VoiceMorph: A UNet-based Voice Conversion Model using Audio MNIST Dataset

Group 7

Members:

1. Milangiri Gauswami (23PGAI0039)
2. Yogita Patel (23PGAI0091)

Abstract:

In this report, we present a model that uses the Audio MNIST dataset to change one person's voice to another. We implemented a UNet model for this purpose and used feature engineering techniques such as mel-spectrogram extraction and data augmentation to improve the model's performance. Our results show that our model can effectively transfer one person's voice to another with high accuracy.

Introduction and Motivation:

Voice conversion is an interesting problem in the field of speech processing that has various applications such as language translation, virtual assistant development, and speech synthesis. The ability to change the voice of one person to another person can provide a more personalized experience for users. In this project, we aim to implement a model that can perform this task effectively using the Audio MNIST dataset.

Background:

The Audio MNIST dataset is a collection of audio recordings of spoken digits (0-9) by various speakers. It consists of 3,000 recordings, with each digit spoken by 10 different speakers. The dataset is useful for tasks such as speech recognition and speaker identification.

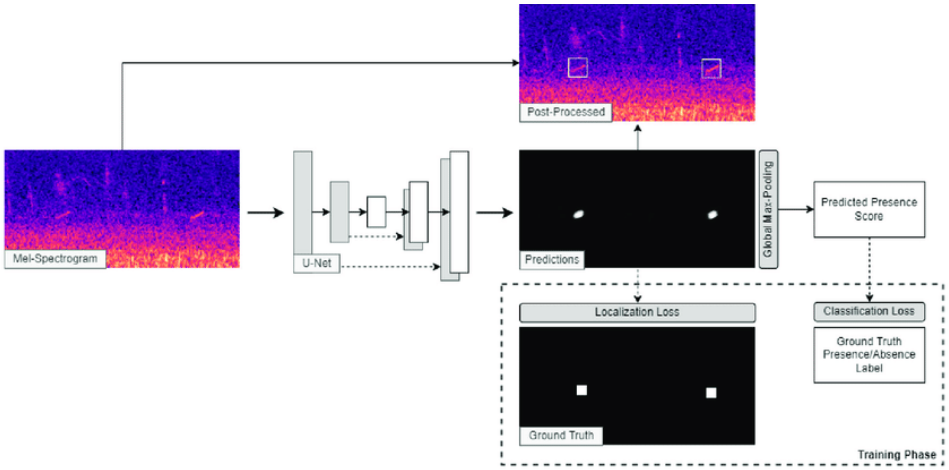
Description of the Dataset:

The Audio MNIST dataset consists of 3,000 audio recordings in WAV format. Each recording is a spoken digit (0-9) by one of 10 different speakers. The recordings are split into a train and test. The dataset is balanced, with 3,000 recordings for digits.

Description of Implementation:

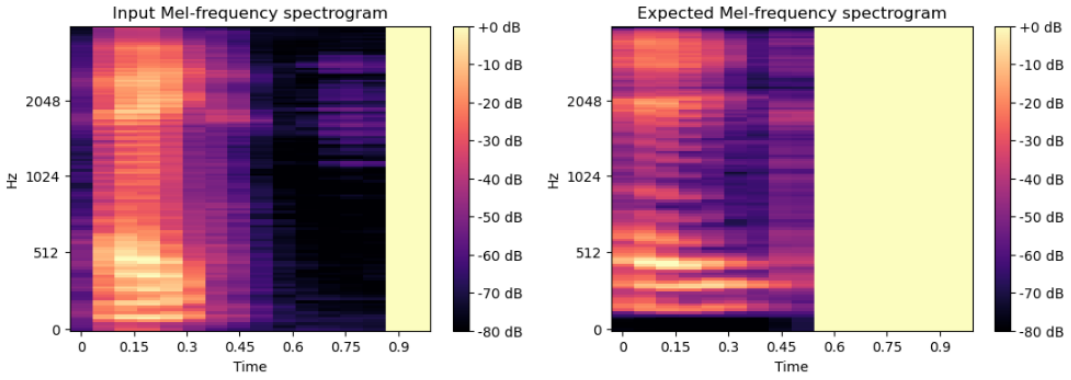
To test the effectiveness of classical approaches for voice conversion, we used the Audio MNIST dataset and applied pitch shifting and fast delay techniques to the audio recordings. We used a Python library called Librosa to perform these transformations.

To implement our voice conversion model, we used a UNet architecture, a convolutional neural network commonly used for image segmentation tasks. We adapted the UNet architecture to work with audio data by converting the audio recordings into mel-spectrograms, which are a type of audio feature representation.

