Prediction-error-based Adaptive SpecAugment for Fine-tuning the Masked Model on Audio Classification Tasks

Xiao Zhang* Haoran Xing* Mingxue Song* Daiki Takeuchi† Noboru Harada† Shoji Makino*
* Waseda University, Japan
E-mail: zhang_x07@toki.waseda.jp, haoranxing@suou.waseda.jp, smx_shue@toki.waseda.jp, s.makino@waseda.jp
† NTT Corporation
E-mail: daiki.takeuchi.ux@hco.ntt.co.jp, noboru.harada.pv@hco.ntt.co.jp

***Abstract*—Spectrogram augmentation (SpecAugment), a data processing method that enlarges datasets without adding new samples, has been widely employed in the fine-tuning process of the masked model for audio classification. In the conventional SpecAugment, the positions of time masking and frequency masking, which directly determine the available information that the model can learn from, are selected randomly. As a result, the random position masking in the conventional SpecAugment may prevent the model from fully utilizing the input information. To this end, we propose a Prediction-error-based Adaptive SpecAugment (PEAS), which incorporates two auxiliary tasks based on reconstruction and then introduces a mask position selector in the fine-tuning process for the masked model. Rather than masking at random positions, the proposed PEAS generates masks on the parts of the spectrograms that are hard to reconstruct for the model. Masking these positions forces the model to learn more generalized audio features, which can effectively prevent the model from learning classification labels by identifying individual special features. Besides, to accelerate the learning of audio features during the early training epochs, we progressively increase the proportion of adaptive masks. Experimental results demonstrate that our proposed PEAS can match or outperform both the conventional method and random masking strategy on the ESC-50 and Speech Commands V2 datasets.**

I. Introduction

The audio classification task aims to enable the machine to automatically recognize and distinguish between different kinds of sounds. Classifying audio segments into predefined categories [1] provides useful information for understanding audio content, which is crucial in several domains, including speech recognition [2], environmental sound classification [3], emotion recognition [4], etc.

Spectrogram augmentation (SpecAugment) [5] is a regularization technique used for audio data processing, aiming to improve the performance of large-scale deep neural network (DNN) models on small datasets [6] [7]. This paper focuses on two groups of SpecAugment strategies [8] [9]: time masking and frequency masking. These techniques simulate scenarios where some data is lost during actual data acquisition, thereby forcing the model to extract meaningful features from incomplete input data.

However, conventional time masking and frequency masking strategies still have limitations due to the predefined random strategy of masked position selection. These positions of masks directly determine the available audio information for feature extraction, which is highly correlated to the performance of audio classification. The random masking strategy prevents the model from fully utilizing the limited data to obtain generalized audio features. Thus, we propose a Prediction-error-based Adaptive SpecAugment (PEAS) that encourages the model to mask where it is most beneficial.

The related work Hard Patches Mining (HPM) [10] in Computer Vision (CV) employs a new training paradigm for Masked Visual Modeling (MVM), which enables the model to identify patches that are difficult to reconstruct, namely, hard patches. It introduces two processes: patch reconstruction and hard patch prediction. They use two models with the same structure: a student model mainly to reconstruct patches and a teacher model mainly to identify the hard patches, respectively. The teacher model can identify the parts of the images that are hard to reconstruct and mask these parts during the pretraining process, thereby bootstrapping the performance of masked image modeling across various downstream tasks. HPM method has proven that hard patches often contain more specific features, and masking these patches can compel the model to discern finer differences between various classes using more generalized image features, thereby enhancing the stability and performance of the model. Since HPM has shown superior performance and considering the similarity between visual images and spectrograms, the HPM idea can be extended to the fine-tuning process of the masked model for audio classification tasks.

Although the HPM method is expected to improve the performance of masked models during fine-tuning, its high training cost and the complex update mechanism for two models limit its application on low-resource terminals. To solve this issue, we employ aforementioned two processes as auxiliary tasks on one masked model during the fine-tuning process inspired by the concept of Multi-Task Learning (MTL) [11], which can decrease the number of parameters and save training time. In addition, block masking has proven effective in SpecAugment, thus we mask consecutive time frames and frequency bins as well. These two auxiliary tasks share the

same parameters, serving for spectrogram reconstruction and hard block prediction, respectively.

