INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

SpeechMix - Augmenting Deep Sound Recognition using Hidden Space Interpolations

Amit Jindal[1] *, *Narayanan Elavathur Ranganatha*[1] *, *Aniket Didolkar*[1] *, *Arijit Ghosh Chowdhury*[1] *, *Di Jin* [2], *Ramit Sawhney* [3], *Rajiv Ratn Shah* [4]

[1]Manipal Academy of Higher Education, India
[2]Massachusetts Institute of Technology, USA
[3]Netaji Subhas Institute of Technology, India
[4]Indraprastha Institute of Information Technology, India

{amitj646, naruarjun, adidolkar123, arijit10}@gmail.com, jindi15@mit.edu,
ramits.co@nsit.net.in, rajivratn@iiitd.ac.in

Abstract

This paper presents SpeechMix, a regularization and data augmentation technique for deep sound recognition. Our strategy is to create virtual training samples by interpolating speech samples in hidden space. SpeechMix has the potential to generate an infinite number of new augmented speech samples since the combination of speech samples is continuous. Thus, it allows downstream models to avoid overfitting drastically. Unlike other mixing strategies that only work on the input space, we apply our method on the intermediate layers to capture a broader representation of the feature space. Through an extensive quantitative evaluation, we demonstrate the effectiveness of SpeechMix in comparison to standard learning regimes and previously applied mixing strategies. Furthermore, we highlight how different hidden layers contribute to the improvements in classification using an ablation study.

Index Terms: speech recognition, data augmentation, mixup, regularization

1. Introduction

Deep learning has achieved high performance in Speech Recognition tasks [1, 2, 3]. However, these deep neural networks tend to contain millions to billions of parameters, and are thus prone to overfitting, due to a lack of sufficient train data [4]. Techniques proposed for improving model generalization in Speech Recognition mostly fall under data augmentation or regularization methods and have proved to be effective. Some of these include altering the shape or property [5, 6] and generating external data for augmentation [7, 8, 9].

Mixup [10] is a data-agnostic augmentation technique that constructs virtual training examples by interpolating pairs of training samples from its vicinal distribution. It can be viewed as a data augmentation approach that creates new data samples based on the original training set. Mixup has been demonstrated to work well on image data [10, 11, 12, 13] and text classification [14, 15]. It has also been explored to improve Speech Recognition with methods such as linear interpolations [16, 17, 18]. Recent techniques like Between-Class Learning [19] mix the input signals by taking auditory perception of sounds into account to generate virtual samples. However, these methods have not explored mixing up meaningful regions of the feature space, which proves to be more effective in our work.

We introduce **SpeechMix**, a data augmentation technique for automatic speech recognition (ASR). It combines the latent representations of two or more samples from a neural model to provide additional training signals. Through experiments, we show improvements of this data augmentation method over standard learning methods that do not employ interpolation based training. More importantly, we also compare it against previous Mixup strategies that only utilize interpolations of the input signals such as the Between-Class learning, and empirically highlight the effectiveness of SpeechMix that owes to capturing a greater breadth of the feature space.

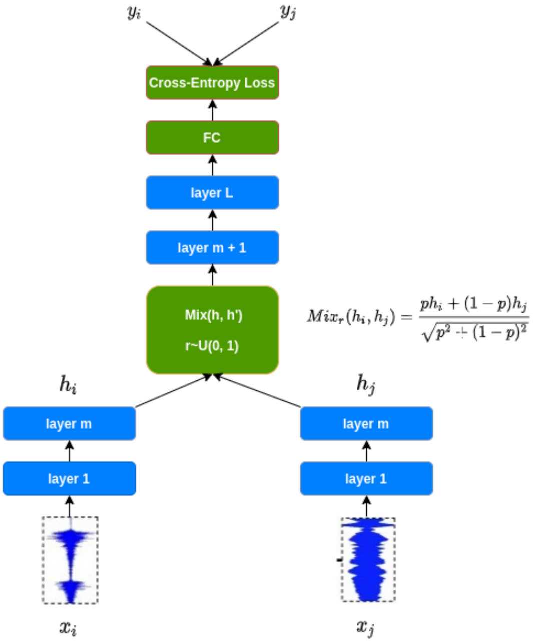

Figure 1: *SpeechMix takes as input two sound waves (x_i and x_j) and their corresponding labels y_i and y_j. The hidden representations at layer m are interpolated and the mixed representation is passed forward through the network.*

2. Related Work

In this section, we briefly go over existing relevant literature on deep speech recognition networks, data augmentation techniques, and interpolation based regularizers.

