INTERMIX: AN INTERFERENCE-BASED DATA AUGMENTATION AND REGULARIZATION TECHNIQUE FOR AUTOMATIC DEEP SOUND CLASSIFICATION

Ramit Sawhney

Scaler, India
ramit.sawhney@scaler.com

Atula Tejaswi Neerkaje

Manipal Institue of Technology, India
atula.neerkaje@learner.manipal.edu

ABSTRACT

In this paper, we present InterMix, an interference-based regularization and data augmentation strategy for automatic sound classification. InterMix creates virtual training examples by creating an interference-based mixed representation for a sampled phase difference and mixup ratio. InterMix can be used to train sound classification models with the ability to generate a vast amount of training samples. These are significantly varied compared to that of other mixup strategies due to the introduction of phase difference, a continuous variable. While building on other mixup strategies which use linear interpolation, we perform mixup based on the formula of interference. We demonstrate the utility of InterMix in comparison to standard learning techniques and previously applied mixing strategies through a quantitative analysis. We also demonstrate that InterMix is more robust towards adversarial attacks compared to standard learning and other mixup strategies.

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

1. INTRODUCTION

Deep learning has demonstrated exceptional performance in speech-recognition tasks in the recent past [1]. Large networks with billions of parameters have a broad hypothesis space, and will overfit if the size of the training set is not large enough [2]. Techniques that improve model generalization like data augmentation and regularization have proven to be effective in speech recognition tasks [3], and other tasks like image classification [4]. Examples of such methods include altering shape or property [5, 6], and generating external data for augmentation [7, 8, 9]. Mixup [10] is a data-agnostic augmentation technique and a form of vicinal risk minimization using virtual training samples. It can be viewed as a data augmentation approach that creates new data samples based on linear interpolation of the original training data. Mixup has shown to work well on image data [10, 11, 12], text classification [13] and speech recognition [14]. Other recent methods like between-class learning [6] mix the input signals by taking auditory perception of sounds into account. While this method has shown to be effective, it does not explore the concept of mixup based on the phase difference between sounds, which results in varied sound waves due to interference (Figure 1).

Another significant concern with deep models is that their input-output mappings are discontinuous to a significant extent [15], potentially making them vulnerable to adversarial examples. Adversarial examples are samples with imperceptible perturbations which can deceive a classifier into a wrong prediction [15]. Similar concerns arise in sound recognition [16, 17], a domain with a variety of applications [18]. Moreover, such networks are trained on crowd-sourced data, which may contain sensitive information, and could potentially be targeted by membership inference [19] and model inversion [20] attacks.

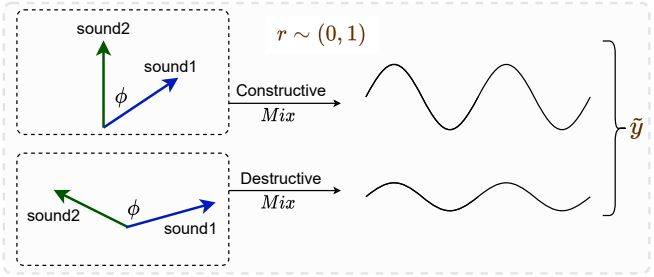

Fig. 1. Interference-based learning involves the mixing of sounds controlled by a mixing ratio r, with the sounds having a phase difference ϕ. The phase differences create varied representations of the same mixed sounds, with mixed label $\tilde{y}$.

We present **InterMix**, an interference-based data augmentation and regularization technique for automatic sound classification. InterMix involves an interference-based combination of hidden representations in a neural network to create virtual training signals. We show improvements on using this data augmentation and regularization method over standard learning methods that do not leverage mixup based training. We also compare it against other mixup strategies such as between-class learning, and highlight the effectiveness of InterMix that captures the property of interference in the feature space, while providing better privacy safeguards.

