMM topology with a swivel structure in each GMM-HMM, which is able to cluster the similar frames into the same hidden state. Mel-frequency cepstral coefficients (MFCCs) are adopted as the acoustic features. Suppose we have M acoustic scene classes and each scene class is model by a GMM-HMM with N hidden states. The parameters of GMM-HMMs can be learned by the Baum-Welch algorithm [21]. In the decoding stage, one scene audio recording is represented by a sequence of hidden states with the corresponding segment for each decoded hidden state. Finally, all $J = M \times N$ hidden states with acoustic segments are collected together as the initialization of ASMs.

2.1.2. ASM/HMM Training

By the initial segmentation, each scene audio is transcribed as a sequence of ASMs. In order to capture the correlation between different acoustic scenes, it is desired to model each ASM with a left-to-right HMM using iterative training procedure. Then Baum-Welch estimation is applied to update the parameters of 3-state HMMs using the training data. After the estimation, the Viterbi decoding is performed to re-transcribe the training recordings to new ASM sequences, which are regarded as new labels/transcriptions for the training recordings and used to train the HMMs in the next iteration. This process is repeated until the training data labels are stably converged. The whole training procedure is summarized in Algorithm 1.

Algorithm 1 Procedure of ASM Training

Initial Segmentation
- **Step 1:** Each scene class is model by a GMM-HMM.
- **Step 2:** ASM units are initialized by all hidden states.
- **Step 3:** Transcribe the training recordings as T_0.

ASM/HMM Training
- **Step 4:** Each ASM is modeled by a GMM-HMM and $i = 0$.
- **Step 5:** Update GMM-HMM parameters using T_i.
- **Step 6:** Re-transcribe the training recordings into T_{i+1}.
- **Step 7:** i++, go to **Step** 5 for next iteration until converged.

2.2. Latent Semantic Analysis

After each scene audio is transcribed into a sequence of ASMs, we characterize each sequence by a vector, which uses the text vectorization techniques. The dimension of these vectors is equal to the total number of useful features based on unigram and bigram counts. Moreover, their co-occurrences could be seen as a rough syntax in acoustic scenes and reflect connections

to internal sound events. Although, the usage of bigram counts can achieve the expectation of the discrimination capability of feature vectors, the sparsity problem will arise. In general, SVD is used for the feature dimension reduction.

LSA has been successfully applied to information retrieval, question and answer systems and clustering in the field of text processing. In this work, we use LSA to extract features from the term-document matrix of the training set. The rows correspond to the transcribed scene audio recordings and the columns of this term-document matrix are related with the ASM units. Like lexical constraints, the constrains of acoustic segments can be typically described by the ASM n-grams. Suppose there is an ASM transcription (S_1, S_2, S_3). The statistics of unigram terms are derived form S_1, S_2, S_3 and the statistics of bigram terms are derived from (S_1, S_2), (S_2, S_3) to account for left and right contexts. If there are J terms in an scene corpus and assume all unigrams and bigrams exist, each column is a vector with the dimension of $K = J \times (J + 1)$.

In the text retrieval, the term counts are often composed of two parts: term frequency (TF) and inverse document frequency (IDF) [22]. The former is the frequency of occurrence of individual word in the text and the latter reflects the frequency of a word in all text. For example, the "a" word that appears in the text is high, but these words are less informative than that of professional vocabulary. Therefore, IDF is to help us to reflect the importance of the word, and then to correct the word eigenvalues expressed only by word frequency. The TF of ASM term j in the i-th scene transcript is given by

$$TF_{j,i} = \frac{c_{j,i}}{\sum_{k=1}^{K} c_{k,i}} \tag{1}$$

where $c_{j,i}$ is the count of j in ASM transcription i. The IDF is given by

$$IDF_j = \log \frac{L+1}{L(j)+1} \tag{2}$$

where L is the number of training scene transcripts and $L(j)$ is the total number of times that ASM unit j appears in the training scene transcripts. Finally, each element in the matrix W is given by

