Frequency Axis Pooling Method for Weakly Labeled Sound Event Detection and Classification

Miao Liu* and Jing Wang* Yujun Wang† Lidong Yang‡
* Beijing Institute of Technology, Beijing, China
E-mail: liumiao424@163.com, wangjing@bit.edu.cn Tel/Fax: +86-10-18810318269
† Xiaomi Inc., Beijing, China
E-mail: wangyujun@xiaomi.com
‡ Inner Mongolia University of Science and Technology, Baotou, China
E-mail: yld_nkd@imust.edu.cn

***Abstract*—Recently, the convolutional recurrent neural network (CRNN) has been widely used in weakly labeled sound event detection (SED) and audio tagging (AT) tasks. However, it is possible that the information of frequency dimension is not well used in the existing network design, which may cause information loss or redundancy. We propose a frequency axis pooling method to further boost the representation power of CRNN. Based on the existing pooling functions, the frequency axis pooling is applied on the feature map before recurrent neural network (RNN) input in CRNN. Compared to frequency axis no-pooling method, our method assigns different weights to different frequency dimensions during compressing, which can better compress frequency information and reduce information redundancy. To evaluate the proposed method, three commonly used pooling functions on frequency axis are compared on the Dcase2017 task4 dataset. The experimental results show that reasonable compression of frequency information helps to improve the performance of AT and SED tasks significantly. Among them, the frequency axis pooling based on linear softmax performs the best on both tasks.**

I. Introduction

Sound event detection is a task that detects not only event categories but also the start and end times of sound events in audio stream, which has a wide range of application scenarios, including environmental and security monitoring [1], [2], [3], [4] in real life.

The traditional SED method is based on strong labeled data with the onset and offset time of each sound event which is very tedious to be obtained manually. Therefore, many weakly-labeled datasets have been published and applied to SED research. The typical solution for weakly-labeled SED is multiple instance learning (MIL) [5]. In MIL, we do not know the label of each instance, only the label of the bag containing many instances is provided. A bag containing one or more positive instances is considered as a positive bag, otherwise negative. If the instance is treated as a frame and the bag is treated as an audio clip, it is very consistent with the weakly labeled SED task. When the neural network gives the sound event predictions on each frame, a pooling function on the time axis is needed to compress the frame-level predictions and aggregate them into clip-level predictions. In weakly labeled SED, the earliest used pooling functions are max pooling [6] and average pooling [7]. Later, Kong et al. [8] proposed an attention pooling function, which has been adopted in several works [9], [10]. Hong et al. [11] proposed gated multi-head attention pooling (GMAP) for MIL, which can attend to the information of events from different heads. In [12], pooling is applied over both time and frequency axis during feature extraction. In [13] and [14], authors used a two-dimensional pooling to weigh and pool the neural network output.

Similarly, many researchers are paying attention to the network construction of weakly labeled SED. CNN was introduced for large-scale audio classification [15]. Lu et al. [16] proposed a multi-scale RNN model that has the benefits of modeling both the fine-grained and long-term dependencies. A convolutional recurrent neural network (CRNN) [17] with learnable gated linear units (GLUs) non-linearity applied on the log mel spectrogram is proposed by Xu et al.. Since it exhibits strong performance on audio tagging and localization at the same time, the CRNN structure has been favored by most scholars in weakly labeled SED task once it was proposed. Yan [18] made improvements on this basis and proposed a novel region multi-scale based attention method. Recently, Hong et al. [19] proposed a CNN-based spatial and channel wise attention (SCA) to explore the effect of attention for weakly labeled audio tagging. Their work proves that optimizing the network structure and using audio information rationally within the network are very beneficial to improve the accuracy of weakly labeled SED task.

