Enhancing Forensic Audio Investigations: A Multi-Speaker Speech Understanding and Analysis System - Project Report

Himani (M23CSA516)
m23csa516@iitj.ac.in

Ankit Kumar Chauhan (M23CSA509)
m23csa509@iitj.ac.in

Abstract

The need for effective multi-speaker audio analysis is essential for applications like forensic audio investigations, surveillance, and customer service. In this paper, we present a Multi-Speaker Speech Understanding and Analysis System that combines speech separation, enhancement, emotion classification, gender recognition, and speaker identification into a unified framework. The system uses the VoxCeleb dataset for speaker recognition and the SepFormer model for speech separation and enhancement. We employ Convolutional Neural Networks (CNNs) for emotion classification and Long Short-Term Memory (LSTM) networks for gender prediction. The effectiveness of the system is demonstrated on a custom dataset generated by mixing VoxCeleb2 audio files, highlighting its potential for forensic applications. Our results indicate that the system performs robustly in multi-speaker environments, offering reliable predictions for each task.

1. Introduction

Forensic audio analysis often requires processing complex recordings with overlapping speech from multiple speakers. Traditional systems struggle to accurately separate voices and analyze speech content, especially in real-world scenarios. To address this challenge, we propose an integrated system for Enhancing Forensic Audio Investigations, which performs speech separation, enhancement, emotion recognition, gender classification, and speaker identification. The system is designed to handle multi-speaker environments, leveraging the VoxCeleb dataset for training and evaluation.

We focus on separating overlapping speech using the SepFormer model, enhancing audio clarity, and then performing emotion and gender classification using deep learning models. Additionally, the VoxCeleb dataset is used to identify and verify speakers, making the system well-suited for forensic investigations where speaker identification is crucial.

2. Methodology

Our system integrates speech separation, enhancement, emotion classification, gender classification, and speaker identification into a unified pipeline for forensic audio investigations. The system consists of several modules, each designed to handle specific tasks:

2.1. Dataset Preparation

We used the VoxCeleb2 dataset, a large-scale collection of speech from various speakers, including both male and female voices. Audio from VoxCeleb was mixed [2] to create multi-speaker environments. The mixing process simulates real-world scenarios where multiple speakers are overlapping, a common challenge in forensic audio analysis.

2.2. Speech Separation

The first step in processing the mixed audio is speech separation, performed using the SepFormer model [5]. The model utilizes a waveform-based separation approach and isolates each speaker's audio. The separation process is crucial for accurate emotion and speaker analysis, as overlapping speech can significantly degrade classification accuracy.

2.3. Speech Enhancement

After speech separation, the system applies the SepFormer enhancement model [5] to improve the clarity of the separated audio. This model reduces background noise and enhances the quality of the speech, making it more suitable for emotion and gender classification.

2.4. Feature Extraction

The system extracts MFCC (Mel-Frequency Cepstral Coefficients) [6], Zero Crossing Rate (ZCR), and Root Mean Square Error (RMSE) features from the audio. These features are crucial for emotion and gender classification. The extracted features are normalized using a pre-trained StandardScaler.

2.5. Gender Classification

We used an LSTM-based model for gender classification, trained on MFCC features extracted from the audio. The model predicts whether the speaker is male or female.

The Gender Classification Dataset[1] contains .wav audio files organized into two folders: Female and Male. The Female folder consists of 5,768 audio files, while the Male folder contains 10,400 audio files. The audio clips range from 1.5 to 5 seconds, with an average length of 3.26 seconds. The dataset was obtained online and is intended for building machine learning models that classify gender based on voice. Approximately 1,000 files are duplicates, which need to be handled during preprocessing.

2.6. Emotion Classification

The emotion classification model uses a CNN architecture [7]. The model is trained to predict one of the following emotions: Neutral, Calm, Happy, Sad, Angry, Fear, Disgust, Surprise. The model uses convolutional layers for feature extraction, followed by fully connected layers for classification.

