Speaker Recognition using Neural Network

Course Project - DA 623

Presented by
Posa Mokshith (210101077)

Department of Computer Science and Engineering
IIT Guwahati

8th May 2025

Speaker identification task

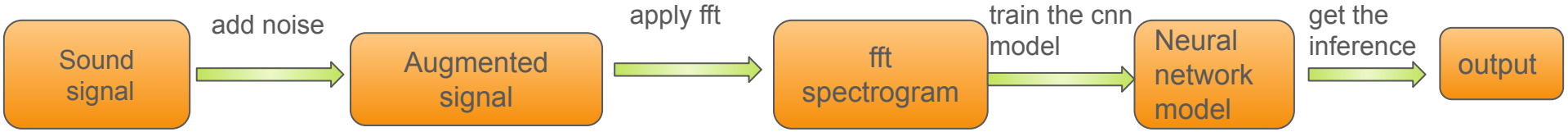
What is Speaker Identification?

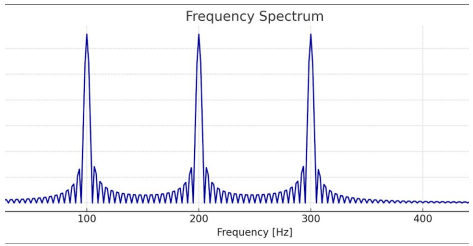
- Speaker identification is the process of determining which registered speaker is speaking from a set of known speakers, based on their voice characteristics.
- The system labels spoken utterances with the correct speaker ID from a predefined group

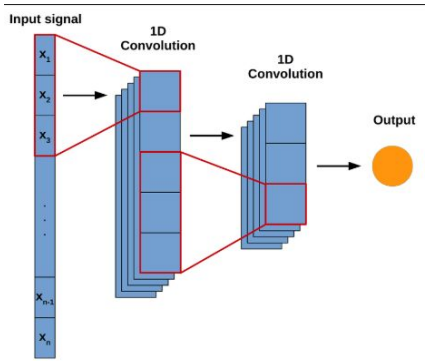
Key Challenges

- Variability in speech due to emotion, health, background noise, and recording devices.
- Differences in speaking style: fixed phrases vs. spontaneous speech

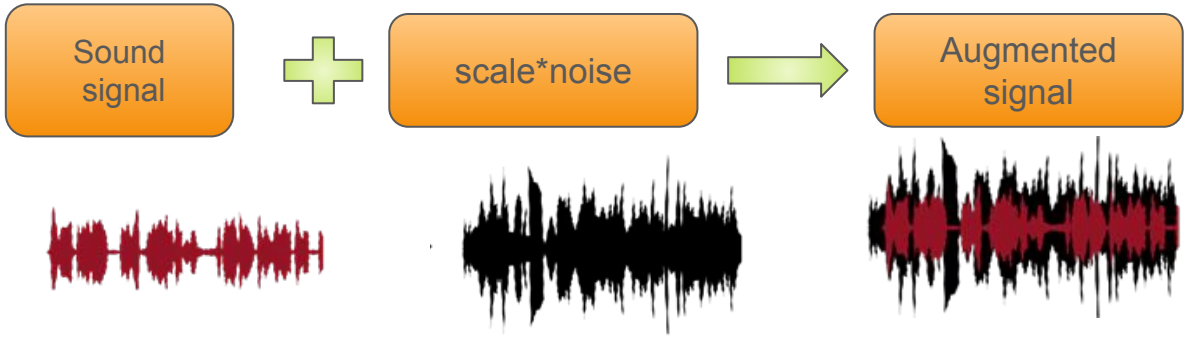
Implementation overview







Data augmentation



Purpose of Data Augmentation

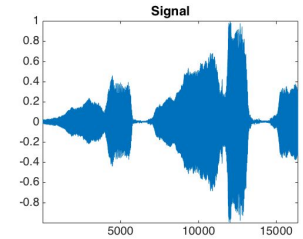
- Increases the diversity and size of training data without collecting new recordings.
- Helps neural networks generalize better and become more robust to real-world variability