Copyright © 2020 ISCA

http://dx.doi.org/10.21437/Interspeech.2020-3147

2.1. Deep Speech Recognition

Deep learning has been extensively used for speech recognition in the recent past. Piczak [20] proposed to apply CNNs to the log-mel features extracted from raw waveforms. Aytar et al. [7] proposed a sound recognition network using 1- D convolutional and pooling layers named SoundNet and learned the audio feature using a large number of unlabeled videos. Tokozume and Harada [6] proposed a network using both 1-D and 2-D convolutional and pooling layers named EnvNet. Furthermore, Tokozume et al. [19] proposed an architecture called EnvNet-v2, with a higher number of layers and a higher sampling rate. We utilize these diverse and high performing architectures to perform a thorough comparison of our proposed method.

2.2. Data Augmentation for Speech

Similar to any other supervised machine learning task, one of the critical challenges to speech recognition is the lack of adequate volume of training data. One of the most standard and vital data augmentation methods is cropping [20, 7, 6, 13]. Salamon and Bello [5] proposed the usage of additional training data created by time stretching, pitch shifting, dynamic range compression, and adding background noise chosen from an external dataset. One major disadvantage of adding external data is the quality [21]. *SpecAugment* [22] is a recent data augmentation technique that consists of warping features, masking blocks of time steps, and masking blocks of frequency channels. Some methods also use acoustic data transformation techniques like audio signal speed alteration [23], applying noises, introduction of artificial reverberation into the records [24]. However, in noise based augmentation methods, the system often ends up learning more about the nature of the noise than the phonetic combinations present in the samples [21].

2.3. Interpolation Based Regularizers

Interpolation-based regularizers like Mixup [19, 11] have been successful for image classification problems, by overlaying two input images and combining image labels as virtual training data and have achieved state-of-the-art performances across a variety of tasks. For speech, in particular, linear interpolations of the input samples have been explored as regularization techniques [16, 17, 18]. Most notably, Tokozume et al. [19] have explored a mixing strategy called Between-Class learning (BC learning) using the *EnvNet-v2* architecture, and surpassed human performance on ESC-50 dataset [25].

We present **our work** within the same settings as Tokozume et al. [19]. We demonstrate how **SpeechMix**, a mixing strategy that utilizes the interpolation of hidden states, outperforms BC learning, and standard learning regimes throughout the experimental setup.

3. Methodology

In BC learning [19], mixup occurs in training examples before they are sent as input to the model. **SpeechMix** augments BC learning by employing mixup of hidden states as shown in Figure 1. We will describe how this is achieved and why this approach warrants a performance improvement over BC learning.

3.1. SpeechMix

The main idea of Mixup [10] is that given two labeled data points (x_i, y_i) and (x_j, y_j), where x_i and x_j are two samples of different classes randomly selected and y_i and y_j are the one-hot representation of the label. The algorithm creates virtual training samples by linear interpolation:

$$\tilde{x} = mix(x_i, x_j) = rx_i + (1-r)x_j \tag{1}$$

$$\tilde{y} = mix(y_i, y_j) = ry_i + (1-r)y_j \tag{2}$$

where $r \sim U(0,1)$ is the mixing ratio and $\tilde{y}$ is defined as the mixed label. Models that take these virtual training samples are usually trained to output the mixed label. When it comes to sound data, the Equation (3) should be adopted as a replacement of Equation (1) since it takes into account the relationship between energy and amplitude, i.e., sound energy is proportional to the square of the amplitude:

$$\tilde{x} = mix(x_i, x_j) = \frac{rx_i + (1-r)x_j}{\sqrt{r^2 + (1-r)^2}} \tag{3}$$

The mixup formula used in BC learning is derived by taking auditory perceptions of sound into account [19]. The mixing ratio r is transformed into the sound mixing ratio p and the mixup method is updated as follows:

$$mix(x_i, x_j) = \frac{px_i + (1-p)x_j}{\sqrt{p^2 + (1-p)^2}} \tag{4}$$

$$where\ p = \frac{1}{1 + 10^{\frac{G_i - G_j}{20}} \cdot \frac{1-r}{r}}$$

where G_i and G_j are the sound pressure levels of x_i and x_j in $[dB]$. These are calculated via A-weighting [27].