2. RELATED WORK

In recent years, deep learning has been widely used for speech recognition. Dai et al. [21] proposed a deep CNN that uses raw time-domain waveforms as inputs. Aytar et al. [7] proposed a sound recognition network using 1-D convolutional and pooling layers named SoundNet and learned the audio features using a large corpus of unlabeled videos. Piczak et al. [22] proposed to apply CNNs to the log-mel features extracted from raw waveforms. Li et al. [23] proposed DS-CNN which consisted of two stacked CNNs trained on log-mel inputs. Tokozume and Harada [24] introduced EnvNet, a network that uses both 1-D and 2-D convolutional and pooling layers. Further, Tokozume et al. [6] proposed an architecture called EnvNet-v2, with more layers and a higher sampling rate.

One of the challenges to supervised learning is the scarcity of labeled training examples. Data Augmentation is a frequently used technique to increase the diversity of training samples without collecting new data. For image data, some common techniques are

zooming, flipping, shifting and distortion [25]. In speech recognition, one of the most standard data augmentation methods is cropping [24, 22]. Kanda et al. [26] explored an elastic spectral distortion-based augmentation method. Other methods include Vocal Tract Length Perturbation (VTLP) [27], Stochastic Feature Mapping (SFM) [28] and other acoustic data transformation techniques like audio signal speed alteration [9], and applying noises and artificial reverberation into the records [29]. Further, Salamon and Bello [5] used augmentation techniques such as time stretching, pitch shifting, dynamic range compression, and adding background noise chosen from an external dataset. Notably, Park et al. [30] introduced SpecAugment, which consists of warping features, masking blocks of time steps, and masking blocks of frequency channels. Moreover, previous work has shown that privacy risk in deep models is relieved to a considerable extent due to augmentation methods [31].

Interpolation-based augmentation and regularization techniques like Mixup [6, 12] have achieved state-of-the-art performances across a variety of tasks. The essence of Mixup involves taking a weighted average of two input samples and their respective one-hot labels as virtual training data [14]. Tokozume et al. [6] and Jindal et al. [32] explore such mixup strategies using the EnvNet-v2 architecture, which surpassed human performance on the ESC-50 dataset [33]. In **our work**, we build on the same base settings and seek to show that a mixup technique involving an interference-based formula, which relies on the creation of phase differences, broadens the feature space and improves regularization. It provides better privacy safeguards by creating varied training signals that maintain the properties of the mixed sounds.

3. METHODOLOGY

In BC Learning [6], the mixup is based on linear interpolation by taking the auditory perceptions of sound into account. Speechmix [32] goes one step further and performs interpolation in the hidden space. **InterMix** builds on these settings to perform an interference-based mixup, based on the phase difference between the sounds.

3.1. InterMix

Given are two labelled data points (x_i, y_i) and (x_j, y_j), belonging to different classes, y_i and y_j, represented by one-hot encoding. The Mixup [10] of these two samples is defined as follows:

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

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

where $r \sim U(0,1)$ is the mixing ratio. By taking into account the auditory perceptions of sound and the relation between energy and amplitude [6, 32], the sound mixing ratio p is obtained. The revised mixup equation using p is given as:

$$\begin{aligned} mix(x_i, x_j) &= \frac{px_i + (1-p)x_j}{\sqrt{p^2 + (1-p)^2}} \\ p &= \frac{1}{1 + 10^{\frac{G_i - G_j}{20} \cdot \frac{1-r}{r}}} \end{aligned} \tag{3}$$

Here, G_i and G_j are sound pressure levels of x_i and x_j in dB, calculated using A-weighting [34].

Given a model $f(x, \theta)$ with M layers, where the model parameters are represented by θ, we propose interference-based mixing which happens at the m^{th} layer, $m \in [0, M]$. The layer m is sampled randomly for each batch of example pairs with equal probability from a given layer set $L = \{L_1, L_2, \cdots\}$ where $L_i \in [0, M]$.

