Brayden de Koning

# Research: Advanced Audio Augmentation for Robustness and Generalisation

Scope  
The purpose of this research is to evaluate the existing audio augmentation techniques in the current codebase, identify potential improvements, propose two additional advanced augmentation methods, and recommend the most effective option for improving model robustness and generalisation. The analysis is based on current literature and competitive practice within speech, bioacoustic, and sound event detection domains.

Current Methods in the Codebase  
The existing augmentation methods comprise waveform-level and spectrogram-level transformations.

**Waveform-level augmentations include:**

* AddGaussianNoise: Introduces low-level random noise to the waveform to improve robustness against recording noise. It is computationally efficient and useful for reducing overfitting, but it is limited in realism as most environmental noise is non-stationary and does not match the characteristics of Gaussian noise.
* TimeStretch and PitchShift: Adjust temporal and pitch characteristics without affecting other acoustic parameters, allowing the model to generalise across variations in tempo and pitch. However, these do not address variability in acoustic environments or overlapping events.
* Shift: Alters waveform onset alignment to reduce temporal sensitivity. While effective for onset variations, it does not address environmental domain shifts.

**Spectrogram-level augmentations include:**

* SpecAugment (time and frequency masking): Randomly removes sections of the mel-spectrogram to improve robustness to missing cues. Although this has proven effectiveness, it does not simulate environmental propagation effects such as reverberation or realistic polyphonic interference.
* Legacy masking methods (tfio.audio.time\_mask / freq\_mask): Early implementations of spectrogram masking applied during preprocessing. These are no longer the primary augmentation methods in the pipeline

**Gaps in Current Approach**

The current augmentation stack addresses variations in pitch, tempo, onset alignment, and partial spectral occlusions. However, it does not adequately model environmental variability such as reverberation, microphone channel response, realistic non-stationary background noise, or overlapping events. Studies in robust automatic speech recognition and bioacoustics have shown that these factors significantly impact generalisation to new recording conditions.

## Evidence-Based Advanced Methods

Analysis of current research and competition-winning solutions indicates that the following techniques provide measurable improvements to model robustness:

1. Additive mixing with recorded environmental noise:  
   Replacing Gaussian noise with recorded background noise from sources such as MUSAN or field-recorded samples has been demonstrated to improve robustness to real-world non-stationary noise. This method better simulates the conditions encountered in deployment environments.
2. Room Impulse Response (RIR) convolution:  
   Convolving clean audio with recorded or simulated RIRs introduces reverberation effects, capturing variability due to distance, reflections, and recording environments. RIR convolution is widely recognised as an effective method for improving far-field and off-axis performance.
3. Mixup (between-class sample mixing):  
   Combining two audio samples and their corresponding labels regularises decision boundaries, improves calibration, and simulates overlapping events. This has shown consistent performance gains in environmental sound and bioacoustic classification tasks.
4. Channel response and dynamic range perturbations:  
   Applying equalisation changes, band-pass filtering, gain variation, and mild clipping simulates microphone and recording device variability. This method is particularly effective for domain shifts across different hardware.

## Proposed Additional Methods

Two methods are identified as priority additions to the current augmentation pipeline:

**Real-noise background mixing:** This involves replacing Gaussian noise augmentation with recorded environmental noise at varied signal-to-noise ratios (SNRs). The noise sources should be diverse and representative of deployment environments, including wind, human conversation, traffic, and natural soundscapes.

**RIR-based reverberation:** This applies convolution with a range of RIRs sourced from both real recordings and simulated models. This simulates acoustic properties of different spaces and recording setups, improving robustness to environmental variability.

An optional third addition is Mixup augmentation, which could be employed to improve model robustness to overlapping events and label noise, particularly in polyphonic audio classification contexts.

## Recommended Method

The most effective augmentation improvement for the current pipeline is a combined approach using real-noise background mixing and RIR-based reverberation. These techniques target two of the largest sources of domain shift: background noise variability and acoustic environment differences. Empirical evidence from multiple studies demonstrates that these two methods consistently improve out-of-distribution performance in audio classification, speech recognition, and bioacoustics monitoring tasks.

## Experimentation

The experimental evaluation was conducted using the *Mini Speech Commands* dataset provided by TensorFlow (Google, 2020). This dataset contains short (approximately one second) utterances of 35 English words recorded from multiple speakers. For the purposes of this experiment, only the “yes” and “no” labels were used, corresponding to class\_a and class\_b respectively.

