NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

School of Computer Science and Engineering
College of Engineering

Keyword Spotting

Small Footprint Model under Noisy Far-field Environment

Student: Pang Jin Hui

Supervisor: Assoc Prof Chng Eng Siong

Abstract

Building a small memory footprint keyword spotting model is important as it typically runs on mobile devices with low computational resources. In real life, noisy environment with some reverberations is degrading the performance of a keyword spotting (KWS) model. We proposed a novel feature interactive convolution (ConvMixer) model with small parameters for single-channel and multi-channel utterance. Moreover, we proposed a centroid-based awareness component to improve the multi-channel system by providing some additional spatial geometry information in the latent feature projection space.

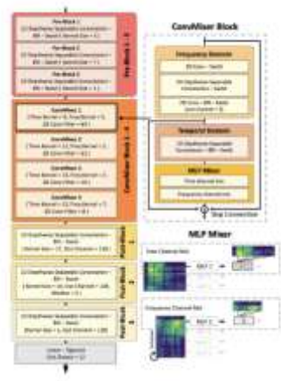
Introduction

Small models face a tough challenge in KWS task in noisy environments.

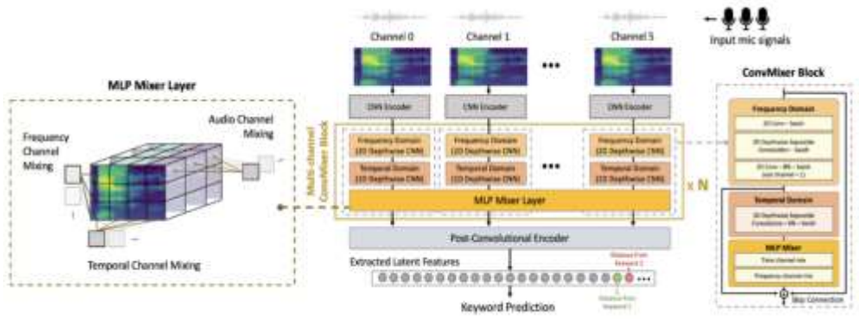
- **ConvMixer** – The model can outperform some novel small neural network models like ResNet and attain robustness in noisy far-field single-channel and multi-channel signals.
- **Centroid-based Awareness** – Use the distance between input and keywords to estimate the centroid vector for better performance.

Approach

- **Single-channel model**



- **Multi-channel model**



Result

Our proposed model achieved the best performance among SOTA small footprint models.

- **Single-channel model**

- **Multi-channel model with centroid-based awareness**

Conclusion

The accuracy of single-channel ConvMixer in Google Speech Command is 98.20% and in an average of four different SNRs in far-field is 76.94%. For multi-channel ConvMixer, it reaches the best accuracy of 94.3% and the lowest score of 0.152 with the use of centroid-based awareness in MISP dataset.