Impact on Model Performance

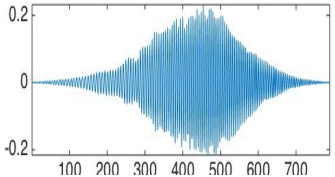
- Augmentation leads to significant improvements in speaker identification accuracy and robustness, especially in noisy or mismatched condition

Apply FFT signal

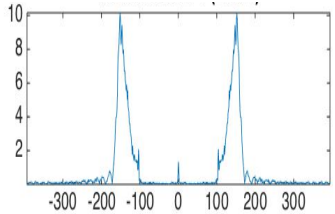
The Fast Fourier Transform (FFT) converts time-domain audio signals into frequency-domain representations, revealing spectral components critical for analyzing vocal characteristics.
This transformation enables Identification of pitch, harmonics, and formants unique to each speaker.



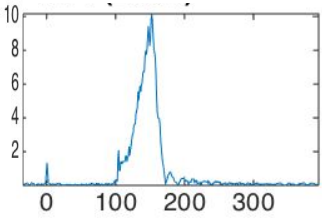
Sound signal



Windowed signal

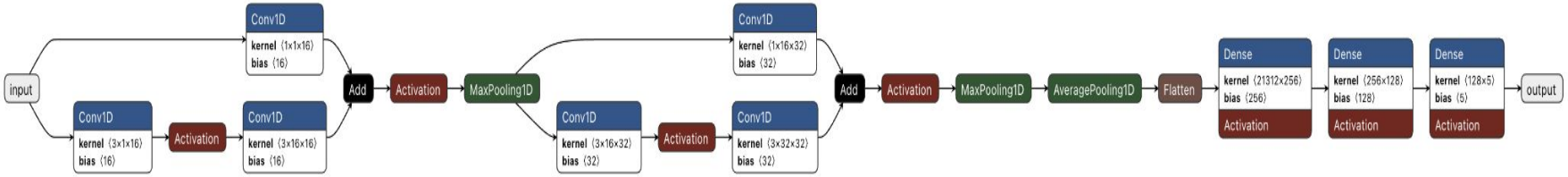


fft spectrogram



Positive frequencies in fft spectogram

CNN architecture



- **FFT Spectrogram Input**
  Converts time-domain audio to frequency domain using FFT, capturing pitch, harmonics, and formants essential for speaker-specific features.
- **CNN with Residual Blocks**
  Two residual blocks improve gradient flow and training stability. Each block includes convolution, batch normalization, ReLU, and identity skip connections.
- **Classification Layer**
  Flattened features pass through dense layers to a softmax classifier for speaker ID. This pipeline handles noise and variability introduced by augmentation.

Flattening & Dense Layers

- Flattened output passed through fully connected (dense) layers.
- Non-linear activations improve model expressiveness.