First, we introduce phase shifts $\phi_i \sim U[-\pi/2, \pi/2]$ to x_i, and $\phi_j \sim U[-\pi/2, \pi/2]$ to x_j, to obtain phase shifted data samples $\tilde{x}_i$ and $\tilde{x}_j$ respectively. Following existing work [32], we choose to mix the samples at the m-th layer. Let the m-th layer be denoted as $f_m(h_{m-1}, \theta)$, where h_{m-1} is the feature representation from the previous layer. Then, for two inputs $\tilde{x}_i$ and $\tilde{x}_j$, the hidden representations at the m-th layer are h_m^i and h_m^j, respectively. These hidden representations are first weighted by p, as given by:

$$\begin{aligned} \widetilde{h}_m^i &= \frac{ph_m^i}{\sqrt{p^2 + (1-p)^2}} \\ \widetilde{h}_m^j &= \frac{(1-p)h_m^j}{\sqrt{p^2 + (1-p)^2}} \end{aligned} \tag{4}$$

Since our approach takes in raw audio signals as input, the signal phase does not get discarded [35]. Hence, we mix the two representations on the manifold characterized by the formula of interference, by taking into account the phase difference $\phi = \phi_i - \phi_j$, given as:

$$\widetilde{h}_m = \widetilde{h}_m^i + \widetilde{h}_m^j + 2\sqrt{\widetilde{h}_m^i \widetilde{h}_m^j} \cos\phi \tag{5}$$

3.2. Optimization

Let N denote the batch size, M the number of layers, and C the number of classes. The mixed label is computed using Equation 2. We follow existing work [32, 11] and minimize the KL divergence between the mixed label and the softmax of the final output layer $\widetilde{h}_M$. The loss function, $\mathcal{L}$ is as follows:

$$\mathcal{L} = \frac{1}{N} \sum_{i=0}^{N} D_{KL}(\widetilde{y}^i || softmax(\widetilde{h}_M^i)) \tag{6}$$

$$D_{KL}(\widetilde{y}^i || softmax(\widetilde{h}_M^i)) = \sum_{j=0}^{C} \widetilde{y}_j^i log \frac{\widetilde{y}_j^i}{\{softmax(\widetilde{h}_M^i\}_j}$$

3.3. Why InterMix Works

Linear interpolation based manifold mixup creates smoother boundaries in the feature space between two class distributions [12]. However, interference-based mixup provides augmented samples by mixing with a varied range of phase differences, which creates representations of varied feature intensities while maintaining the implicit properties of the mixed sounds, and thus provides a better regularizing effect. InterMix has the potential to create a larger number of training signals which are significantly varied compared to other mixup methods. As a result, the effect of sensitive crowd-sourced data while training is further minimized, which makes InterMix significantly robust towards adversarial samples (Section 5.3).

4. EXPERIMENTS

4.1. Datasets and Preprocessing

The datasets used in the experiments are ESC-10, ESC-50 [33] and UrbanSound8K [5]. ESC-50 consists of 2000 samples belonging to

Table 1. We tabulate the error rates to present a comparison between Standard Learning, BC Learning [6], Speechmix [32], and InterMix. For the ESC-50 and ESC-10 datasets, we performed 5-fold cross-validation to show the standard error.