For model training, we used a combination of four widely-used speech emotion recognition datasets[4]: Ravdess, Crema, Savee, and Tess. These datasets contain audio recordings(approx 12k) of speakers expressing a variety of emotions such as happy, sad, angry, fearful, surprised, disgusted, and neutral. The audio files are in .wav format, and we extracted features like MFCC, ZCR, and RMSE to train our emotion recognition model. This diverse collection of labeled data allowed us to build a robust model capable of accurately classifying emotional states from speech.

2.7. Speaker Identification

For speaker identification, the system (Figure 1) uses the VoxCeleb dataset2 and the ECAPA-TDNN model [3]. The model compares the features extracted from the test audio against a reference speaker database and provides a matching score for speaker identification.

3. Experiments and Results

3.1. Model Training for Gender Classification

3.1.1 Model Architecture

- The gender classification model uses an LSTM-based architecture, which is well-suited for sequence data like audio. The model consists of the following layers:
- LSTM Layer: The LSTM layer processes the input sequence (MFCC features) and produces a hidden state.
- Fully Connected Layers: After the LSTM, the model includes two fully connected layers. The first layer maps the LSTM output to 1024 units with ReLU activation, followed by the second layer, which outputs the final prediction using a Softmax activation to classify the input as either Male or Female.
- Dropout: Dropout layers are applied to reduce overfitting.

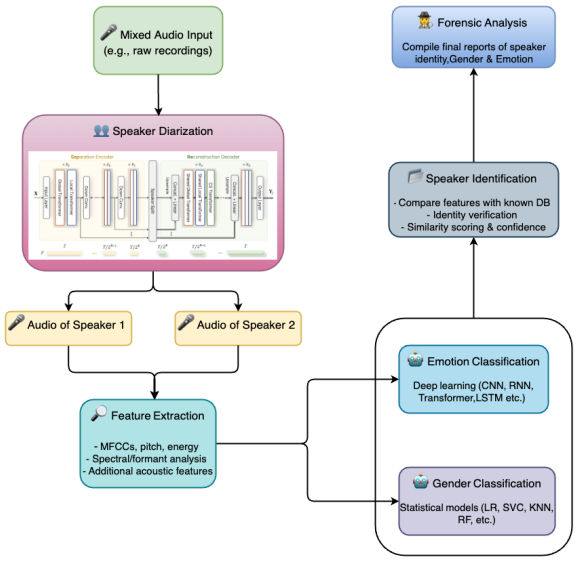

Figure 1. System Architecture

3.1.2 Hyperparameters

- Input Size: The input to the model is the MFCC feature vector, with 20 coefficients extracted from the audio.
- Hidden Size: The LSTM hidden state has a size of 1024 units.
- Num Layers: The LSTM has 2 stacked layers to capture complex sequential dependencies.
- Num Classes: The output layer has 2 units (representing Male and Female).
- Optimizer: Adam optimizer is used to minimize the loss function.
- Loss Function: Binary Cross-Entropy Loss (BCELoss) is used because the task is binary classification (Male vs. Female).
- Learning Rate: 0.001.
- Batch Size: 32.
- Epochs: The model is trained for 5 epochs.

The graph (Figure 2) shows the training, validation, and test loss over 5 epochs:

1. Training Loss decreases consistently, indicating the model is learning and improving over time.
2. Validation Loss initially increases at epoch 2 but eventually stabilizes, showing minor overfitting or instability during training.
3. Test Loss remains constant at 0, possibly indicating that the test set was too small or the model didn't generalize effectively during the test phase

