NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

School of Computer Science and Engineering
College of Engineering

Sound Event Detection

with Human and Emergency Sounds

Student: Lee Yan Zhen

Supervisor: Assoc Prof Chng Eng Siong

Background

Sound Event Detection (SED) is the task of recognizing sound events and their respective onset and offset timestamps in audio clip recordings. It has useful implementations in smart homes and autonomous vehicles, such as detecting ambulance sirens and prompting vehicles to adjust their routes to give way accordingly.

Problem Statement

Existing SED systems analyse audio clips as a whole and have yet to fully explore the impacts of applying rolling segmentation windows on the audio clips prior to analysis. This is despite the fact that doing so may yield potential benefits, especially in real-life applications, where audio clips could be hours long and processing them as a whole is not feasible. However, just analysing non-overlapping audio segments may not yield the best outcomes. Hence, there is a need to develop a method to amalgamate the frame-wise predictions post-analysis.

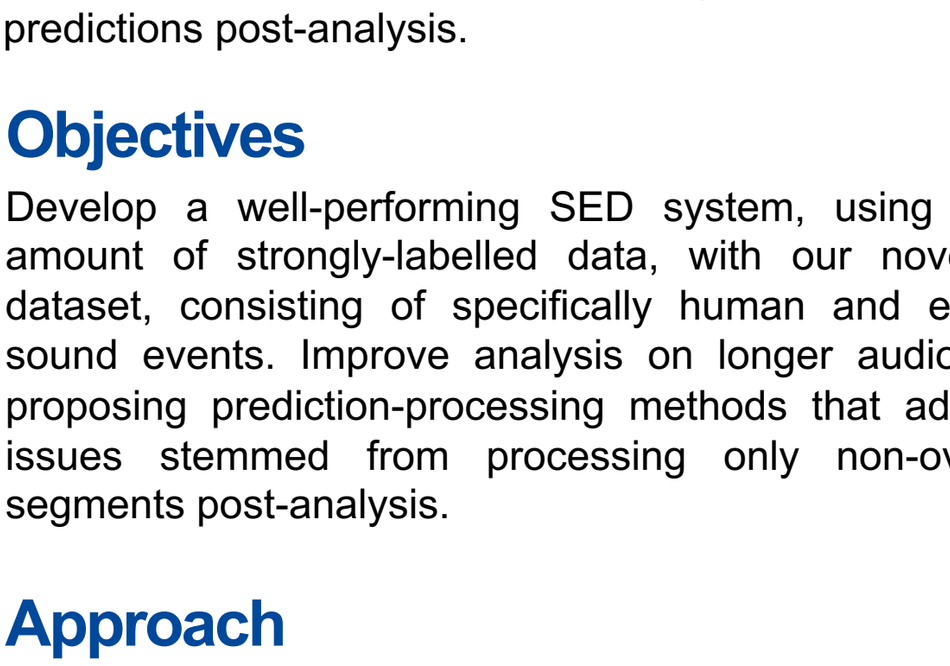
Objectives

Develop a well-performing SED system, using a limited amount of strongly-labelled data, with our novel project dataset, consisting of specifically human and emergency sound events. Improve analysis on longer audio clips by proposing prediction-processing methods that address the issues stemmed from processing only non-overlapping segments post-analysis.

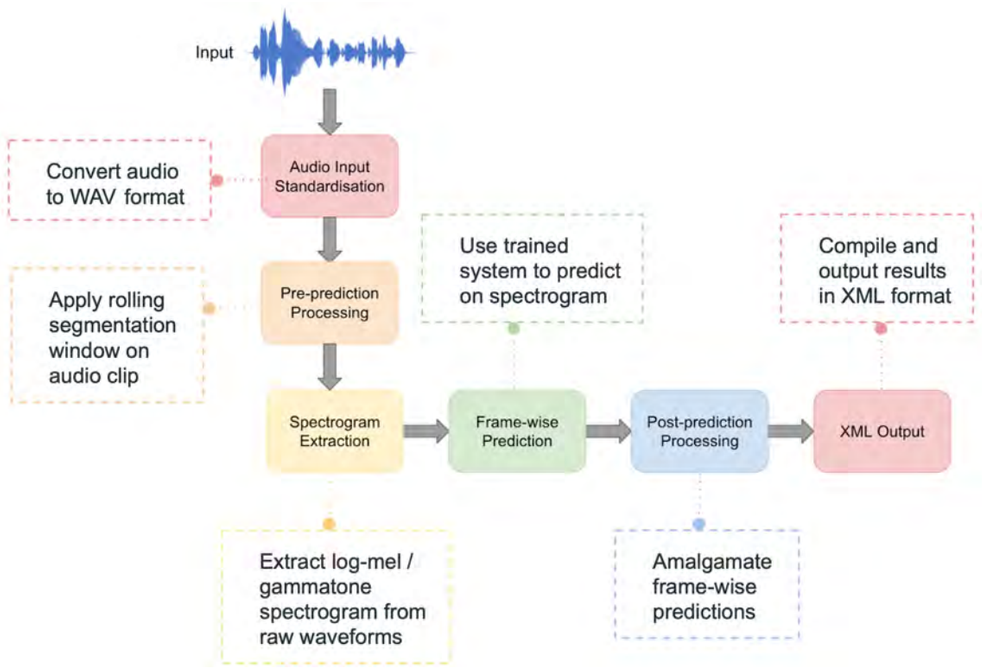
Approach

There are two main components of the SED system, which are namely the training system and prediction system. The training system is the foundation for the development of our SED system while the prediction system implements the trained system and analyses audio input that is independent of the audio clips in the development set.

Training System



Prediction System



Our proposed post-prediction processing methods include frame-wise averaging and using a voting scheme. The former sums the frame-wise predictions together and then averages them based on the number of overlaps present in a particular segment. The latter sums binarized frame-wise predictions and thresholds based on the number of overlaps present.

Results

Our proposed frame-wise averaging method improved the performance of the SED system consistently, proving especially useful for longer audio clips.

Model	Processing Type	Threshold Type	Error Rate	F1-Score	Precision	Recall	Process Time (Seconds)
CNN-9 -GRU -Attention	None	Non-optimised	0.555	0.624	0.718	0.552	25.124
		Optimised	0.601	0.635	0.619	0.651	
	Frame-wise averaging	Non-optimised	0.557	0.618	**0.749**	0.525	35.931
		Optimised	0.574	0.639	0.645	0.634	
	Voting	Non-optimised	0.658	0.572	0.624	0.528	43.302
		Optimised	0.711	0.560	0.574	0.547	
CNN-9 -Transformer -Attention	None	Non-optimised	0.552	0.628	0.712	0.562	25.947
		Optimised	0.561	0.648	0.650	0.646	
	Frame-wise averaging	Non-optimised	**0.549**	0.622	0.747	0.533	33.682
		Optimised	0.550	**0.648**	0.667	0.628	
	Voting	Non-optimised	0.562	0.643	0.663	0.623	38.923
		Optimised	0.675	0.628	0.569	**0.700**	
CNN-9 -Conformer -Attention	None	Non-optimised	0.560	0.620	0.666	0.580	18.227
		Optimised	0.591	0.629	0.617	0.642	
	Frame-wise averaging	Non-optimised	0.560	0.611	0.699	0.611	24.645
		Optimised	0.563	0.635	0.646	0.624	
	Voting	Non-optimised	0.662	0.562	0.615	0.518	31.836
		Optimised	0.683	0.557	0.600	0.520	

Analysis Results on Longer Audio Clips