In the paper, we propose a frequency axis pooling method to improve the performance on weakly labeled sound event detection and classification tasks. Based on the commonly used pooling functions, experiments have been conducted to apply three pooling functions on the frequency dimension of the feature map after CNN output and before RNN input in CRNN to find the best choice for SED tasks. This does not only help the CRNN structure to better compress and utilize the information of shallow audio features without generating information redundancy, but also reduces the size of the feature map and the amount of calculation. Furthermore, a combined experiment of the pooling function on frequency and time axis is designed to compare the performance of frequency axis pooling on different time axis pooling in order to get the best result.

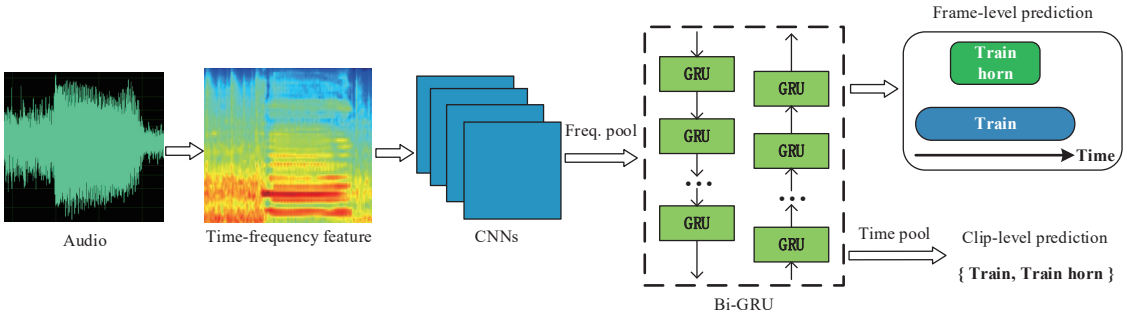


Fig. 1. Overview of sound event detection system adding the proposed method.

TABLE I
DEFINITION OF THREE POOLING FUNCTIONS.

Pooling function	Definition
Average pooling	$y^k = \frac{1}{n}\Sigma_i y_i^k$
Linear softmax	$y^k = \frac{\Sigma_i (y_i^k)^2}{\Sigma_i y_i^k}$
Attention	$y^k = \frac{\Sigma_i y_i^k w_i^k}{\Sigma_i w_i^k}$

II. POOLING FUNCTION

In this section, we briefly introduce the existing pooling function. In traditional weakly labeled SED, the role of pooling function is to aggregate frame-level predictions into clip-level predictions on the time axis. Wang [20] made a detailed comparative analysis of five pooling functions. Among them, linear softmax and attention pooling function achieve a strong performance for both audio tagging and localization.

Let $y_i^k \in [0,1]$ be the frame-level probability of a certain sound event type k at the i-th frame, where $k \in \{1,...,K\}$ and K is the number of events. y_i^k is output after CRNN and w_i^k is the weights for each frame if necessary. Let $y^k \in [0,1]$ be the aggregated clip-level probability of the same event k. y^k is output after the pooling function. We list the definitions of the three commonly used pooling functions to be compared in Table I.

Finally, the model is trained to minimize the binary cross entropy loss over all events, which is defined as:

$$\min -\frac{1}{K}\sum_k (t_k \log y_k + (1-t_k)(1-\log y_k)) \quad (1)$$

where $t_k \in [0,1]$ is the clip-level ground truth of the sound event k.

III. PROPOSED METHOD

An overview of our proposed method based on CRNN system is illustrated in Fig.1. We apply pooling function not only on the time axis, but also on the frequency axis.

The time-frequency characteristics of a piece of audio are represented by X. A convolutional layer can be represented as:

$$Y = W * X + b \quad (2)$$

where $*$ is the convolution operator, W and b represent the filter kernel and bias respectively. The input $T \times F$ feature $X \in R^{T\times F}$ is transformed into feature maps $Y \in R^{C\times T\times F}$ after convolution by the filter kernel adding a new dimension C, where T represents the number of time dimensions, F represents the number of frequency dimensions and C represents the number of channels .