Model	Learning	Error Rates (%)		
		ESC-50	**ESC-10**	**UrbanSound8K**
M18 [21]	Standard	31.5±0.5	18.2±0.5	28.8
	BC Learning	26.7±0.1	14.2±0.9	26.5
	Speechmix	**24.3±0.2**	**12.4±0.5**	25.1
	InterMix (Ours)	25.4±0.5	12.6±0.5	**25.1**
SoundNet5 [7]	Standard	33.8±0.2	16.4±0.8	33.3
	BC Learning	27.4±0.3	13.9±0.4	30.2
	Speechmix	25.6±0.2	11.6±0.3	27.4
	InterMix (Ours)	**25.1±0.3**	**10.6±0.3**	**26.5**
EnvNet [24]	Standard	29.2±0.1	12.8±0.4	33.7
	BC Learning	24.1±0.2	11.3±0.6	28.9
	Speechmix	22.5±0.3	9.3±0.4	**26.5**
	InterMix (Ours)	**22.5±0.3**	**9.1±0.2**	26.8
PiczakCNN [22]	Standard	27.6±0.2	13.2±0.4	25.3
	BC Learning	23.1±0.3	9.4±0.4	23.5
	Speechmix	22.1±0.3	8.4±0.2	22.1
	InterMix (Ours)	**21.9±0.2**	**8.3±0.4**	**21.1**
EnvNet-v2 [6]	Standard	25.6±0.3	14.2±0.8	30.9
	BC Learning	18.2±0.2	10.6±0.6	23.4
	Speechmix	16.2±0.3	8.5±0.4	21.6
	InterMix (Ours)	**15.8±0.4**	**8.2±0.4**	**21.4**
EnvNet-v2 +Augmentation	Standard	21.2±0.3	10.9±0.6	24.9
	BC Learning	15.1±0.2	8.6±0.1	21.7
	Speechmix	13.1±0.2	**7.1±0.1**	20.8
	InterMix (Ours)	**12.9±0.4**	7.2±0.1	**20.5**
Human		18.7	4.3	

50 classes, while its subset, ESC-10 consists of 400 samples belonging to 10 classes. UrbanSound8K consists of 8732 samples belonging to 10 classes. The pre-processing steps [6] are given as: Given T as the input length of the model. Each sound is padded with $T/2$ seconds of zeros on each side. During the training phase, a T second section is then randomly cropped. In the testing phase, $10T$ sections were cropped as inputs to the model, with the outputs combined via average pooling. The input data was regularized into a range of -1 to +1 by dividing it by 32,768, the full range of 16-bit recordings.

4.2. Sound Classification Models

We compare InterMix with Standard Learning, BC Learning [6], and Speechmix [32] on the sound classification models given below.

- SoundNet5 [7]: Deep CNN which transfers discriminative visual knowledge from well-established visual recognition models into the sound modality.
- M18 [21]: Deep CNN with time-domain wave-transforms as input to the model.
- PiczakCNN [22]: CNN applied to two to three channels of data, which consist of the arrangement of long-mel features along with the time axis and delta long-mel features.
- EnvNet [24]: End-to-end CNN for environmental sound classification. A fixed sample of T seconds sampled at 16kHz is fed to the network. To make the network convolve in both time and frequency, the direction of the convolution is switched in between.
- EnvNet-v2 [6]: Similar to EnvNet, with a higher sampling rate of 44kHz and with more convolutions.

4.3. Experimental Settings

We use Nesterov's accelerated gradient with a momentum of 0.9, weight decay of 0.0005, and batch size 32, as described in [6, 32]. This set-up is used for BC learning, Speechmix, and InterMix, with the number of epochs twice as that for Standard learning. We perform 5-fold cross-validation on the ESC-10 and ESC-50 datasets to obtain standard error rates (%). Scale augmentation with a factor randomly sampled from $[0.8, 1.25]$ is used along with gain augmentation factor sampled from $[-6dB, +6dB]$ before zero padding and before inputting to the network, respectively. We found the best performing layer set as $L = \{1, 2, 3, 4\}$.

5. RESULTS

5.1. Performance Comparison

We compare the results in Table 1. We observe that InterMix generally outperforms other techniques across the given models and datasets. The best-performing model is the augmented EnvNet-v2. The highest relative improvement is observed in the ESC-10 dataset

Table 2. Ablation study of InterMix components on the EnvNet-v2+Augmentation architecture.

Variant	Error Rates (%)		
	ESC-50	**ESC-10**	**UrbanSound8K**
Vanilla InterMix	13.6	7.9	21.0
+Hidden Mixup	13.3	7.7	20.8
+p-weighting	12.9	7.2	20.5

Table 3. We tabulate the error rates for adversarial examples from the ESC-10 dataset for various values of epsilon (ϵ).