We propose **SpeechMix**, where the neural network is trained on interpolations of the hidden states. Let $g(., \theta)$ denote the classification model used, where θ denotes the model parameters. Assuming this model has M layers, we choose to mix the hidden representations at the m-th layer, $m \in [0, M]$. These interpolated hidden representations at layer m are fed to the upper layers using the previously discussed mixup strategy. Mathematically the m-th layer is denoted as $g_m(., \theta)$, hence the hidden representation of the m-th layer is $h_m = g_m(h_{m-1}, \theta)$. The 0-th layer is considered as the input layer. Hence, for two samples x_i and x_j, $h_0^i = x_i, h_0^j = x_j$ and the following hidden representations are as follows:

$$h_l^i = g_l(h_{l-1}^i, \theta), l \in [1, m] \tag{5}$$

$$h_l^j = g_l(h_{l-1}^j, \theta), l \in [1, m] \tag{6}$$

These hidden representations at the m-th layer are mixed using equation (4). We denote this mixed representation as $\tilde{h}_m$. Mixup at the m-th layer is thus defined as follows:

$$\tilde{h}_m = \frac{ph_m^i + (1-p)h_m^j}{\sqrt{p^2 + (1-p)^2}} \tag{7}$$

The continued forward pass after the mixed hidden representation has been generated is defined as follows:

$$\tilde{h}_l = g_l(\tilde{h}_{l-1}, \theta), l \in [m+1, M] \tag{8}$$

The layers chosen for mixup are denoted by S where $S = \{S_1, S_2, ...\}$ where each $S_i \in [0, M]$. The layer m where mixup occurs is chosen randomly from S with equal probability given to each layer in S and sampled separately for each pair of examples that are mixed.

Table 1: *Comparison between Standard Learning, BC Learning and SpeechMix using error rates(%). We performed 5-fold cross validation on ESC-10 and ESC-50 to show the standard error.*

		Error rates (%)		
Model	**Learning**	**ESC - 50**	**ESC - 10**	**UrbanSound**
SoundNet5 [7]	Standard Learning	33.8 ± 0.2	16.4 ± 0.8	33.3
	BC Learning	27.4 ± 0.3	13.9 ± 0.4	30.2
	SpeechMix (Ours)	**25.6 ± 0.2**	**11.6 ± 0.3**	**27.4**
M18 [26]	Standard Learning	31.5 ± 0.5	18.2 ± 0.5	28.8
	BC Learning	26.7 ± 0.1	14.2 ± 0.9	26.5
	SpeechMix (Ours)	**24.3 ± 0.2**	**12.4 ± 0.5**	**25.1**
EnvNet [6]	Standard Learning	29.2 ± 0.1	12.8 ± 0.4	33.7
	BC Learning	24.1 ± 0.2	11.3 ± 0.6	28.9
	SpeechMix (Ours)	**22.5 ± 0.3**	**9.3 ± 0.4**	**26.5**
PiczakCNN [20]	Standard Learning	27.6 ± 0.2	13.2 ± 0.4	25.3
	BC Learning	23.1 ± 0.3	9.4 ± 0.4	23.5
	SpeechMix (Ours)	**22.1 ± 0.3**	**8.4 ± 0.2**	**22.1**
EnvNet-v2 [6]	Standard Learning	25.6 ± 0.3	14.2 ± 0.8	30.9
	BC Learning	18.2 ± 0.2	10.6 ± 0.6	23.4
	SpeechMix (Ours)	**16.2 ± 0.3**	**8.5 ± 0.4**	**21.6**
EnvNet-v2 + Augmentation	Standard Learning	21.2 ± 0.3	10.9 ± 0.6	24.9
	BC Learning	15.1 ± 0.2	8.6 ± 0.1	21.7
	SpeechMix (Ours)	**13.1 ± 0.2**	**7.1 ± 0.1**	**20.8**
Human		18.7	4.3	-