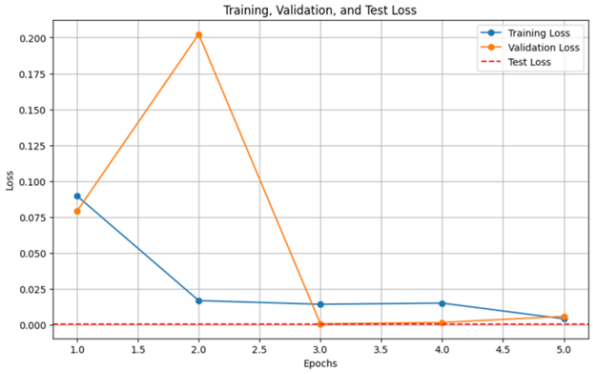

Figure 2. Loss Graph: Gender Prediction Model

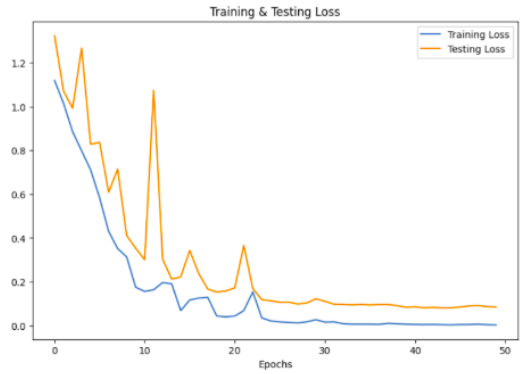

Figure 3. Loss Graph: Emotion Recognition Model

3.2. Model Training for Emotion Classification

3.2.1 Model Architecture

- The model used for emotion recognition is a 1D Convolutional Neural Network (CNN). The architecture consists of multiple convolutional layers followed by max pooling and dropout layers to prevent overfitting. The final output is passed through a fully connected layer with softmax activation to classify emotions into 7 categories.
- Input Shape: (X_train.shape[1], 1).
- Conv1D Layers:
  - 512 filters with kernel size 5, followed by batch normalization and max pooling.
  - 256 filters with kernel size 5, followed by batch normalization and max pooling.
  - 128 filters with kernel size 3, followed by batch normalization and max pooling.
- Dropout: Dropout layers (with rate 0.2) are added after certain layers to prevent overfitting.
- Fully Connected (Dense) Layers:
  - Dense layer with 512 units and ReLU activation.
  - Output layer with 7 units (for the 7 emotion categories) and softmax activation.

3.2.2 Hyperparameters

- Optimizer: Adam
- Loss Function: Categorical Cross-Entropy
- Metrics: Accuracy
- Batch Size:64
- Epochs: 50

3.2.3 Analysis of the Graphs

1. Training & Testing Loss (Figure 3):
   - Training Loss: Starts at 1.2 and decreases sharply to approximately 0.05 by epoch 50, indicating rapid learning and improvement in the model's ability to fit the training data.

Figure 4. Accuracy Graph: Emotion Recognition Model

   - Testing Loss: Starts at around 1.1 and decreases to about 0.1 by epoch 50, following a similar trend but remaining slightly higher than the training loss, which suggests some difficulty in generalizing to the test set.
2. Training & Testing Accuracy (Figure 4):
   - Training Accuracy: Begins at about 50% and reaches approximately 95% by the end of training. This indicates the model is fitting the training data well.
   - Testing Accuracy: Starts at around 50% and reaches about 92% by the end of training, showing that the model generalizes well to unseen data, though there's a small drop compared to training accuracy.

These numbers show that the model is learning effectively with high training accuracy (95%), good test accuracy (92%), and relatively low loss in both the training (0.05) and testing (0.1) datasets, indicating that the model does not overfit and generalizes well to unseen data.

The Confusion Matrix (Figure 5) shows how well the model performs in classifying the 7 emotions (Angry, Disgust, Fear, Happy, Neutral, Sad, and Surprise). The diagonal values indicate the number of correct predictions for each emotion, while off-diagonal values represent misclas-