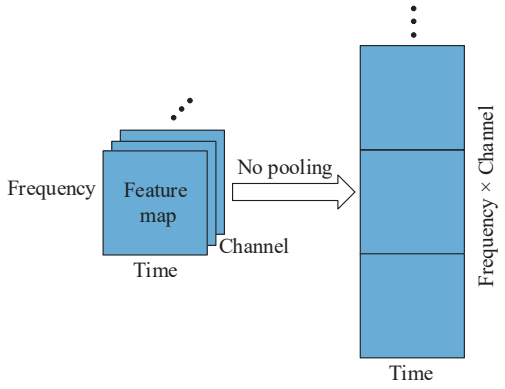
However, the input data $Y' \in R^{C'\times T}$ of RNN is usually two-dimensional. Therefore, the dimension reduction is often performed after the output of the CNN part, which has two commonly used methods, dimensional merging or dimensional compression. As shown in Fig.2(a), Wang [20] merged the C and F dimensions in Y into one new dimension C' without pooling method, where C' equals $F \times C$. For dimensional compression, Kong [10] took the average pooling along the F dimension and removed the F dimension in Y to achieve dimension reduction. This method can be represented as:

$$Y' = g(Y) = \frac{1}{F}\sum_{i=1}^{F} Y_i \quad (3)$$

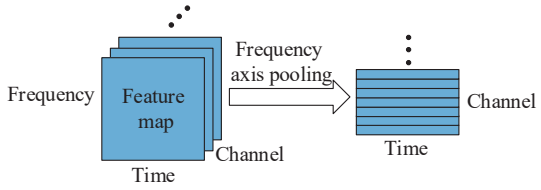
But neither of them conducted in-depth research on this part. Since the pooling functions we introduce in section 2 is similar to the dimension reduction on Y, we introduce those pooling functions to frequency axis. That is, to replace $g()$ in formula (3) with pooling functions such as linear softmax or attention in Table I. They can be represented as:

$$Y' = \frac{\sum_{i=1}^{F} Y_i^2}{\sum_{i=1}^{F} Y_i} \quad (4)$$

$$Y' = \frac{\sum_{i=1}^{F} Y_i W_i}{\sum_{i=1}^{F} W_i} \quad (5)$$



(a) Frequency axis no-pooling method.



(b) Frequency axis pooling method

Fig. 2. Illustration of dimension reduction on feature map with frequency axis no-pooling and pooling method.

The linear softmax pooling function on the frequency axis computes Y' as a weighted average of Y_i's, where the weights are equal to Y_i's themselves. The attention pooling function on the frequency axis is also a weighted average, where the weights W_i for each frequency dimension are learnable and modeled by a dedicated layer in neural network.

As shown in Fig.2(b), we name the reduction of frequency dimension in CRNN as frequency axis pooling. Frequency axis pooling compresses the frequency dimension of the feature map output by CNN and then sends the compressed feature map that only retains the time and channel dimensions into RNN. Meanwhile, the reduction of time dimension is named as time axis pooling in our paper. Fig.1 clearly shows their own position in the system.

In our experiment, the frequency dimension represents the shallow features (Fbank) of audio clips. The importance of information on different frequencies is different. Simple average compression makes the gradient distributed evenly across all frequences in back propagation, which may cause information loss. But linear softmax and attention pooling function will avoid this problem [20]. Meanwhile, since Fbank is artificially extracted feature, there is not much information available after CNN learning. Therefore, if the channel and frequency dimensions are spliced without compression, it may cause information redundancy and affect the deep feature extracted by CNN.