$$w_{j,i} = TF_{j,i} \times IDF_j. \tag{3}$$

The term-document matrix W calculated by TF-IDT is quite sparse due to the sparsity problem of bigrams. The dimension of vector K is determined by the number of unigrams and bigrams. Therefore, singular value decomposition is used for dimensionality reduction of the $K \times L$ matrix W. Define the SVD of matrix W as

$$W = U \Sigma V^T. \tag{4}$$

The matrix W is decomposed into the product of three matrices: the left-singular $K \times K$ matrix U, the right-singular $L \times L$ matrix V and the diagonal $K \times L$ matrix Σ consisting of singular values of W. The rank of W is R ($R \leq \min(K, L)$) and U describes an orthonormal basis in the domain. The first r largest singular values and the first r rows of their corresponding U matrices are taken out to form a mapping space U_r ($r \times K$). The mapping space converts the original matrix W into a lower-dimensional "concept" space by $W_r = U_r \times W$ and W_r serves as the training data for the vector-based classifier. The value of r is determined by the percentage of sum of squares of the singular values we need.

In the testing stage, we first perform LSA using the TF values calculated based on Eq. (1) and the decoded ASMs of one testing recording. Please note that IDF values calculated in the training stage are directly adopted here. Then a new term-document matrix W^{test} of one testing recording is actually a K-dimensional vector ($L = 1$). The final vector fed to the classifier is generated by $U_r \times W^{\text{test}}$.

2.3. Vector-Based Classifiers

In this study, two vector-based classifiers, namely SVM and DNN, are investigated to generate the acoustic scene classes. The multi-class SVM is based on one-against-one approach, in which suppose there are Q classes, $Q(Q-1)$ independent binary SVM classifiers are trained for all pairs of categories. The posterior probabilities are reckoned by the algorithm built in libSVM [23]. In addition, DNN has proven to be more efficient in variety of multi-class problem. The back-propagation algorithm with stochastic gradient decent method is used to update DNN parameters. In this study, we design a simple DNN structure to classify the vectors due to the limited training data.

3. Experiments and Analysis

3.1. Dataset and Feature Extraction

The experiments are conducted on DCASE2018 Task 1A [24], 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 audio recording into mono recoding. Then, 60-dimensional MFCC features including the corresponding delta and delta-delta features are extracted by applying 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 includes 6122 segments while the test subset has 2518 segments.

3.2. Experiments on Different Settings

In this subsection, we explore the different configurations of the proposed hybrid approach. First, we use hidden states of GMM-HMM of each scene to achieve initial segmentation and find the generic ASM units. Then, each ASM unit is modelled by a left-to-right HMM with 3 states. Each state has 50 Gaussian mixtures. By iteration, the new transcription of every scene audio and the corresponding ASM units are created. Second, LSA and SVD are performed on the entire training dataset to obtain the IDF and mapping space, respectively, which can be directly adopted in the testing stage. Finally, the transformed features are fed into classifiers. The SVM employs the "one-against-one" voting scheme [25] to classify acoustic scenes. The DNN used here has three hidden layers and every hidden layer has 512 neurons with the fixed dropout rate 0.2. The parameters of DNN are learned by using the SGD [26] algorithm. The initial learning rate is set to 0.1 and 70 epochs are conducted. Based on the DNN classifier, we discuss the following critical issues: 1) ASM resolution; 2) dimensionality reduction in SVD.

3.2.1. ASM Resolution

The number of ASM units reflects the acoustic coverage in terms of characterizing the sound space. Depending on the nature of acoustic scene, too few ASM units are not sufficient to represent their internal variations while too many ASM units lead to a large computational complexity and overfitting problem due to limited training data. In order to explore the impact of the number of ASM units on the classification results, Table 1

lists the experimental results for different ASM units ranging from 10 to 40. All systems use 80% of the singular values in SVD to extract input vector of DNN. From Table 1, it is clear to observe that a middle-level resolution with 20 ASM units is conductive to correctly comprehend.