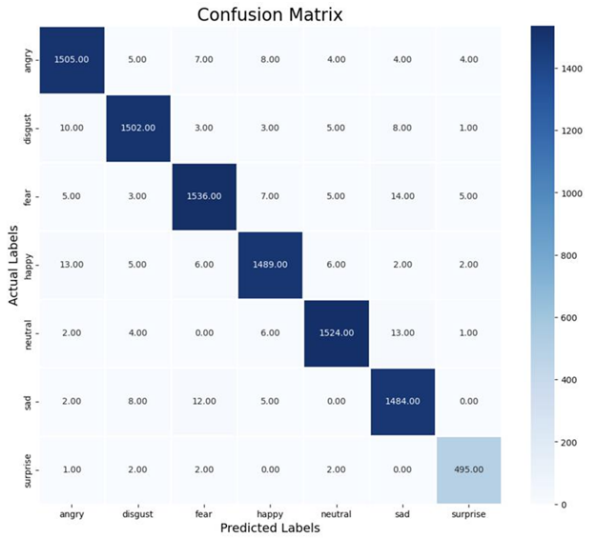

Figure 5. Confusion Matrix: Emotion Recognition Model

sifications. For example, 1505 instances of Angry were correctly classified, while 5 were incorrectly predicted as Disgust. The model performs well overall, with the highest miss-classification occurring for Surprise, which has a significant number of misclassifications.

4. End-to- End Application on Streamlit

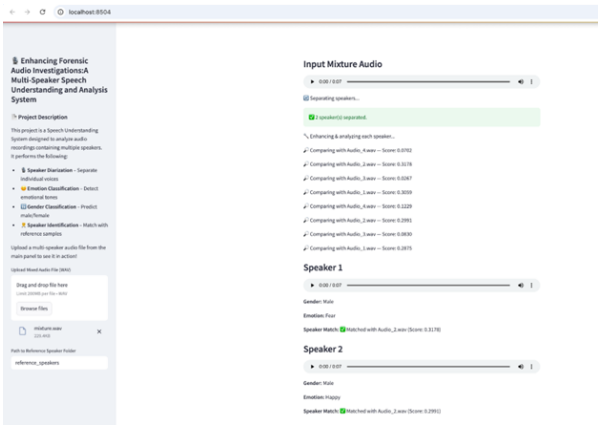

Figure 6. Application deployed on Streamlit

The Streamlit-based application (Figure 6) allows users to upload a multi-speaker audio file, which is then processed step by step: first, the system separates the voices using SepFormer, then classifies the emotion and gender of each speaker, and finally matches the speakers with reference samples, displaying a matching score. The results of each step, including separated audio, gender, emotion, and speaker match, are displayed for the user to interact with, making it ideal for tasks such as forensic audio investigations.

5. Discussion

The system performs well in separating and analyzing multi-speaker audio. However, one of the challenges encountered during the analysis is that speaker matching relied on a basic method of organizing audio files into folders, without proper labeling for more accurate matches. Specifically, the audio samples were placed in a single folder and matched accordingly, which worked in our controlled setup. However, for real-world applications, a more refined and systematic labeling and organization of audio samples is required to ensure accurate speaker identification and better matching results. The current approach lacks detailed metadata, which could affect its performance when applied in more dynamic and uncontrolled environments. For practical deployment, the system would benefit from incorporating more advanced techniques for speaker labeling and better organization of reference data.

6. Conclusion

In this work, we develop a multi-speaker speech understanding and analysis system that integrates speech separation, emotion recognition, gender classification, and speaker identification. The system demonstrated robust performance, effectively separating voices and accurately predicting emotions and genders from audio. However, for real-world applications, the current speaker matching process requires more refined labeling and systematic organization of audio data to ensure precise identification. The system provides a solid foundation for forensic audio analysis, and future work will focus on improving the speaker matching process, incorporating real-time processing, and expanding the data set to handle more diverse and challenging acoustic scenarios.

7. Links

Github Public Repo link: https://github.com/m23csa516/speechunderstanding_Poject/tree/main

References

[1] Gender recognition by voice dataset. https://www.kaggle.com/datasets/murtadhanajim/gender-recognition-by-voiceoriginal/data. 2

[2] Librimix: Data generator script. https://github.com/JorisCos/LibriMix/blob/master/generate_librimix.sh. 1

[3] Microsoft unispeech. https://github.com/microsoft/UniSpeech/tree/main/downstreams/speaker_verification. 2

[4] Speech emotion recognition dataset. https://www.kaggle.com/datasets/dmitrybabko/speech-emotion-recognition-en/data. 2

[5] Speechbrain sepformer-whamr. https://huggingface.co/speechbrain/sepformer-whamr. 1

[6] S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366, 1980. 1

[7] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5200–5204, 2016. 2