3.2. Optimization

We denote n as the number of samples in a mini-batch, r as the mixing ratio, m as the layer at which mixup occurs and S as the set of layers eligible for mixup. For each mini-batch, m is sampled randomly from S. We consider two random mini-batches (x_i, y_i) and (x_j, y_j) of data . These two mini-batches undergo the SpeechMix process as described above in Equations (7) and (8) respectively. The mixed label is calculated according to Equation (2). We minimize the KL-divergence between the mixed label represented as $\tilde{y}$ and softmax of the generated outputs ($\tilde{h}_M$). The loss is as follows:

$$L = \frac{1}{n}\sum_{i=0}^{n} D_{KL}(\tilde{y}^i || softmax(\tilde{h}^i_M)) \tag{9}$$

$$where$$

$$D_{KL}(\tilde{y}^i || softmax(\tilde{h}^i_M)) = \sum_{k=0}^{c} \tilde{y}^i_k log \frac{\tilde{y}^i_k}{\{softmax(\tilde{h}^i_M)\}_k}$$

where c is the number of classes.

3.3. Probing SpeechMix: Why does it work?

Mixing of sounds in the input layer physically makes sense as humans can recognize two sounds and perceive which sound is more prominent in a digitally mixed sample. However, we also need to take into account how the machine perceives mixed data. CNNs can learn features directly from raw waveforms which is evident by their ability to filter out frequencies [6, 26, 28]. Therefore their activations encode higher level information. We could think of lower level activations of a DNN as speaker-adapted features [29], while the upper layer activations could be thought of as performing class-based discrimination. SpeechMix utilizes interpolations of hidden layers by adding them as training signals, which helps us use these features to regularize our classification model better. All points not observed during training that are in-between the class-representations end up being assigned low-confidence scores [11]. This obtains smoother decision boundaries at multiple levels of representation in the neural network. Hence, mixing of upper layers' activations rather than just the input layer will lead to a smoother decision boundary in that feature space.

Table 2: *Statistics of sound classification datasets.*

Dataset	**Classes**	**Samples**	**Duration**
UrbanSound8k	10	8732	9.7 hours
ESC-50	50	2000	2.8 hours
ESC-10	10	400	33 min

4. Experiments

4.1. Dataset and Preprocessing

We used ESC-50, ESC-10 [25], and UrbanSound8K [5] to train and evaluate the models, of which the statistics are summarized in Table 2. We pre-process the data as follows [19]:

Let $\mathbb{T}$ be the input length of a network. In the training phase, the sound is padded with $\mathbb{T}/2$ seconds of zeros on each side. A $\mathbb{T}$ second section is then randomly cropped from the padded sound. In the testing phase, $\mathbb{T}/2$ seconds of zeros was added as padding on each side of the sound. In contrast to the training phase, 10 $\mathbb{T}$ second sections were cropped from the padded sound at regular intervals. These 10 crops then act as the input to the network, and the outputs are combined via average pooling. Input data was regularized into a range of -1 to $+1$ by dividing it by 32,768, that is, the full range of 16-bit recordings.

4.2. Sound Classification Models

We compare SpeechMix with Standard Learning and BC Learning strategies[19] on various types of strong sound recognition models to show its general effectiveness.

EnvNet [6] is an end-to-end CNN for classification of environmental sounds. A fixed $\mathbb{T}$ second section of sound data sampled at 16 kHz is classified and is fed in as raw waveforms. The direction of convolution is switched in between, making the network convolve in both time and frequency. **EnvNet-v2** [19] is a extension of EnvNet. It is the same architecture but with a higher number of convolution layers and a higher sampling rate of 44kHz for the data. **SoundNet 5** [7] has a deep convolution architecture which transfers discriminative knowledge from visual recognition networks into sound networks. **M18** [26] is a very deep CNN that directly uses time-domain waveforms as input. It can efficiently optimize over very long sequences. **PiczakCNN** [20] is a CNN applied to two or three channels of data. The channels consist of the arrangement of log-mel features along the time axis, delta log-mel features i.e. the first temporal derivative of the (static) log-mel features.