Learning	Epsilon (ϵ)			
	$\mathbf{1e^{-2}}$	$\mathbf{1e^{-3}}$	$\mathbf{1e^{-4}}$	$\mathbf{1e^{-5}}$
Standard	64.5	37.7	36.5	36.2
BC Learning	47.2	28.7	26.9	26.8
InterMix (Ours)	**42.9**	**26.2**	**25.1**	**25.0**

for EnvNet-v2, with a relative improvement of 42.2% with respect to standard learning and 22.6% with respect to BC learning. For ESC-50 and UrbanSound8K, the relative improvements with respect to BC learning are 13.2% and 8.5% respectively. We also observe relative improvements of 2.4% and 0.9% on these datasets with respect to Speechmix. When InterMix is trained with augmentation, we observe relative improvements of 14.6%, 16.3%, and 5.5% with respect to BC learning on ESC-50, ESC-10, and UrbanSound8K.

5.2. Ablation Study

In Table 2, we study the impact of individual components of InterMix. The Vanilla InterMix variant consists of the normal r-weighting scheme, with interference-based mixing performed in the input space. We note significant performance improvements with the addition of p-weighting, and our observations align with existing studies on mixup strategies [32, 6] which suggest the importance of considering the difference in sound pressure levels between the samples. We further observe that performing the interference-based mixup in the hidden space offers additional performance improvements, likely because the hidden layers provide a greater regularizing effect by smoothening the mixed representations [32].

5.3. Robustness Towards Adversarial Attacks

We perform a white-box attack [31] using the Fast Gradient Sign Method (FGSM) [36] on the trained models. This method uses the gradients of the loss function with respect to the input, scaled by a constant epsilon (ϵ) to produce an adversarial example that potentially causes a misclassification. The constant epsilon (ϵ) controls how noisy the adversarial sample is. Given an input x, original label y, model parameters θ, cost function $\mathcal{L}$, and a small constant epsilon ϵ, the adversarial sample x_{adv}, is obtained as follows:

$$x_{adv} = x + \epsilon \cdot sign(\nabla_x \mathcal{L}(\theta, x, y)) \tag{7}$$

We feed the adversarial examples to the trained models and evaluate their error rates for various values of epsilon (ϵ), as summarized in Table 3. We observe that mixing-based learning techniques are significantly robust towards such adversarial attacks, as a consequence of learning rich between-class feature spaces. InterMix provides a further regularizing effect by creating phase differences between the sounds, and then by performing mixup using an interference-based formula. Further, we observe that increase in error rates for higher values of ϵ is lower for InterMix compared to other learning strategies. The robustness towards adversaries shows that InterMix reduces the reliance on sensitive training data by using virtual samples.

6. CONCLUSION

We introduced InterMix, an interference-based learning strategy that uses the concept of phase differences to create varied mixed representations of training signals. It can outperform data augmentation and regularization techniques which use linear interpolation based mixup. We further observed InterMix can learn models which are more robust towards adversarial attacks, through the introduction of phase difference to create varied mixed representations that do not belong in the training set, which improves generalization and has a shielding effect on sensitive training data.

7. REFERENCES

[1] Juncheng Li, Wei Dai, Florian Metze, Shuhui Qu, and Samarjit Das, "A comparison of deep learning methods for environmental sound detection," in *2017 IEEE International conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 126–130.

[2] Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang, "Learning and generalization in overparameterized neural networks, going beyond two layers," *arXiv preprint arXiv:1811.04918*, 2018.

[3] Thai-Son Nguyen, Sebastian Stueker, Jan Niehues, and Alex Waibel, "Improving sequence-to-sequence speech recognition training with on-the-fly data augmentation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7689–7693.

[4] Connor Shorten and Taghi M Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, pp. 1–48, 2019.

[5] Justin Salamon and Juan Pablo 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] Yuji Tokozume, Yoshitaka Ushiku, and Tatsuya Harada, "Learning from between-class examples for deep sound recognition," *arXiv preprint arXiv:1711.10282*, 2017.

[7] Yusuf Aytar, Carl Vondrick, and Antonio Torralba, "Soundnet: Learning sound representations from unlabeled video," *arXiv preprint arXiv:1610.09001*, 2016.