Table 1: *Performance comparisons with different ASM units.*

ASM units	10	20	30	40
Accuracy	61.6%	66.1%	64.5%	62.7%

3.2.2. Dimensionality Reduction in SVD

In our proposed approach to extract the matrix from ASM sequences, all information are useful. Besides, it leads to a sparse matrix. One common technique is to retain only the top singular values in a matrix. In other words, the input dimension for vector-based classifier is determined by the percentage of sum of squares of the singular values. In Table 2, we discuss the effects of retained information in SVD by controlling the percentage. These experiments take 20 ASM units and corresponding dimension of each vector is 401 before dimension reduction. Please note that $K = 401$ is smaller than $20*21 = 420$ as some bigrams do not exist due to the limited training data. The results show that a proper dimension reduction (80%) can reduce data redundancy and make different scenes more distinguishable.

Table 2: *Performance comparisons with different reduced dimensions in SVD for vector-based classifier.*

Percentage	100%	90%	85%	80%	75%
Accuracy	65.0%	65.8%	65.7%	66.1%	65.8%

3.2.3. Overall Comparison

As shown in Table 3, compared with the CNN-based approach officially provided by DCASE2018 [27], the proposed "Hybrid-DNN" approach using a simple DNN architecture for vector-based classifier can obtain a remarkable improvement of accuracy from 59.7% to 66.1%. From the results of "Hybrid-SVM" and "Hybrid-DNN", DNN demonstrates its superiority over SVM for text-based classifier with a significant performance gap of 12.8%, which also inspires us to investigate more advanced text-based classifier in the future work.

Table 3: *Overall comparison of different approaches.*

System	CNN [27]	Hybrid-SVM	Hybrid-DNN
Accuracy	59.7%	53.3%	66.1%

3.3. Results Analysis

In Figure 2 and Figure 3, the CNN-based system [27] confuses between the *tram* and *bus* scenes, but our Hybrid-DNN system can distinguish these two scenes. For visual analysis, we show the decoded ASM sequences of examples from *tram* and *bus* scenes in Figure 2 and Figure 3 using the Hybrid-DNN approach with the best configuration mentioned above. Moreover, each decoded ASM unit is accompanied by the explicit

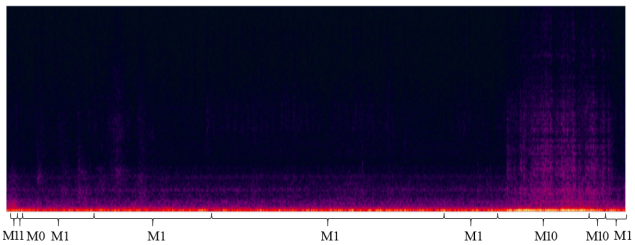

Figure 2: *The spectrogram and ASM sequence of an example recording from the tram scene. This example was misclassified by CNN-based approach [27] as the bus scene but correctly classified by our Hybrid-DNN approach.*

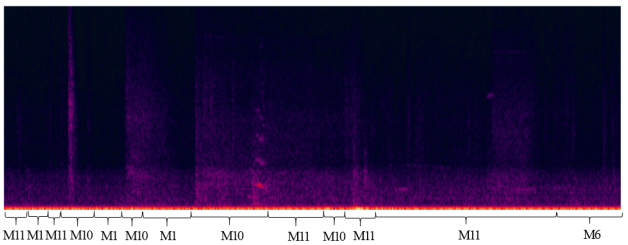

Figure 3: *The spectrogram and ASM sequence of an example recording from bus scene. This example was misclassified by CNN-based approach [27] as the tram scene but correctly classified by our Hybrid-DNN approach.*

segment. Each audio recording is transcribed by the dictionary of 20 ASM units named from M0 to M19. Based on the spectrograms and human listening of these two examples, the acoustically similar parts can be represented by the same ASM unit, like M10. The differences between two scenes can be captured by other ASM units such as M1 for *tram* scene and M11 for *bus* scene. In this way, our approach is able to explicitly show the correlation of segments in different scenes and give better classification results than CNN-based approach.