4.3. Experimental Settings

We use the same hyperparameters as described by [19]. That is, we use Nesterovs accelerated gradient using a momentum of 0.9, weight decay of 0.0005, and mini-batch size of 64. We then use this setup for SpeechMix and BC learning to train it for twice the number of epochs compared with that in the Standard Learning. We also perform a 5-fold cross validation on the ESC-10 and ESC-50 datasets to show standard error rates (%). Scale augmentation with a factor randomly selected from $[0.8, 1.25]$ is used along with gain augmentation with a factor randomly selected from $[-6dB, +6dB]$. Scale augmentation is performed with EnvNet-v2 before zero padding using linear interpolation, and gain augmentation was performed just before inputting to the network.

4.4. Results

The results are summarized in Table 1. The proposed SpeechMix improves performance across all datasets and all architectures. Our best performing model is the Env-Net-v2 coupled with Augmentation (mentioned in Section 4.3) and SpeechMix. We obtain the highest relative improvements on the ESC-10 dataset, across all models. EnvNet-v2 combined with SpeechMix shows the highest improvement as compared to BC learning (19.8%) and standard learning (40.2%). Furthermore, for ESC-50 and UrbanSound8K, we obtain relative improvements of 10.98% and 7.69%.

Table 1 also highlights how SpeechMix complements existing data augmentation techniques. We obtain a % increase of 13.2%, 8.06% and 4.14% on ESC-50, ESC-10 and UrbanSound respectively. We also surpass human performance on ESC-50 by 42.7%. We empirically elucidate how a linear combination of higher level features helps in learning better class of representations irrespective of the model or dataset.

4.5. Ablation Study

Mixup Strategy: We compare different mixup formulae such as Mixup (Equation 1), BC learning (Equation 4) and SpeechMix (Equation 7). As shown in Table 3, accounting for the deeper activations using SpeechMix has a significant contribution to the performance.

SpeechMix Layers Set: When using SpeechMix, we randomly select a layer to perform mixup from a set of layers S. We investigate different sets of layers S for SpeechMix using EnvNet-v2 on ESC-10 and ESC-50 datasets and results are shown in Table 3. In our experiments for SpeechMix, the layers for mixing were: input (layer 0), the output from first max-pool (layer 1), second max-pool (layer 2), third max-pool (layer 3) and the fourth max-pool (layer 4). When no mixup is performed, the model error rate for ESC-50 and ESC-10 was 25.6% and 14.2% respectively. Our model achieves the best performance at $\{1, 2, 3\}$. This set of layer learns discriminative features such as frequency response which is quite similar to human perception [19].

Table 3: *Ablation analysis with EnvNet-v2 on ESC-50 and ESC-10. We report the average error rate of 5 trials.*

		Error rates (%)	
Comparison of	**Setting**	**ESC50**	**ESC10**
Mixup Strategy	Mixup [10]	20.1	12.5
	BC Learning [6]	18.2	10.6
	SpeechMix (Ours)	**16.2**	**8.5**
SpeechMix Layers Set	{0}	18.2	10.6
	{0, 1}	18.2	10.6
	{0, 1, 2}	17.9	9.25
	{0, 1, 2, 3}	16.5	9.5
	{0, 1, 2, 3, 4}	18.1	9.8
	{1}	16.3	10.5
	{1, 2}	18	11.5
	{1, 2, 3}	**16.2**	**8.5**
	{1, 2, 3, 4}	18.5	8.75
	{2}	18.75	9.1
	{2, 3}	18.25	11.5
	{2, 3, 4}	18.5	10.5
	{3}	18.75	9.5
	{3, 4}	18.25	10.7
Standard Learning		25.6	14.2

5. Conclusion

We proposed SpeechMix a data augmentation and regularization method for deep sound recognition. SpeechMix is a generalization of BC learning as it interpolates latent representations of hidden states. We validate the effectiveness of SpeechMix as a regularizer and show improvement across various architectures on datasets of different scale and class distributions. We demonstrate that SpeechMix learns a better discriminative feature space over existing augmentation approaches.