To determine which blocks to mask in the spectrograms, we propose a mask position selector. In PEAS strategy, each of the masks has a block shape along the time and frequency axes, respectively. The blocks with a larger mean prediction error are desired to be masked. The proposed PEAS enables the model to identify the hard blocks in the input spectrograms, which are believed to have specific information for audio classification. Thus, masking these parts is expected to force the model to learn more generalized features from the input spectrogram, thus improving the efficacy of the masked model for audio classification tasks.

We apply Self-Supervised Audio Spectrogram Transformer (SSAST) [12] as the backbone of proposed PEAS. SSAST is a masked model that significantly improves performance on various downstream tasks in the filed of audio classification without the need for large amounts of labeled data. Moreover, SSAST also includes generative masked spectrogram patch modeling as one of the joint training tasks, which is the same as one of our proposed auxiliary tasks.

The rest of this paper is organized as follows: Section II provides an overview of related work; Section III details the proposed methodology; Section IV presents the experimental setup and discusses the results; and concludes the paper in Section V.

II. Related Work

A. Time masking and Frequency Masking for SpecAugment

SpecAugment, which is commonly used in various fields [5] [13] [14], can enhance the generalization ability of models at a low cost. Specifically, time masking and frequency masking proposed in [5] can be regarded as an augmentation policy that acts on spectrograms directly. The positions of masks on the spectrogram are important, as they determine the available information for the model to learn. However, SpecAugment selects the starting point from a uniform distribution of the given parameters and generates masks with a random width. This random masking in both consecutive time frames and frequency bins is inefficient for obtaining the input information.

B. Hard Patches Mining for Masked Image Modeling

Hard Patches Mining (HPM) [10] is an approach for masked visual modeling initially deployed in CV pretraining. It has been demonstrated that patches with higher predicted losses are often more discriminative, and thus masking these patches brings a hard situation. HPM introduces two collaboratively trained models. The student model reconstructs the masked patches, while the teacher model predicts the positions of hard patches and generates masks. This iterative process drives the model to continuously generate and tackle difficult tasks, thereby enhancing its overall comprehension of the visual content. The findings in HPM have demonstrated superior performance based on the ViT model, which is a Transformer-based model used for image classification tasks. Therefore, HPM is supposed to be suitable for the masked model with Transformer structure in the audio classification task as well, which takes the spectrograms as input. It is proved that HPM can integrate with existing frameworks and improve performance consistently.

C. Multi-Task Learning

The Multi-Task Learning (MTL) [15] method simultaneously leverages the common information from multiple related tasks to enhance the generalization ability and learning efficiency of the model. By using a shared module for feature extraction, models can efficiently learn the common features, which are then fine-tuned through task-specific layers. The model proposed in [16] utilizes cross-task knowledge for feature extraction from spectrograms and adopts a convolutional neural network with a multi-task learning scheme for different subtasks, which effectively reduces the computational cost of multi-task.

D. Self-Supervised Audio Spectrogram Transformer

Self-Supervised Audio Spectrogram Transformer (SSAST) [12] is a self-supervised pretrained model designed for audio classification tasks. It not only matches but often exceeds the performance of previous supervised models. Also, it exhibits enhanced generalization capabilities. SSAST introduces two auxiliary tasks for joint pretraining, i.e., reconstruction and discrimination of masked patches. It uses both audio and speech datasets for pretraining, which allows it to achieve good performance on both audio classification and speech recognition tasks.

III. Proposed Method

To improve the classification ability of the masked model, we propose a Prediction-error-based Adaptive SpecAugment (PEAS) using a generative masking strategy to calculate the mask positions in the fine-tuning process. In our work, we utilize two auxiliary tasks to calculate the mean prediction errors of each time and frequency blocks, which indicate the positions that are hard to reconstruct, and then generate masks on hard positions via a mask position selector. The overall architecture of the proposed model is shown in Fig. 1.

A. Auxiliary Tasks for MTL Model

The two auxiliary tasks are designed as spectrogram reconstruction and hard block prediction, respectively. Unlike the HPM method, which employs auxiliary tasks on two different models with a same structure, the proposed adaptive masking strategy conducts both tasks on a single model to aid audio classification based on the MTL method.