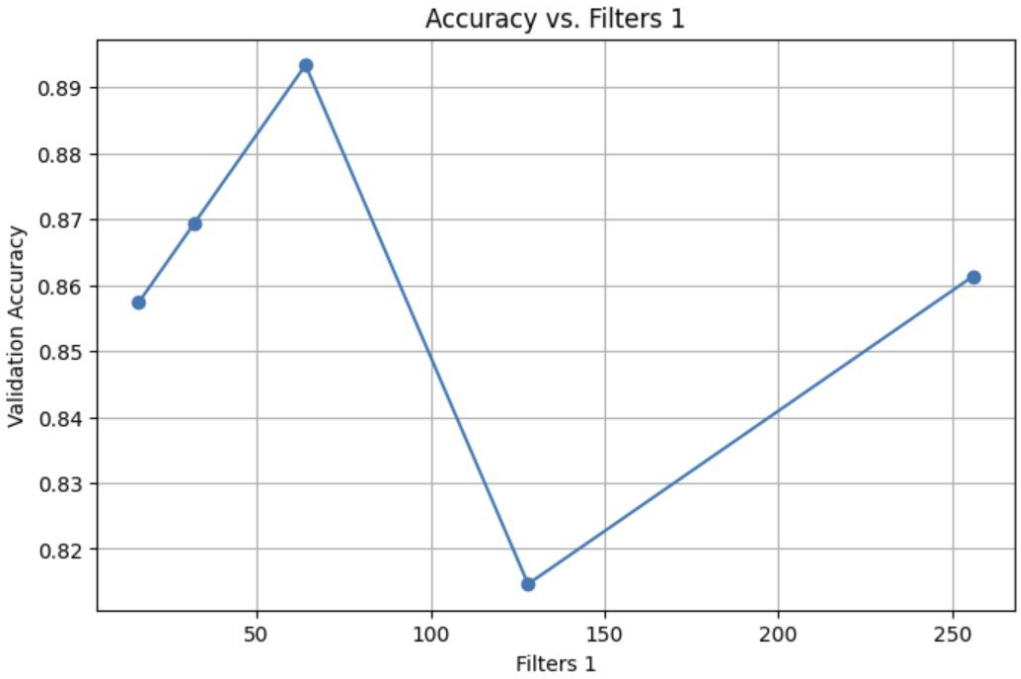
Output Layer

- Softmax activation for multiclass speaker classification.

Interactive Model Training in Kaggle Notebook

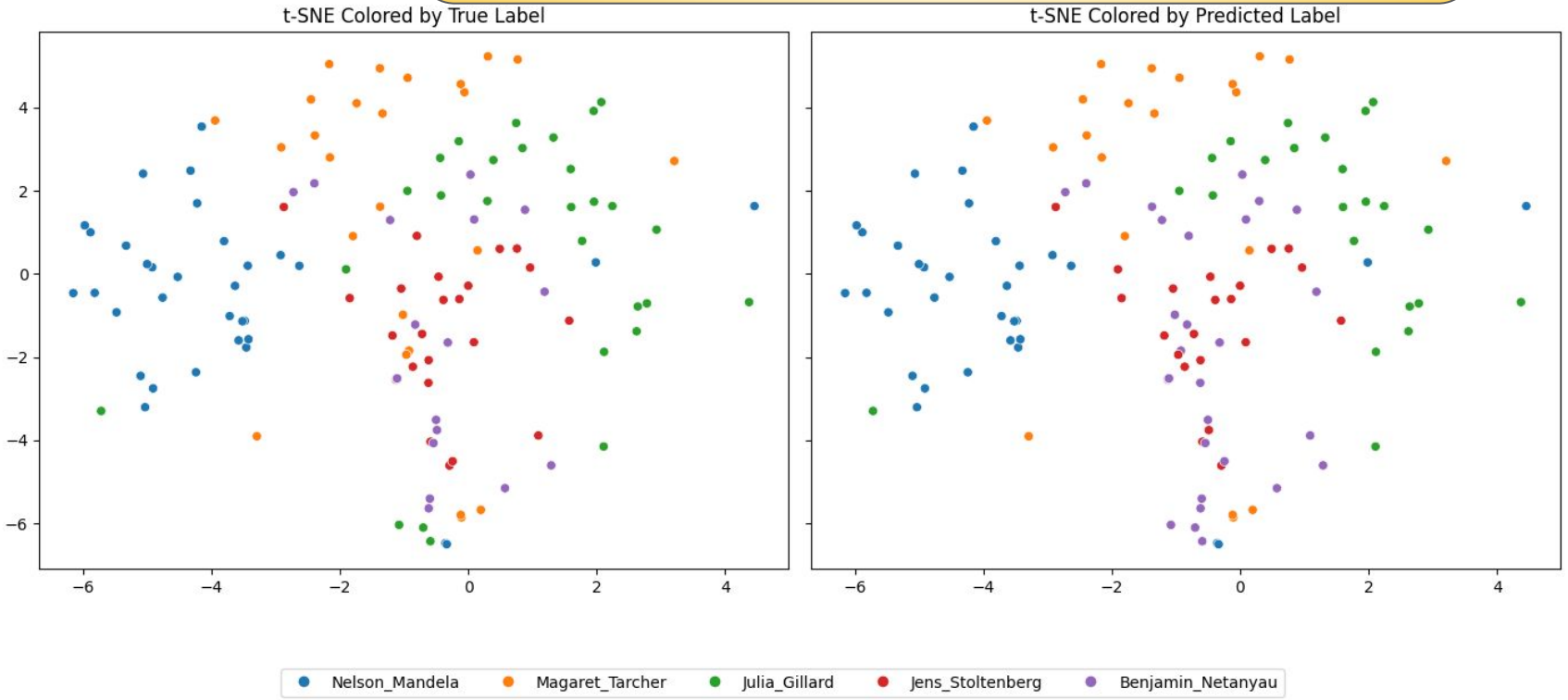
```
🧪 Running sweep: Filters 1
Filters 1 = 16 → Val Accuracy: 0.8573
Filters 1 = 32 → Val Accuracy: 0.8693
Filters 1 = 64 → Val Accuracy: 0.8933
Filters 1 = 128 → Val Accuracy: 0.8147
Filters 1 = 256 → Val Accuracy: 0.8613
```