6. References

[1] L. Deng, G. Hinton, and B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: An overview," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 8599–8603.

[2] P. Harár, R. Burget, and M. K. Dutta, "Speech emotion recognition with deep learning," in *2017 4th International Conference on Signal Processing and Integrated Networks (SPIN)*. IEEE, 2017, pp. 137–140.

[3] D. Yu, G. Hinton, N. Morgan, J.-T. Chien, and S. Sagayama, "Introduction to the special section on deep learning for speech and language processing," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 4–6, 2011.

[4] D. Zou and Q. Gu, "An improved analysis of training over-parameterized deep neural networks," in *Advances in Neural Information Processing Systems*, 2019, pp. 2053–2062.

[5] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, 2017.

[6] Y. Tokozume and T. Harada, "Learning environmental sounds with end-to-end convolutional neural network," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2721–2725.

[7] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," in *Advances in neural information processing systems*, 2016, pp. 892–900.

[8] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5220–5224.

[9] W.-N. Hsu, Y. Zhang, and J. Glass, "Unsupervised domain adaptation for robust speech recognition via variational autoencoder-based data augmentation," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 16–23.

[10] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[11] V. Verma, A. Lamb, C. Beckham, A. Courville, I. Mitliagkis, and Y. Bengio, "Manifold mixup: Encouraging meaningful on-manifold interpolation as a regularizer," *stat*, vol. 1050, p. 13, 2018.

[12] Y. Tokozume, Y. Ushiku, and T. Harada, "Between-class learning for image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5486–5494.

[13] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6023–6032.

[14] H. Guo, "Nonlinear mixup: Out-of-manifold data augmentation for text classification."

[15] H. Guo, Y. Mao, and R. Zhang, "Augmenting data with mixup for sentence classification: An empirical study," *arXiv preprint arXiv:1905.08941*, 2019.

[16] I. Medennikov, Y. Y. Khokhlov, A. Romanenko, D. Popov, N. A. Tomashenko, I. Sorokin, and A. Zatvornitskiy, "An investigation of mixup training strategies for acoustic models in asr." in *Interspeech*, 2018, pp. 2903–2907.

[17] Y. Zhu, T. Ko, and B. Mak, "Mixup learning strategies for text-independent speaker verification," *Proc. Interspeech 2019*, pp. 4345–4349, 2019.

[18] Z. Zhang, S. Xu, S. Cao, and S. Zhang, "Deep convolutional neural network with mixup for environmental sound classification," in *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*. Springer, 2018, pp. 356–367.

[19] Y. Tokozume, Y. Ushiku, and T. Harada, "Learning from between-class examples for deep sound recognition," *arXiv preprint arXiv:1711.10282*, 2017.

[20] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2015, pp. 1–6.

[21] J. Ramrez Snchez, M. Bereau, and J. Lara, *A Survey of the Effects of Data Augmentation for Automatic Speech Recognition Systems*, 10 2019, pp. 669–678.

[22] 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," *arXiv preprint arXiv:1904.08779*, 2019.

[23] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[24] V. Peddinti, G. Chen, D. Povey, and S. Khudanpur, "Reverberation robust acoustic modeling using i-vectors with time delay neural networks," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[25] K. J. Piczak, "Esc: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 1015–1018.

[26] W. Dai, C. Dai, S. Qu, J. Li, and S. Das, "Very deep convolutional neural networks for raw waveforms," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 421–425.

[27] IEC, "Electroacoustics - sound level meters - part 1: Specifications," 2013.

[28] T. N. Sainath, R. J. Weiss, A. W. Senior, K. W. Wilson, and O. Vinyals, "Learning the speech front-end with raw waveform cldnns," in *INTERSPEECH*, 2015.

[29] A. Rahman Mohamed, G. E. Hinton, and G. Penn, "Understanding how deep belief networks perform acoustic modelling." in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 4273–4276.