The first auxiliary task is to reconstruct the original spectrogram. Its process is illustrated as step 1 in Fig. 1. The target of this auxiliary task involves reconstructing the input spectrogram and calculating the reconstruction error to evaluate the reconstruction quality. The value of the reconstruction error can be mapped to the concept of reconstruction difficulty. For each point in spectrograms, a larger reconstruction error

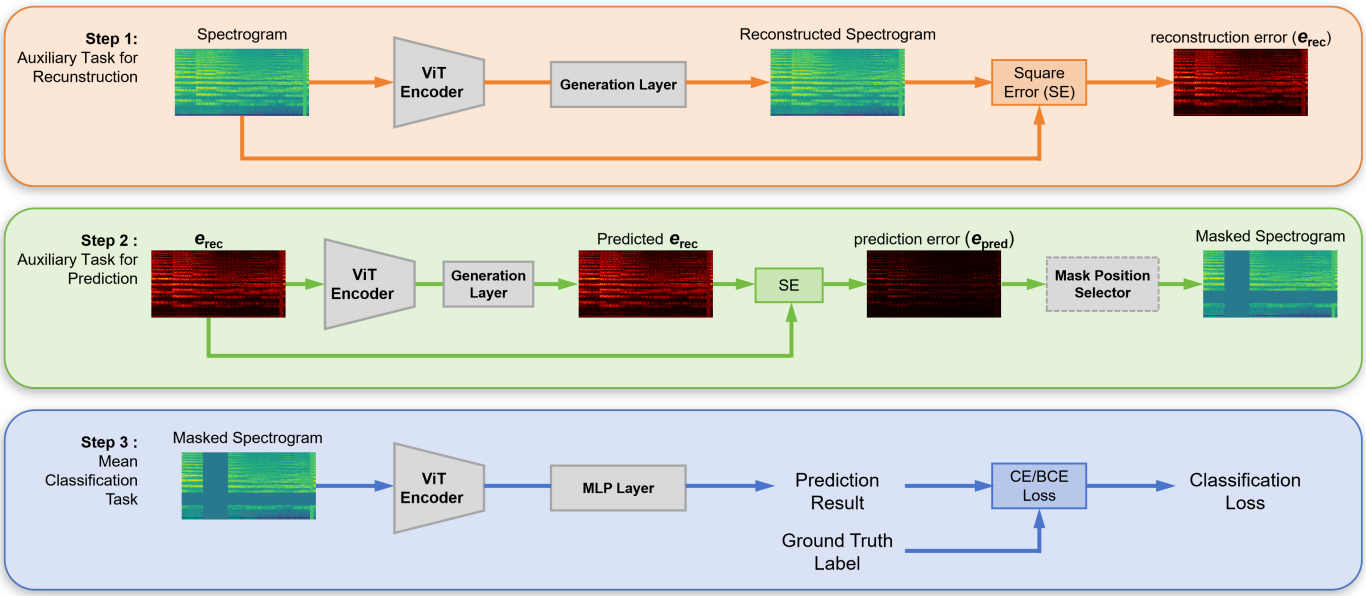

Fig. 1. Overview of the PEAS model process over one training epoch. In Step 1, the reconstruction error e_{rec} is calculated using the original spectrogram. In Step 2, the prediction error e_{pred} is derived from e_{rec}. Finally, Step 3 involves the audio classification task, utilizing a masked spectrogram as input. Each step's input is the output from the preceding step.

indicates a greater prediction deviation, implying increased difficulty in accurate reconstruction.

Assuming that the input spectrograms are 3D tensors with the shape $[B, H, W]$, where B denotes the batch size, H represents the input frame length, and W represents the number of input mel bins. In each iteration, we reconstruct the spectrograms and calculate the point-wise error using the Squared Error (SE) from the input spectrograms.

$$e_{\text{rec}} = (G(E(x)) - x)^2, \tag{1}$$

where x is the original spectrograms, E is the encoder and G is the generation layer. e_{rec} denotes the reconstruction error. As a result, the reconstruction error e_{rec} has the same shape as the input spectrogram x.

The second auxiliary task is to predict the reconstruction difficulty of blocks, as depicted in picture (ii) in Fig. 2 (b). The objective of this task is to enable the model to accurately indicate which positions in the spectrogram have the largest reconstruction error. Subsequently, the model is supposed to generate masks at these positions. We use the point-wise SE function between the true reconstruction error and predicted value of reconstruction error as the output named prediction error for the hard block prediction task:

$$e_{\text{pred}} = (G(E(e_{\text{rec}})) - e_{\text{rec}})^2, \tag{2}$$

where e_{pred} indicates prediction error.