4. Conclusions

The approach we have proposed achieves a competitive performance with simple classification model. Inspired by speech recognition, we create a dictionary for acoustic scene utterances and transcribe each utterance into units in the dictionary to achieve the purpose of capturing similar segments in acoustic scenes. By utilizing a new initial segmentation, the dictionary of ASM units is built, which is then used to conduct ASM/HMM training. Finally, the word-document matrix can be classified by different classifiers back-end, such as SVM and DNN.

5. Acknowledgements

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 Grants No. 61671422 and U1613211, the Key Science and Technology Project of Anhui Province under Grant No. 17030901005, and Huawei Noah's Ark Lab.

6. References

[1] 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 Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 321–329, 2006.

[2] M. Bugalho, J. Portelo, I. Trancoso, T. Pellegrini, and A. Abad, "Detecting audio events for semantic video search," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[3] A. Mesaros, T. Heittola, and T. 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] X. Bao, T. Gao, J. Du, and L.-R. 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] 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.

[6] D. Battaglino, L. Lepauloux, N. Evans, F. Mougins, and F. Biot, "Acoustic scene classification using convolutional neural networks," *IEEE AASP Challenge on Detec*, 2016.

[7] T. H. Vu and J.-C. Wang, "Acoustic scene and event recognition using recurrent neural networks," *Detection and Classification of Acoustic Scenes and Events*, vol. 2016, 2016.

[8] H. Jallet, E. Cakır, and T. Virtanen, "Acoustic scene classification using convolutional recurrent neural networks," *the Detection and Classification of Acoustic Scenes and Events (DCASE)*, pp. 1–5, 2017.

[9] http://www.cs.tut.fi/sgn/arg/dcase2017/challenge/task-acoustic-scene-classification/.

[10] S. Mun, S. Park, D. K. Han, and H. Ko, "Generative adversarial network based acoustic scene training set augmentation and selection using svm hyper-plane," *Proc. DCASE*, pp. 93–97, 2017.

[11] Y. Sakashita and M. Aono, "Acoustic scene classification by ensemble of spectrograms based on adaptive temporal divisions," Tech. Rep., DCASE2018 Challenge, Tech. Rep., 2018.

[12] http://dcase.community/challenge2018/task-acoustic-scene-classification/.

[13] 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, 1988, pp. 501–541.

[14] 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, 2007.

[15] 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.

[16] 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.

[17] J. Reed and C.-H. Lee, "A study on music genre classification based on universal acoustic models." in *ISMIR*, 2006, pp. 89–94.

[18] 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.

[19] 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.

[20] M. E. Wall, A. Rechtsteiner, and L. M. Rocha, "Singular value decomposition and principal component analysis," in *A practical approach to microarray data analysis*. Springer, 2003, pp. 91–109.

[21] D. Elworthy, "Does baum-welch re-estimation help taggers?" in *Proceedings of the fourth conference on Applied natural language processing*. Association for Computational Linguistics, 1994, pp. 53–58.

[22] D. Hull, "Improving text retrieval for the routing problem using latent semantic indexing," in *SIGIR'94*. Springer, 1994, pp. 282–291.

[23] C.-W. Hsu, C.-C. Chang, C.-J. Lin *et al.*, "A practical guide to support vector classification," 2003.

[24] http://dcase.community/challenge2018/task-acoustic-scene-classification-results-a/.

[25] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE transactions on Neural Networks*, vol. 13, no. 2, pp. 415–425, 2002.

[26] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010*. Springer, 2010, pp. 177–186.

[27] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," *arXiv preprint arXiv:1807.09840*, 2018.