Blocks:	5
Activatio…	relu
Vary Para…	Filters 1
Filters 1	16
Filters 2	32
Filters 3	64
Filters 4	128
Filters 5	128
Filters 6	128
Conv 1	2
Conv 2	2
Conv 3	3
Conv 4	3
Conv 5	3
Conv 6	3

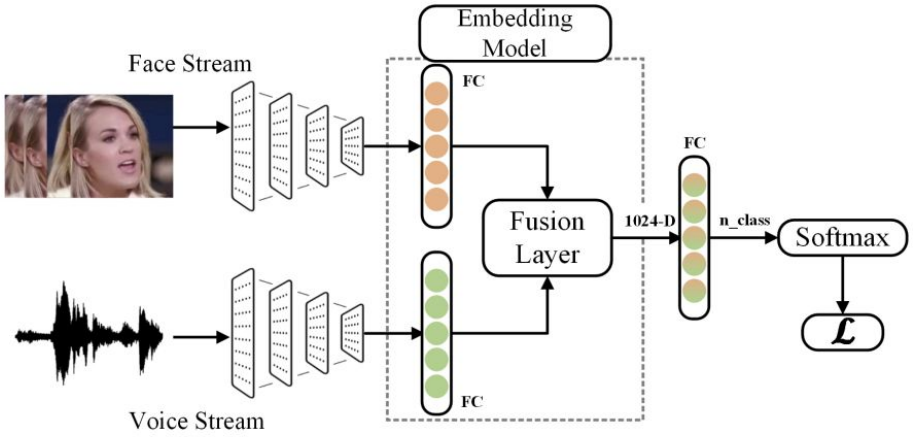


T-SNE plots

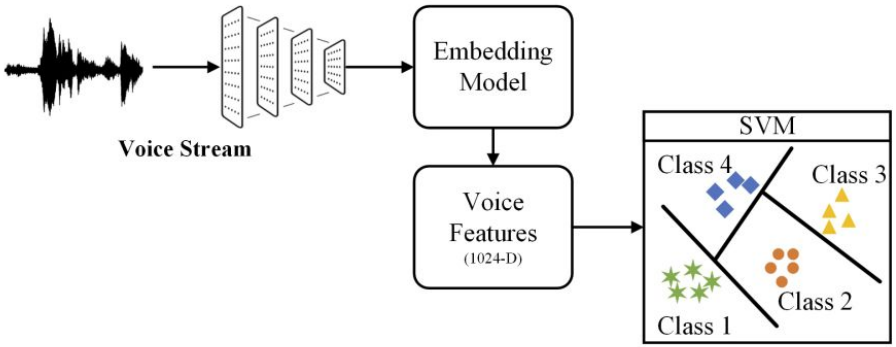
T-SNE plots visualize high-dimensional fft spectrogram in 2D space. They reveal clear clustering patterns, indicating effective speaker separation by the model.