The final classification task and two auxiliary tasks introduced above are employed on a pretrained SSAST-base model provided in [12]. The update of the model mainly relies on the loss of the main classification task. We use the masked spectrograms as input to predict the audio labels and update the whole model using the cross-entropy (CE) loss or binary cross-entropy (BCE) loss, while the auxiliary tasks indirectly bootstrap the training of the classification task by generating masks. This approach prevents the model from using special information to learn classification labels, thus improving the model's learning ability and generalization performance. Even if the model parameters are not directly updated with the loss from the auxiliary tasks, these tasks can still act as a regularizer through the masking strategy, preventing the model from overfitting.

B. Mask Position Selector

With the auxiliary tasks, the mask position selector is designed to identify hard blocks in the spectrograms and generate masks at those positions. As depicted in picture (iii) in Fig. 2 (b), we compute the average error for each block along the time frames and frequency bins separately, using the expected maximum mask width as a unit of calculation with a stride of 1, and then we rank the results. This allows the model to identify the predicted time and frequency blocks with higher average errors, i.e., the blocks that are hard to reconstruct. Generating masks with similar shape to those of the SpecAument method in these blocks can maximize the coverage of difficult information.

IV. Experiment

We validated the proposal for audio classification on the ESC-50 dataset and the Speech Commands V2 (SC-V2) dataset. The SSAST model was employed as the backbone of the audio classification model, along with the aforementioned

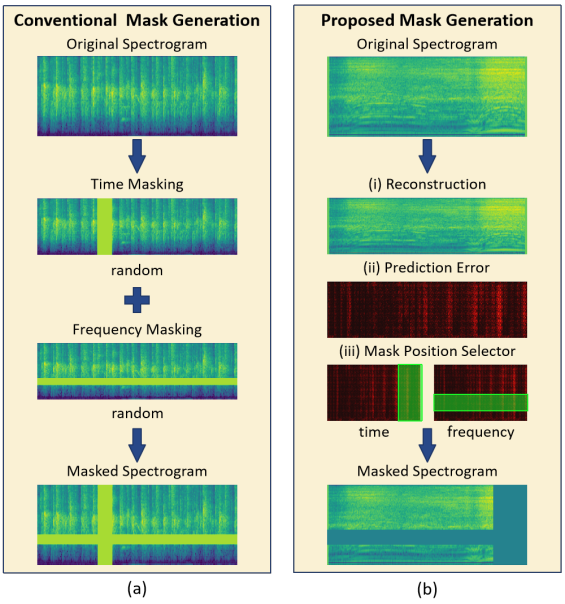

Fig. 2. The comparison between the conventional random masking method and the proposed PEAS masking strategy. In (a), the traditional SpecAugment masking method is shown. In (b): (i) Reconstruction of the original spectrogram, (ii) Prediction of difficulty, where brighter colors represent more challenging parts, and (iii) Illustration of hard block selection using a sliding window.

two auxiliary tasks. For better classification performance, we introduced the progressive strategy into the training process.

To verify the effectiveness of the proposed PEAS in the field of SpecAugment, we compared the performance of the SSAST model with the conventional SpecAugment method and the proposed adaptive SpecAugment method. To ensure the consistency of approaches, we adopted the same experimental settings as those used in the original SSAST paper for parameters as long as possible.

To further validate the superiority of the proposed method, we compared the performance of the proposed masking position selector and the random masking strategy. In addition, we explored the different numbers of masking blocks' impacts. We also generated multiple masks for evaluation to further illustrate the superiority of adaptive masking generation over the random masking strategy.

A. Datasets

We validated the proposed PEAS on the ESC-50 [17] and SC-V2 [18] datasets. The ESC-50 dataset contains 2,000 environmental sound clips (5 seconds each) across 50 classes, with 40 samples per class, and we used 5-fold cross-validation. The SC-V2 dataset includes over 105,000 one-second speech commands from diverse speakers, covering 35 classes, using separate validation and evaluation sets.

TABLE I
EXPERIMENTAL SETUP

Settings	ESC-50	SC-V2
Num of Classes	50	35
Spectrogram Size (T/F)	512 / 128	128 / 128
Initial Learning Rate	1e-4	2.5e-4
Batch Size	10	16
Epochs	25	30
Ground Truth Loss	Cross Entropy	Binary Cross Entropy
Pretraining	Librispeech/AudioSet	Librispeech/AudioSet
Time Masking	96	48
Frequency Masking	24	48
Mix up	0	0.6

B. Training Settings

Table I shows detailed training parameters employed in our experiments. We used batch sizes of 10 and 16 for the ESC-50 and SC-V2 datasets, respectively, and the remaining settings were consistent with those in SSAST. The backbone SSAST model was pretrained on the AudioSet [19] and Librispeech [20] datasets using 400 patch-level masks. All experiments were repeated at least three times, and the average accuracy was reported.