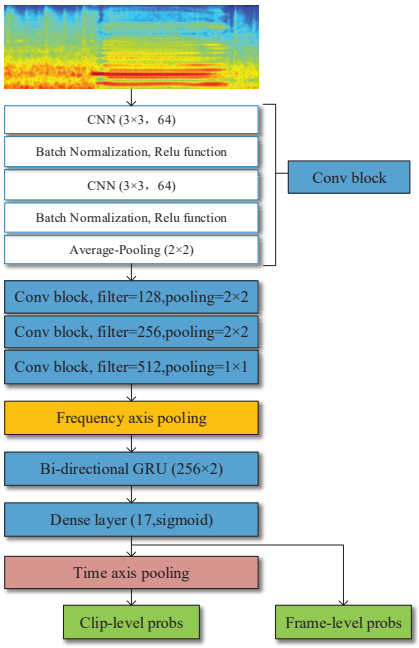


Fig. 3. Illustration of CRNN structure with the frequency axis pooling layer added between CNN and RNN. The size of all convolutional kernels is 3×3.

IV. Experiments

A. Dataset

We evaluate the proposed method on the dataset of task 4 in the DCASE2017 challenge [21], which is a subset of AudioSet [22]. The subset contains 17 classes of events divided into two categories: "Warning" and "Vehicle". The data consists of a training subset with 51172 audio clips, a development subset with 488 audio clips and an evaluation set with 1103 audio clips. The training subset is weakly labeled and no timestamps are provided. The development and evaluation subsets are both weakly and strongly labeled for evaluation. Most of these audio clips have duration of 10 seconds.

B. Experimental setup

As shown in Fig.3, we use CRNN to build our model and add frequency axis pooling layer between CNN and RNN.

We use log mel spectrogram (Fbank) as input feature. To begin with, all audio clips are resampled to 32 kHz. The frame length is 40 ms with the frame shift of 10 ms. Each chunk has 1000 frames and 64 mel bins. The CNN part includes 4 convolutional blocks. Each block consists of 2 convolutional layers and an average-pooling. Batch normalization and ReLU function is applied after each convolutional layer. In RNN part, the bi-directional gated recurrent units (GRU) [23] is used to obtain past and future time information. During model training, we use the Adam [24] optimizer with the initial learning rate of 0.001 and reduce it to 0.0001 after 50000 iterations. The mini batch size is 32. The loss function is binary cross entropy based on clip-level labels. SpecAugment [25] and Mixup [26] is used in all experiments to prevent

TABLE II
PERFORMANCE OF DIFFERENT SYSTEMS FOR AUDIO TAGGING (AT) TASK ON THE DEVELOPMENT SET AND EVALUATION SET.

Development set (Frequency axis + time axis)	F1	P	R
No-pooling + attention [20]	- - -	- - -	- - -
No-pooling + linear softmax [20]	- - -	- - -	- - -
Average + attention [10]	0.581	0.575	0.587
Average + linear softmax	0.623	0.650	0.598
Attention + attention	0.598	0.635	0.564
Attention + linear softmax	0.606	0.635	0.580
Linear softmax + attention	0.616	0.631	**0.603**
Linear softmax + linear softmax	**0.627**	**0.670**	0.589
Evaluation set (Frequency axis + time axis)	F1	P	R
No-pooling + attention [20]	0.492	0.487	0.497
No-pooling + linear softmax [20]	0.495	0.469	0.523
Average + attention [10]	0.640	0.637	**0.642**
Average + linear softmax	0.647	0.687	0.612
Attention + attention	0.641	**0.697**	0.593
Attention + linear softmax	0.638	0.676	0.599
Linear softmax + attention	**0.648**	0.668	0.630
Linear softmax + linear softmax	0.647	0.696	0.606

TABLE III
PERFORMANCE OF DIFFERENT SYSTEMS FOR SOUND EVENT DETECTION (SED) TASK ON THE DEVELOPMENT SET AND EVALUATION SET.