[8] Wei-Ning Hsu, Yu Zhang, and James 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.

[9] Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "Audio augmentation for speech recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[10] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

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

[12] Vikas Verma, Alex Lamb, Christopher Beckham, Aaron Courville, Ioannis Mitliagkis, and Yoshua Bengio, "Manifold mixup: Encouraging meaningful on-manifold interpolation as a regularizer," *arXiv preprint arXiv:1806.05236*, vol. 7, 2018.

[13] Hongyu Guo, "Nonlinear mixup: Out-of-manifold data augmentation for text classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 4044–4051.

[14] Yingke Zhu, Tom Ko, and Brian Mak, "Mixup learning strategies for text-independent speaker verification.," in *Interspeech*, 2019, pp. 4345–4349.

[15] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[16] Moustafa Alzantot, Bharathan Balaji, and Mani Srivastava, "Did you hear that? adversarial examples against automatic speech recognition," *arXiv preprint arXiv:1801.00554*, 2018.

[17] Nicholas Carlini and David Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," in *2018 IEEE Security and Privacy Workshops (SPW)*. IEEE, 2018, pp. 1–7.

[18] Vinod Subramanian, Emmanouil Benetos, Ning Xu, SKoT McDonald, and Mark Sandler, "Adversarial attacks in sound event classification," *arXiv preprint arXiv:1907.02477*, 2019.

[19] Michael Lomnitz, Nina Lopatina, Paul Gamble, Zigfried Hampel-Arias, Lucas Tindall, Felipe A Mejia, and Maria Alejandra Barrios, "Reducing audio membership inference attack accuracy to chance: 4 defenses," *arXiv preprint arXiv:1911.01888*, 2019.

[20] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, 2015, pp. 1322–1333.

[21] Wei Dai, Chia Dai, Shuhui Qu, Juncheng Li, and Samarjit 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.

[22] Karol 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.

[23] Shaobo Li, Yong Yao, Jie Hu, Guokai Liu, Xuemei Yao, and Jianjun Hu, "An ensemble stacked convolutional neural network model for environmental event sound recognition," *Applied Sciences*, vol. 8, no. 7, pp. 1152, 2018.

[24] Yuji Tokozume and Tatsuya 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.

[25] Luis Perez and Jason Wang, "The effectiveness of data augmentation in image classification using deep learning," *arXiv preprint arXiv:1712.04621*, 2017.

[26] Naoyuki Kanda, Ryu Takeda, and Yasunari Obuchi, "Elastic spectral distortion for low resource speech recognition with deep neural networks," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*. IEEE, 2013, pp. 309–314.

[27] Navdeep Jaitly and Geoffrey E Hinton, "Vocal tract length perturbation (vtlp) improves speech recognition," in *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language*, 2013, vol. 117.

[28] Xiaodong Cui, Vaibhava Goel, and Brian Kingsbury, "Data augmentation for deep neural network acoustic modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 9, pp. 1469–1477, 2015.

[29] Vijayaditya Peddinti, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Reverberation robust acoustic modeling using i-vectors with time delay neural networks," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[30] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[31] Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, Yann Ollivier, and Hervé Jégou, "White-box vs black-box: Bayes optimal strategies for membership inference," in *International Conference on Machine Learning*. PMLR, 2019, pp. 5558–5567.

[32] Amit Jindal, Narayanan Elavathur Ranganatha, Aniket Didolkar, Arijit Ghosh Chowdhury, Di Jin, Ramit Sawhney, and Rajiv Ratn Shah, "Speechmix-augmenting deep sound recognition using hidden space interpolations," *Proc. Interspeech 2020*, pp. 861–865, 2020.

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

[34] IEC IEC, "61672-1: 2013 electroacoustics-sound level meters-part 1: Specifications," 2013.

[35] Jakob Abeßer, "A review of deep learning based methods for acoustic scene classification," *Applied Sciences*, vol. 10, no. 6, 2020.

[36] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.