Multimodal learning for speaker identification



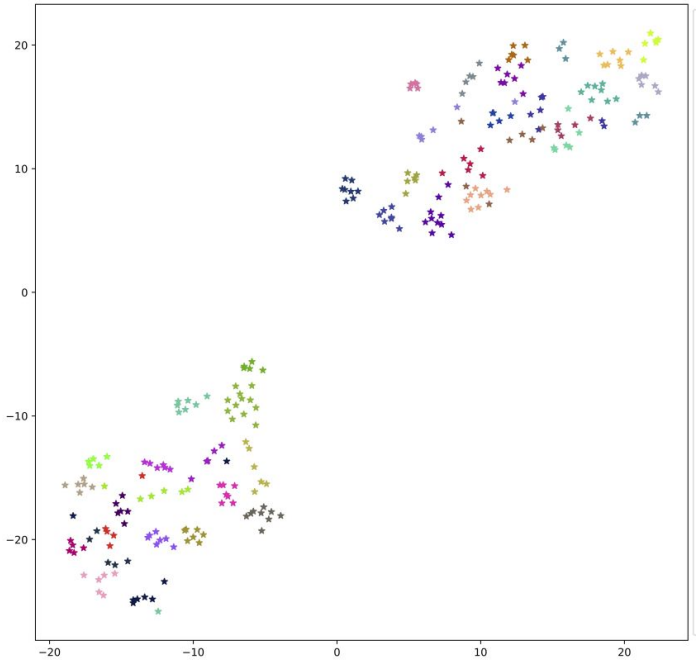
(a) Proposed Two-branch network



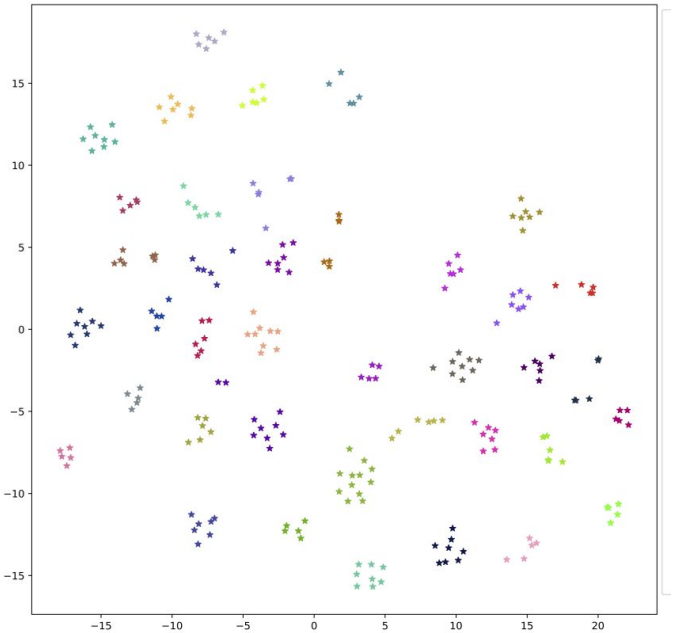
(b) Testing strategy with single modality

T-SNE plots for comparison

features from VGGVOX Network



Features from 2 branched Network



References

Speaker Recognition in Realistic Scenario Using Multimodal Data

Kaggle notebook is Here

Youtube Link is here