Development set (Frequency axis + time axis)	F1	ER
No-pooling + attention [20]	- - -	- - -
No-pooling + linear softmax [20]	- - -	- - -
Average + attention [10]	0.537	**0.65**
Average + linear softmax	0.545	0.675
Attention + attention	0.542	0.662
Attention + linear softmax	0.519	0.695
Linear softmax + attention	0.546	0.673
Linear softmax + linear softmax	**0.550**	**0.652**
Evaluation set (Frequency axis + time axis)	F1	ER
No-pooling + attention [20]	0.401	1.025
No-pooling + linear softmax [20]	0.437	0.843
Average + attention [10]	0.584	0.68
Average + linear softmax	0.578	0.663
Attention + attention	0.587	0.667
Attention + linear softmax	0.561	0.682
Linear softmax + attention	**0.593**	0.665
Linear softmax + linear softmax	0.583	**0.661**

from overfitting. Since thresholds need to be applied to the predictions to obtain the presence or absence of sound events, automatic thresholds optimization algorithm [10] based on gradient descent (GD) of Adam is adopted, instead of applying the same threshold on all events. All experiments are repeated 3 times with random network initialization and the average result of each model is reported as the final result.

Three effective pooling functions of average, attention and linear softmax are experimented on the frequency axis. In order to further verify the effectiveness of frequency axis pooling, attention and linear softmax pooling functions are experimented on the time axis for combined experiments.

C. Results and Analysis

The performance of audio tagging in DCASE2017 task4 was evaluated with the F1 score on the clip level; sound event detection was evaluated with the F1 score and error rate (ER) on 1-second segments.

1) Audio Tagging (AT): Table 2 presents the F1, Precision (P) and Recall (R) results for audio tagging on the development set and evaluation set. The first column represents the pooling method of frequency axis and time axis in models, separated by symbol +. The symbol - - - indicates that the results are not presented in paper.

Analyzing the F1 of evaluation set, the results of frequency axis pooling method are significantly better than that of no-pooling method. As we mentioned in section III, it may be because that frequency dimension representing shallow features doesn't do pooling in deep network but retains all its information, which will cause information redundancy. However, the results of three pooling functions of average, attention and linear softmax on the frequency axis aren't much different. Linear softmax is slightly better than other two methods. Among them, the highest F1 on evaluation set of 0.648 can be obtained when frequency axis adopts linear softmax and time axis adopts attention pooling function.

2) Sound Event Detection (SED): The results of F1 and ER for sound event detection on the development set and evaluation set are given in Table 3.

In terms of the evaluation set, the experimental phenomenon of SED task is consistent with that of AT task. The results of pooling method on the frequency axis are also much better than that of no-pooling method, which indicates that frequency pooling is beneficial for improving the accuracy of both AT and SED tasks. Comparing three pooling functions on the frequency axis, linear softmax is better both in terms of F1 and error rate. On one hand, the highest F1 on evaluation set of 0.593 can be obtained when frequency axis adopts linear softmax and time axis adopts attention pooling function. On the other hand, the lowest ER on evaluation set of 0.661 can be obtained when both the frequency and time axis adopt linear softmax function. It is noticed that the best results on F1 and ER do not appear on the same structure, although they both use linear softmax as frequency axis pooling function. We will address this in the future work.

V. CONCLUSIONS

In this paper, we propose a frequency axis pooling method for weakly labeled sound event detection and classification. The experimental evaluation on the DCASE2017 Task4 dataset shows that the proposed method of frequency axis pooling outperforms the frequency axis no-pooling method both in terms of AT and SED tasks. Among them, linear softmax performs the best. Given that not merely weak labeled SED tasks need to pay attention to frequency axis pooling, the proposed method can be applied on strong labeled SED or other audio tasks in our future work.

ACKNOWLEDGMENT

This work is supported by National Nature Science Foundation of China (No.62071039 and No.61620106002).

REFERENCES

[1] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Scream and gunshot detection and localization for audio-surveillance systems," in *2007 IEEE Conference on Advanced Video and Signal Based Surveillance*, 2007, pp. 21–26.

[2] Q. N. Viet, H. Kang, S. Chung, S. Cho, K. Lee, and T. Seol, "Real-time audio surveillance system for ptz camera," in *2013 International Conference on Advanced Technologies for Communications (ATC 2013)*, 2013, pp. 392–397.