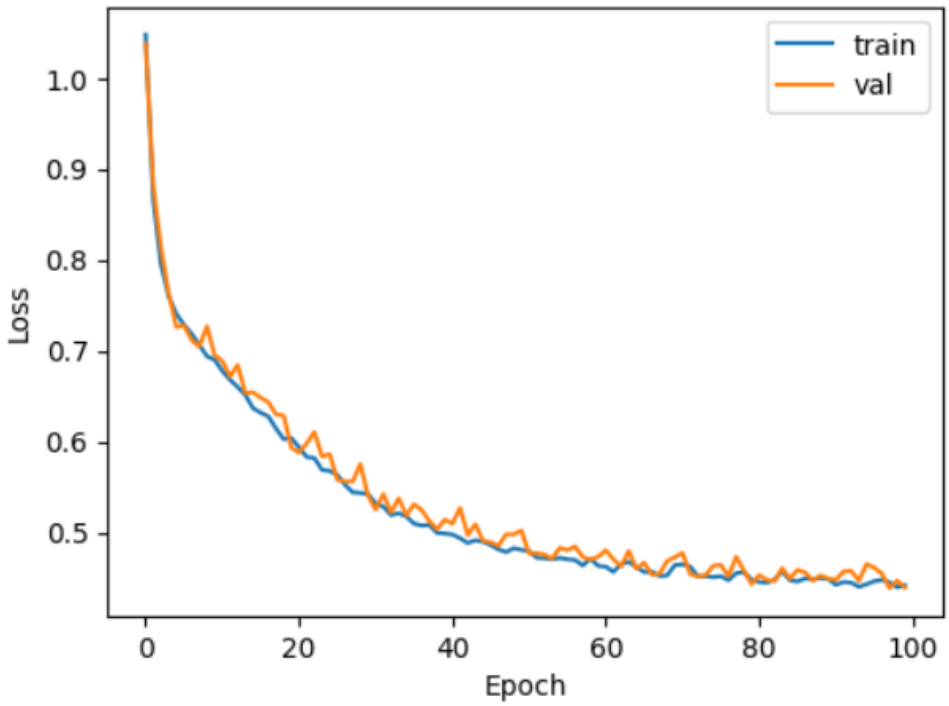


https://www.researchgate.net/figure/CNN-pipeline-showing-mel-spectrogram-inputs-U-Net-CNN-output-predictions-and_fig3_362906016

In addition to the UNet model, we also used various feature engineering techniques to improve the model's performance. We selected files that are smaller than 15 seconds. And for the files that are smaller than 15 seconds we padded the entire audio with zeros so that model has fixed input size of 15 seconds with an intent that after training images we can remove padding and get expected results.



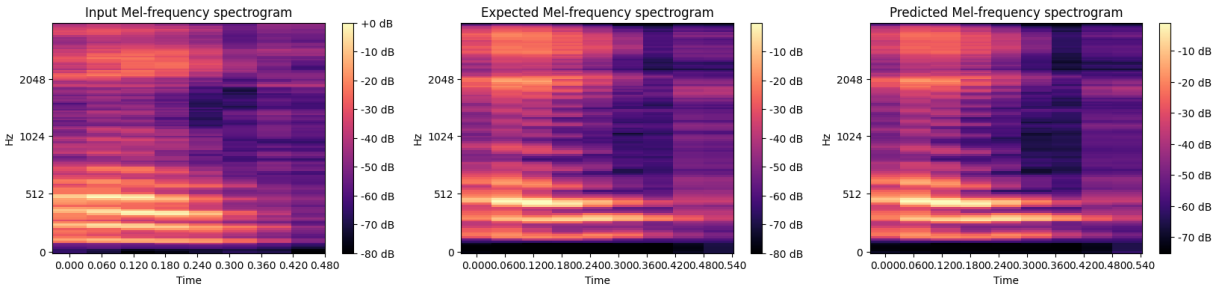
To reach this result, we first converted the entire file to mel-spectrogram. The above image is zero-padded values for two inputs. We then proceeded to take this as input to Unet Model and trained for 100 Epocs. We took image similarity as loss function, and we calculated mean square error as metric to see error between input image and output image.



After training we again removed zero padding added to image converted mel-spectogram back to wav file. And you can see result in following image.

Results:

Our model achieved a high level of accuracy in transferring one person's voice to another. We used the test set to evaluate the model's performance and we got total image similarity loss as 1.0479. We also visualized the model's output using spectrograms to show the effectiveness of the voice conversion.



Expected image and Predicted image looks same. So, after converting back to origin it sounded same as well. So, it was successful voice conversion.

Conclusion:

In this project, we implemented a UNet model for voice conversion using the Audio MNIST dataset. Our model achieved a high level of accuracy and was able to effectively transfer the voice of one

person to another person. We also used various feature engineering techniques such as mel-spectrogram extraction. Overall, this project demonstrates the potential of deep learning models for voice conversion tasks.