VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)
CHENNAI

# SPEECH SEPARATION USING U-NETS

NIKHIL CHAPRE DR. MERCY RAJASELVI BEAULAH P

## OBJECTIVES

The primary objective of using U-Nets for speech separation is to improve the separation performance of the model. This can involve experimenting with different network architectures, training procedures, and loss functions to achieve better separation performance. Also the need is to focus on techniques which helps in tackling problems like room reverberation, de-synchronization during pre-processing.

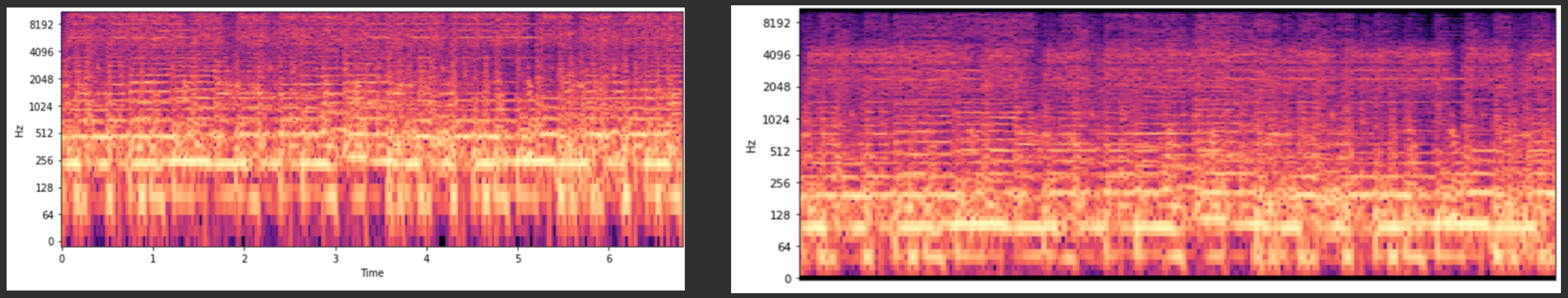
## INTRODUCTION

Speech Separation is one of the most fundamental tasks in the domain of signal processing. This problem revolves around the idea of an acoustic environment where multiple sounds are present and the task is to separate the target speech signal from the surrounding noise/interference or non-relevant speakers. The proposed method utilizes a supervised learning approach and is trained on a large dataset of mixed speech signals, where the model is capable of separating speech signals even with different levels of complexity of external factors such as noise and room reverberation

## METHODOLOGY

First, we start with building a vocal separation model, where our goal is to isolate the singing voice from a mixed music track containing multiple instruments and background noise.
We further extend this baseline model to solve the mixed speaker separation problem which involves separating multiple speech signals from a mixed audio signal, where each speech signal belongs to a different speaker.
Lastly, we will use evaluation metrics such as SDR, SIR and SAR to record and tabulate the performance of our model and fine-tune it to get better results.

## RESULTS: SPECTROGRAM ANALYSIS



On the left side we have our target/actual spectrogram for vocals of a sample song that we used for testing and on the right we have the output/predicted spectrogram from our model after fine-tuning for the vocal component using only the mixture. These spectrograms can then be converted again to audio files to demonstrate the difference in quality.

## CONCLUSION

Overall, this thesis shows how we addressed the problem of Mixed Speaker Separation by using Contemporary methods and filled the research gap between existing work already present on this topic. Also using introducing random noise and reverberation into the mixture can act as a Data Augmentation technique as it leads to unique solutions for masks.
We also discussed how binary separation or two speaker mixtures work better than multiple or unknown number of speakers for our model. This is due to the fact that the training data was limited in that scope.

## REFERENCES

[1]. Wang, D., & Chen, J. (2018). Supervised speech separation based on deep learning: An overview. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26(10), 1702-1726.
[2]. Gao, R., & Grauman, K. (2021, June). Visualvoice: Audio-visual speech separation with cross-modal consistency. In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 15490-15500). IEEE.

## FUTURE WORK

Using Dedicated U-Net models for each component of a song/speaker can help achieve greater accuracy on multiple speaker problem.
Using both Audio and Video embeddings we can provide additional cues for easier speech separation and separate speaker in video debates as an example.

## CONTACT INFORMATION

NAME: NIKHIL CHAPRE
REGISTRATION NUMBER: 19BCE1315
EMAIL ID:
nikhil.chandrashekhar2019@vitstudent.ac.in
MOBILE NO: 9693382527