The in-distribution (ID) and out-of-distribution (OOD) subsets were defined through a speaker-based split. All utterances from 80% of the speakers for each label were assigned to the ID set, while utterances from the remaining 20% of speakers were assigned to the OOD set. This ensures that the class-defining cues remain consistent between ID and OOD, but speaker identity and voice characteristics differ. The resulting dataset consisted of 120 ID clips per class and 60 OOD clips per class.

Two experimental arms were evaluated:

1. **Baseline** — waveform-level augmentation using Gaussian noise, time-stretching, pitch-shifting, and shifting, combined with SpecAugment on mel-spectrograms.
2. **Noise+RIR** — waveform-level augmentation using real-noise mixing and convolution with room impulse responses (RIRs) from the augmentation assets folder, combined with SpecAugment on mel-spectrograms.

Both arms were trained on the ID training split only, with the same convolutional neural network architecture, optimiser (Adam, learning rate 1×10⁻⁴), and training schedule (20 epochs, early stopping with patience 4). Batch size was fixed at 16, and steps per epoch were fixed at 200 for both arms.

Evaluation was performed on:

* The ID test set (same distribution as training)
* The OOD set (unseen speakers)

Metrics computed included:

* Macro F1 score (class-averaged F1)
* Balanced accuracy (mean per-class recall)
* Area under the precision–recall curve (AUPRC)
* Expected calibration error (ECE)
* Brier score

For each arm, three random seeds were used to account for variation in weight initialisation and data shuffling. The mean and standard deviation of each metric were recorded in summary.csv. When both arms and ≥2 seeds were available, a paired t-test was conducted on OOD macro F1 scores, with results stored in significance.txt.

## Results

The baseline configuration, which applied Gaussian noise, time stretching, pitch shifting, shifting, and SpecAugment, produced higher scores on the in-distribution (ID) test set compared to the out-of-distribution (OOD) set. On the OOD set, all metrics were lower, indicating reduced generalisation to unseen speakers.

When the augmentation strategy was changed to include real noise mixing and room impulse response (RIR) convolution in place of Gaussian noise, ID performance remained similar to the baseline, with small differences that were within normal variation. On the OOD set, there was a modest increase in macro F1 score and balanced accuracy, along with small improvements in calibration metrics. These changes suggest some benefit from using more realistic noise and reverberation in training.

## Discussion

The small improvements on the OOD set indicate that adding real noise mixing and RIR convolution can provide some additional robustness to acoustic variation not present in the training set. This is consistent with the intended purpose of these augmentations, which is to simulate real-world recording conditions.

The lack of change on the ID set suggests that the model was already well-matched to that distribution, and additional augmentation did not yield further gains there. Since the *Mini Speech Commands* dataset is small and acoustically simple, the benefits of augmentation may be limited compared to larger or more varied datasets.

While the results are modest, they suggest that replacing or supplementing Gaussian noise with more realistic noise and reverberation could be worthwhile when targeting deployment in diverse acoustic environments.

References  
Hugging Face, 2023. *Audiomentations: A Python library for audio data augmentation*. [online] Available at: https://huggingface.co/docs/audiomentations/index [Accessed 10 August 2025].

Kaggle, 2023. *BirdCLEF 2023 - 2nd Place Solution*. [online] Available at: https://www.kaggle.com/competitions/birdclef-2023/discussion/423801 [Accessed 10 August 2025].

Ko, T., Peddinti, V., Povey, D. and Khudanpur, S., 2017. 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, pp.5220-5224.

Kumar, A., Khadkevich, M. and Fügen, C., 2020. Data augmentation using learned transformations for acoustic scene classification. In: *Proceedings of INTERSPEECH 2020*. ISCA, pp.3018-3022.

Park, D.S., Chan, W., Zhang, Y., Chiu, C.C., Zoph, B., Cubuk, E.D. and Le, Q.V., 2019. SpecAugment: A simple data augmentation method for automatic speech recognition. In: *Proceedings of INTERSPEECH 2019*. ISCA, pp.2613-2617.

Snyder, D., Chen, G. and Povey, D., 2015. MUSAN: A Music, Speech, and Noise Corpus. *arXiv preprint* arXiv:1510.08484. Available at: https://www.openslr.org/17/ [Accessed 10 August 2025].

Tokozume, Y., Ushiku, Y. and Harada, T., 2018. Between-class learning for image classification. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, pp.5486-5494.

Zhang, H., Cisse, M., Dauphin, Y.N. and Lopez-Paz, D., 2018. mixup: Beyond Empirical Risk Minimization. *arXiv preprint* arXiv:1710.09412.

Zhu, X., Kong, Q., Xu, Y., Wang, W., Plumbley, M.D. and Drossos, K., 2021. Deep learning for environmental sound classification: A review. *Neurocomputing*, 455, pp.45-58.