SPEECH ENHANCEMENT - EXPLORING OPEN UNMIX

A PREPRINT

Markús Freyr Sigurbjörnsson
Master in Sound and Music Computing
Pompeu Fabra University
markus.sigurbjornsson01@estudiant.upf.edu

March 29, 2021

***K*eywords** Speech Enhancement · Open Unmix

1 Introduction

Speech is the most sophisticated signal naturally produced by humans. It is the most fundamental way of sharing information. It allows people to express emotions and verbally share feelings. The aim of digital speech processing is to take advantage of digital computing techniques to process the speech signal for increased understanding, improved communication, and increased efficiency and productivity associated with speech activities. The presence of background noise in speech is problematic both for humans and computers alike. The problem of dealing with it is called speech enhancement or noise reduction. The aim of a speech enhancement system is to suppress the noise in a noisy speech signal. I will take a look at a source separation algorithm called Open Unmix and train it on two datasets, one for speech signal and the other one for background noise or interference.

2 Datasets

The data set for speech signals is called The Emotional Voices Database. I chose this dataset as it consists of recordings of four speakers, two male and two female, and each recording is only about 5-6 seconds long. This dataset was built for the purpose of emotional speech synthesis. The emotional styles are neutral, sleepiness, anger, disgust and amused. For the background noises I used the dataset ESC-50, which is a dataset that consists of a collection of 2000 environmental audio recordings suitable for benchmarking methods of environmental sound classification. Each recording is 5 seconds long and is organized into 50 semantic classes. There are five main classes; Animals, Natural soundscapes water sounds, Human, non-speech sounds, Interior/domestic sounds and Exterior/urban noises.

3 Methodology

Prepossessing the data was fairly simple. First of all I go through the audio recordings of speech signal and discard the files which are under five seconds and for the rest of the files I trim to be exactly five seconds. I do this because I want all of the audio signal to be the same length which will be useful when mixing the signals and testing my model. I split up both the dataset into training and validation sets, and for the ESC-50 I took each category and split that up accordingly so there would be same distribution for each environmental sound in testing and validation. I fed the Open Unmix algorithm time series data and the spectrogram is computed internally. The audio is then downsampled. There is a normalisation step where, for each frequency bin within the spectrogram will be normalised so they will have a mean of 0 and standard deviation of 1. Next is a compression step and there is a hyperbolic tangent because the data is zero meaned and at this point we have learned relevant information for features to go into the bidirectional lstm. Than the data is fed through standard fully connected layers and finally passes through a relu activation function to achieve the ratio mask. Now all what is left is to multiply the ratio mask to the original signal to recover the target audio.

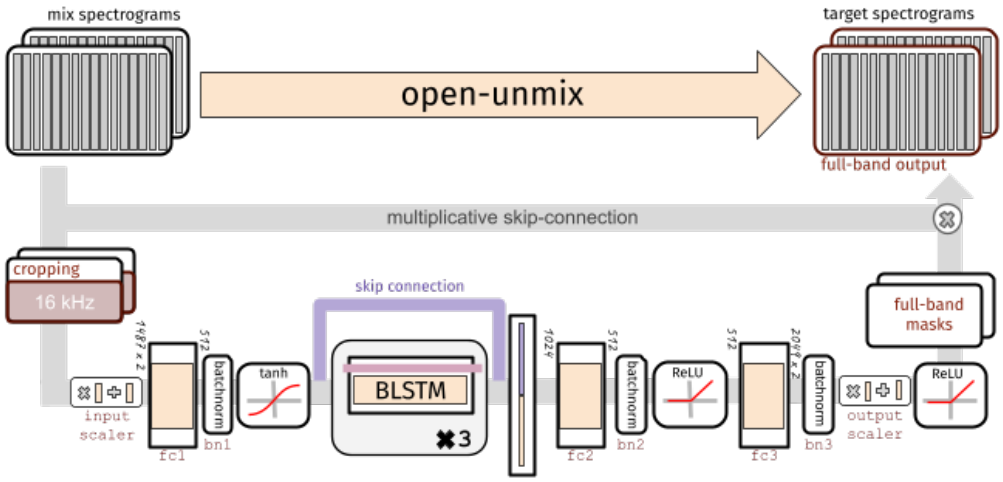

Figure 1: Architecture for Open Unmix

4 results

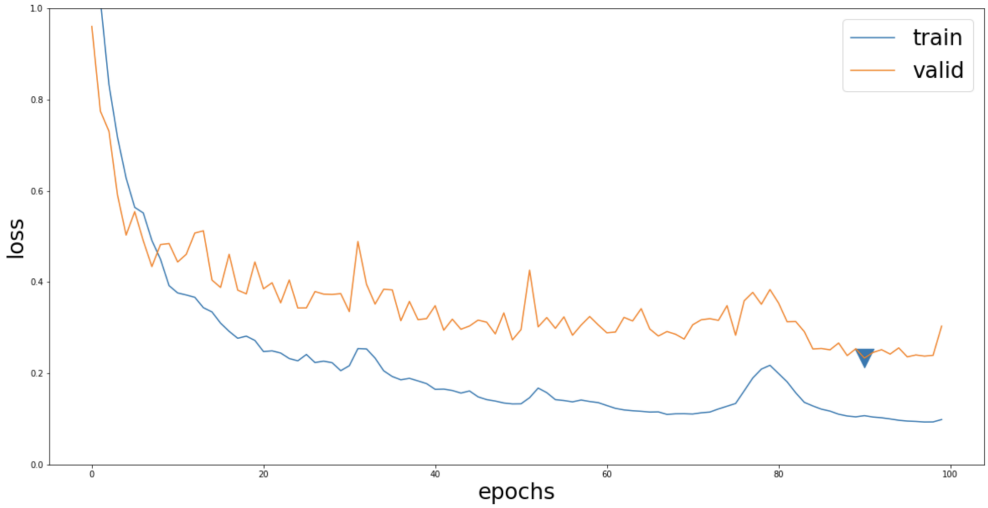

Figure 2: MSE loss for train and validation sets

Training the model took approximately 20 hours and forty minutes. After testing out the model and applying some separations to mixture of signals it was very clear that the model was not very good. The target signal was barely recognisable or not recognisable at all. evaluating the model using my own subjective metrics I give it a very bad score. When evaluating the model I wanted to use the built-in evaluation method in Open Unmix for evaluating SDR and SNR. But the evaluation function needed to take in a musicdb object and I could not figure out how to use my local data. I figured I'd use mir_eval library for evaluation. mir_eval has a function for evaluating source separation. I applied the reference source and estimated source into a evaluation function in mir_eval for five different samples and got back terrible numbers ofr SDR.

-37.95344581 -25.21580209 -39.463604 -42.94104204 -35.01951836

5 Discussion

There have probably been a lot of things I could have done better. I think maybe something could have happened to the speech signals when I was re-sampling them to 44100Hz. Even though this project did not give me very good results I still learned alot.

References

[1] Fabian-Robert Stöter, Stefan Uhlich, Antoine Liutkus, and Yuki Mitsufuji. Open-unmix - a reference implementation for music source separation. *Journal of Open Source Software*, 4:1667, 09 2019.