C. Adaptive Masks Settings

To investigate the superiority of proposed PEAS over random masks and explore the impact of different adaptive mask configurations, we set up control groups with different numbers of mask blocks, i.e., T mask blocks are generated in the time domain and F mask blocks are generated in the frequency domain, respectively, denoted as (T, F). The combination of (T, F) was set to (1, 1), (2, 2) and (3, 3) in our experiments, and the performance with different masking strategies were compared in Table III.

During the first 80% of the training epochs, we generated masks with uniformly distributed widths, and the maximum width is defined by predefined parameters. In the final 20% of the training epochs, we generated masks with a fixed-width which is the expected value of the uniformly distributed widths.

In addition, when the number of generated mask blocks was set to more than one, the total number of mask blocks was the sum of the number of random and adaptive mask blocks. We progressively increased the proportion of adaptive mask blocks according to the training period. The number of adaptive mask blocks N_{a} generated is given by

$$n_{\mathrm{a}} = \min\left(n_t, \left(0.1e + 0.9 \times n_t \times H\left(e - 0.8e_t\right)\right)\right), \qquad (3)$$

where n_{a} is the number of adaptive mask blocks, n_t is the total number of random and adaptive mask blocks generated in the time and frequency domains, respectively, e is the current training epoch, e_t is the total number of training epochs, and H is the Heaviside step function.

TABLE II
CLASSIFICATION ACCURACY OF CONVENTIONAL SPECAUGMENT AND THE PROPOSED PEAS

SpecAugment	Num of masks	ESC-50	SC-V2
Conventional	(1, 1)	88.80%	97.62%
PEAS (Proposed)	**(1, 1)**	**91.15%**	**97.61%**

Num of masks means T mask blocks are generated in the time domain and F mask blocks are generated in the frequency domain, respectively, denoted as (T, F).

TABLE III
PEAS ABLATION STUDY: DIFFERENT MASK BLOCK NUMBERS AND MASKING STRATEGIES

SpecAugment	Num of masks	ESC-50	SC-V2
Conventional	(1, 1)	88.80%	97.62%
Random Masking	(3, 3)	89.35%	96.13%
	(2, 2)	89.65%	97.22%
	(1, 1)	90.00%	97.38%
PEAS (Adaptive Masking)	(3, 3)	90.35%	97.20%
	(2, 2)	90.70%	97.23%
	(1, 1)	**91.15%**	**97.61%**

D. Results

Table II shows the comparison between the conventional SpecAugment method and the proposed PEAS under the same setting. The PEAS with the mask position selector showed better performance than the conventional SpecAugment with a random masking strategy on the small datasets. In particular, this method had a 2.35% improvement in accuracy on the ESC-50 dataset and had comparable performance on the SC-V2 dataset.

Table III shows results of the proposed PEAS with different numbers of mask blocks and different masking strategies. According to the results in Table III, the proposed method shows higher accuracy than the random masking strategy on the ESC-50 dataset and the SC-V2 dataset with identical conditions, regardless of the (T, F) configuration. The proposed strategy improved audio classification accuracy on the ESC-50 and SC-V2 datasets by 1.15% and 0.23%, respectively. In addition, several configurations of mask blocks on the time and frequency axes were compared. When (T, F) were set to (1, 1), the model performed best on the ESC-50 and SC-V2 datasets, with accuracy of 91.15% and 97.61%, respectively.

E. Discussion

The above results showed that the proposed PEAS improved the performance of the SSAST model compared to the conventional SpecAugment method, proving that masking hard blocks can help to improve the audio classification performance of the masked model. The accuracy does not improve on the SC-V2 dataset, which may be due to the fact that the model relies on more stable temporal information and semantic features in the learning of speech data. In addition, since speech is a structured signal, the model uses contextual and linguistic rules to infer ambiguous information, so the effect of the positions of the masks will not be as significant as in scene sounds.

