# InterMix: An Interference-based Data Augmentation And Regularization Technique For Automatic Deep Sound Classification

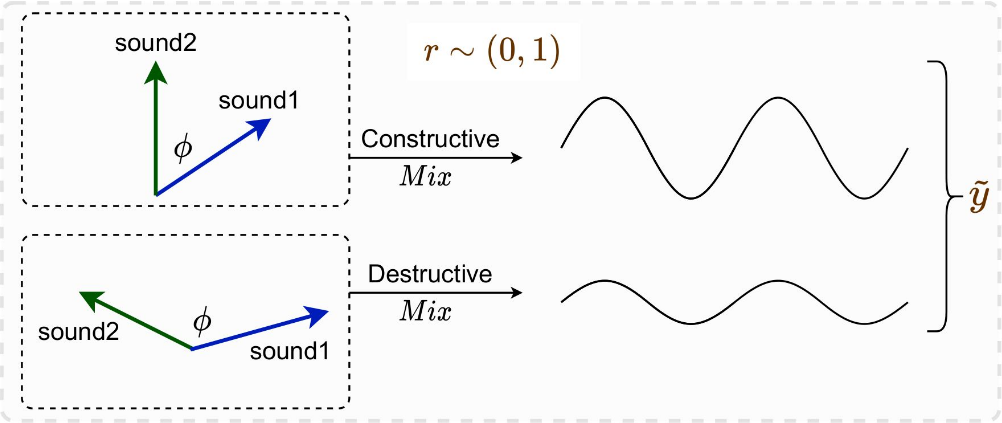
Ramit Sawhney[1], Atula Tejaswi Neerkaje[2]

[1]Scaler, India & [2]Manipal Institute of Technology, India

# This talk in a nutshell

- Interpolation-based augmentation and regularization techniques like Mixup have achieved state-of-the-art performances on various tasks.
- Building on existing work, we present InterMix, an interference-based data augmentation technique for automatic sound classification.
- We compare InterMix against other mixup strategies and highlight the effectiveness of InterMix that uses an interference formula, while simultaneously providing improved privacy safeguards.

# Motivation




Sounds: $(h_1,h_2)$
Labels: $(y_1,y_2)$

Mixup: $r*h_1 + (1-r)*h_2$
$r*y_1 + (1-r)*y_2$

- Mixup involves taking a weighted average of two inputs with their corresponding labels.
- While existing methods of mixup based augmentation have shown to be effective, these do not explore the concept of mixup based on the phase difference between sounds, which results in varied sound waves due to interference

# The approach: InterMix

**How it Works**

- InterMix first introduces phase shifts ($\Phi_1$,$\Phi_2$) to two randomly sampled sounds.
- Next, InterMix mixes the representations of these two sounds using an interference-based formula, taking into account the phase difference ($\Phi_1$- $\Phi_2$).

$$\widetilde{h}_m = \widetilde{h}^i_m + \widetilde{h}^j_m + 2\sqrt{\widetilde{h}^i_m \widetilde{h}^j_m} \cos\phi$$

where,

$$\widetilde{h}^i_m = \frac{p h^i_m}{\sqrt{p^2 + (1-p)^2}}$$

$$\widetilde{h}^j_m = \frac{(1-p) h^j_m}{\sqrt{p^2 + (1-p)^2}}$$

**Why it Works**

- Interference-based mixup provides augmented samples by mixing with a varied phase differences.
- InterMix creates large number of varied training signals - effect of sensitive crowd-sourced data while training is minimized.

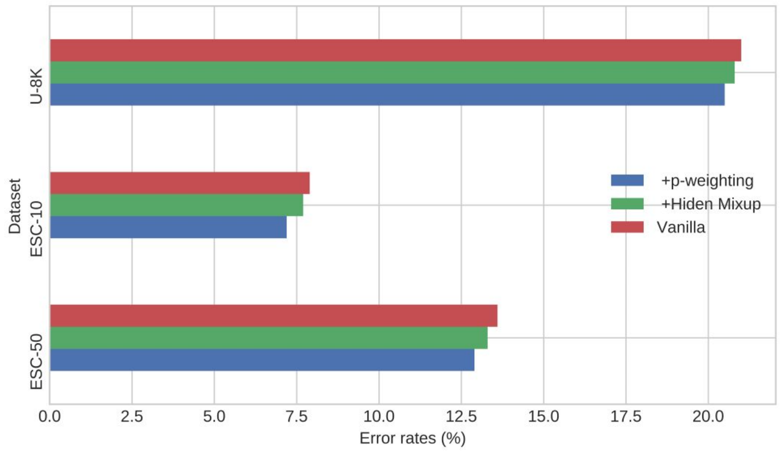
# Performance Comparison

| 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 | |

- InterMix generally outperforms existing techniques across the given models and datasets.
- The best-performing model is the augmented EnvNet-v2.
- We observe relative improvements of 14.6%, 16.3%, and 5.5% with respect to BC learning on ESC-50, ESC-10, and UrbanSound8K respectively

# Ablation: Impact of p-weighting and hidden space

| 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 |

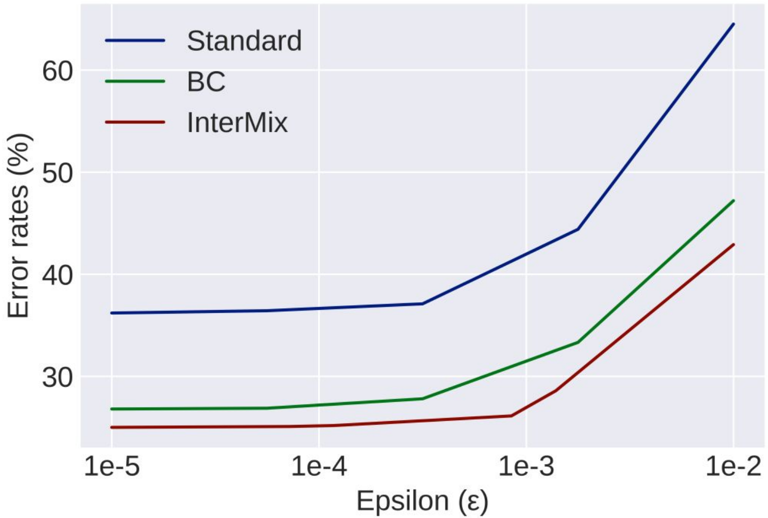



- We observe significant performance improvements on introducing both hidden space mixing and p-weighting.

- Suggests the effectiveness of mixing in the hidden space, and also the importance of considering the difference in sound pressure levels.

# Adversarial Robustness

| Learning | Epsilon ($\epsilon$) | | | |
|---|---|---|---|---|
| | $1e^{-2}$ | $1e^{-3}$ | $1e^{-4}$ | $1e^{-5}$ |
| 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** |




- InterMix provides better privacy safeguards through an improved regularizing effect.

- InterMix reduces the reliance on sensitive training data by using virtual samples.

# Conclusion

- We introduced InterMix, an interference-based learning strategy that uses the concept of phase differences to create varied mixed representations of training signals .

- InterMix achieves competitive performance in comparison to other regularization techniques which use linear interpolation based mixup.

- We further observed InterMix can learn models which are more robust towards adversarial attacks, which improves generalization and has a shielding effect on sensitive training data