[3] S. Chandrakala and S. L. Jayalakshmi, "Generative model driven representation learning in a hybrid framework for environmental audio scene and sound event recognition," *IEEE Transactions on Multimedia*, vol. 22, no. 1, pp. 3–14, 2020.

[4] S. Dimitrov, "Analyzing sounds of home environment for device recognition," in *Proceedings of the European Conference on Ambient Intelligence*, 2014, pp. 1–16.

[5] J. Amores, "Multiple instance classification: Review, taxonomy and comparative study," *Artificial Intelligence*, vol. 201, no. aug., pp. 81–105, 2013.

[6] T. Su, J. Liu, and Y. Yang, "Weakly-supervised audio event detection using event-specific gaussian filters and fully convolutional networks," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 791–795.

[7] A. Shah, A. Kumar, A. G. Hauptmann, and B. Raj, "A closer look at weak label learning for audio events," *CoRR*, vol. abs/1804.09288, 2018. [Online]. Available: http://arxiv.org/abs/1804.09288

[8] Q. Kong, Y. Xu, W. Wang, and M. D. Plumbley, "Audio set classification with attention model: A probabilistic perspective," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 316–320.

[9] Q. Kong, Y. Xu, I. Sobieraj, W. Wang, and M. D. Plumbley, "Sound event detection and time–frequency segmentation from weakly labelled data," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 4, pp. 777–787, 2019.

[10] Q. Kong, Y. Xu, W. Wang, and M. D. Plumbley, "Sound event detection of weakly labelled data with cnn-transformer and automatic threshold optimization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2450–2460, 2020.

[11] S. Hong, Y. Zou, and W. Wang, "Gated Multi-Head Attention Pooling for Weakly Labelled Audio Tagging," in *Proc. Interspeech 2020*, 2020, pp. 816–820. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-1197

[12] 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, 2015.

[13] H. Phan, O. Y. Chén, L. Pham, P. Koch, M. D. Vos, I. McLoughlin, and A. Mertins, "Spatio-Temporal Attention Pooling for Audio Scene Classification," in *Proc. Interspeech 2019*, 2019, pp. 3845–3849.

[14] S. Liu, F. Yang, Y. Cao, and J. Yang, "Frequency-dependent auto-pooling function for weakly supervised sound event detection," *EURASIP J. Audio Speech Music. Process.*, vol. 2021, no. 1, p. 19, 2021. [Online]. Available: https://doi.org/10.1186/s13636-021-00206-7

[15] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "Cnn architectures for large-scale audio classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 131–135.

[16] R. Lu, Z. Duan, and C. Zhang, "Multi-scale recurrent neural network for sound event detection," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 131–135.

[17] Y. Xu, Q. Kong, W. Wang, and M. D. Plumbley, "Large-scale weakly supervised audio classification using gated convolutional neural network," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 121–125.

[18] J. Yan, Y. Song, W. Guo, L. Dai, I. McLoughlin, and L. Chen, "A region based attention method for weakly supervised sound event detection and classification," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 755–759.

[19] S. Hong, Y. Zou, W. Wang, and M. Cao, "Weakly labelled audio tagging via convolutional networks with spatial and channel-wise attention," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 296–300.

[20] Y. Wang, J. Li, and F. Metze, "A comparison of five multiple instance learning pooling functions for sound event detection with weak labeling," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 31–35.

[21] 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," 2017.

[22] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 776–780.

[23] J. Chung, Ç. Gülçehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *CoRR*, vol. abs/1412.3555, 2014. [Online]. Available: http://arxiv.org/abs/1412.3555

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1412.6980

[25] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proc. Interspeech 2019*, 2019, pp. 2613–2617. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-2680

[26] H. Zhang, M. Cissé, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *CoRR*, vol. abs/1710.09412, 2017. [Online]. Available: http://arxiv.org/abs/1710.09412