As the number of mask blocks increases, the classification accuracy decreases. This may be because masking more blocks containing important information prevents the model from acquiring enough necessary learning information, making the learning process less effective. Additionally, introducing multiple masks can lead to excessive fragmentation of the original data, disrupting the continuity of audio signal features in the time-frequency domain. This makes it difficult for the model to extract effective features during training. However, even with multiple masks, the adaptive masking strategy still outperforms the random masking strategy. This indicates that selectively masking hard blocks allows for more effective guidance of the model in feature learning and generalization.

V. CONCLUSIONS

The conventional SpceAugment method for audio classification tasks generates masks for spectrograms in the time and frequency dimensions randomly. This randomness causes the model to acquire less effective information, resulting in an incomplete understanding of the generalized features of the audio. We propose a Prediction-error-based Adaptive SpecAugment (PEAS) that integrates a mask position selector with the progressive generation strategy for the masked model in the field of audio classification. This approach incorporates auxiliary tasks to identify hard blocks where masks are generated, thereby enforcing the model to learn more generalized features and enhancing the overall performance in the fine-tuning phase. We validated our method on the ESC-50 and SC-V2 datasets, which achieved superior or comparable results to the conventional method, demonstrating its robustness and adaptability in handling different audio data types. Furthermore, ablation studies confirm that PEAS bootstraps masked models for better feature learning and accuracy compared with random mask generation. These findings underscore the potential of adaptive masking to enhance the performance of masked models.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Number 23H03423.

REFERENCES

[1] L. Lu, H.-J. Zhang, and H. Jiang, "Content analysis for audio classification and segmentation," *IEEE Transactions on speech and audio processing*, vol. 10, no. 7, pp. 504–516, Oct. 2002.

[2] M. Alzantot, B. Balaji, and M. Srivastava, "Did you hear that? Adversarial examples against automatic speech recognition," *arXiv preprint arXiv:1801.00554*, Jan. 2018.

[3] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *Proc. MLSP*, Nov. 2015, pp. 1–6.

[4] D. Issa, M. F. Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," *Biomedical Signal Processing and Control*, vol. 59, no. 101894, May 2020.

[5] D. S. Park *et al.*, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, Dec. 2019.

[6] M. Caron *et al.*, "Emerging properties in self-supervised vision transformers," in *Proc. ICCV*, May 2021, pp. 9650–9660.

[7] S. Hershey *et al.*, "CNN architectures for large-scale audio classification," in *Proc. ICASSP*, Jan. 2017, pp. 131–135.

[8] J. Han, M. Matuszewski, O. Sikorski, H. Sung, and H. Cho, "Randmasking augment: A simple and randomized data augmentation for acoustic scene classification," in *Proc. ICASSP*, June 2023, pp. 1–5.

[9] G. Kim, D. K. Han, and H. Ko, "Specmix: A mixed sample data augmentation method for training with time-frequency domain features," *arXiv preprint arXiv:2108.03020*, Aug. 2021.

[10] H. Wang, K. Song, J. Fan, Y. Wang, J. Xie, and Z. Zhang, "Hard patches mining for masked image modeling," in *Proc. CVPR*, Apr. 2023, pp. 10 375–10 385.

[11] L. Liebel and M. Körner, "Auxiliary tasks in multi-task learning," *arXiv preprint arXiv:1805.06334*, May 2018.

[12] Y. Gong, C.-I. Lai, Y.-A. Chung, and J. Glass, "SSAST: Self-supervised audio spectrogram transformer," in *Proc. AAAI*, vol. 36, Feb. 2022, pp. 10 699–10 709.

[13] D. S. Park *et al.*, "Specaugment on large scale datasets," in *Proc. ICASSP*, May 2020, pp. 6879–6883.

[14] P. Bahar, A. Zeyer, R. Schlüter, and H. Ney, "On using specaugment for end-to-end speech translation," *arXiv preprint arXiv:1911.08876*, Nov. 2019.

[15] M. Crawshaw, "Multi-task learning with deep neural networks: A survey," *arXiv preprint arXiv:2009.09796*, Sep. 2020.

[16] T. L. Nwe, T. H. Dat, and B. Ma, "Convolutional neural network with multi-task learning scheme for acoustic scene classification," in *Proc. APSIPA*, Dec. 2017, pp. 1347–1350.

[17] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proc. ACMMM*, Oct. 2015, pp. 1015–1018.

[18] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," *arXiv preprint arXiv:1804.03209*, Apr. 2018.

[19] J. F. Gemmeke *et al.*, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. ICASSP*, Mar. 2017, pp. 776–780.

[20] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. ICASSP*, Apr. 2015, pp. 5206–5210.