Gender and Speaker Classification

Mu-Ruei Tseng
Texas A&M University

1 Gender Classification

1.1 Dataset

Mel-Frequency Cepstral Coefficients (MFCCs) are widely used in speech and audio processing to capture the spectral characteristics of a signal. A typical MFCC representation is shown in Figure 1.

The first MFCC feature (MFCC0) is primarily associated with the overall loudness of the audio signal, expressed on a logarithmic scale:

$$MFCC0 = \log \sum_i S[i] \tag{1}$$

where $S[i]$ represents the power spectrum corresponding to the i^{th} Mel filter.

Higher-order coefficients (MFCC1+) encode formant-related information, which is crucial in characterizing vocal tract resonances. Since male voices generally exhibit lower formant frequencies than female voices, MFCC features can effectively represent audio signals for gender classification.

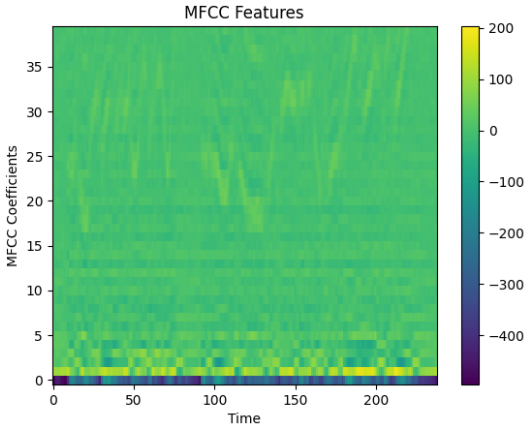

Figure 1: Sample MFCC of an audio file from LibriSpeech dataset (fs=16000Hz).

However, MFCCs inherently contain temporal variations, and audio signals can vary in length. To obtain a consistent representation, we compute the time-averaged MFCCs, which summarize the overall spectral distribution of the entire signal. This approach is particularly beneficial for gender classification, as it focuses on speaker-specific characteristics rather than speech content, simplifying the classification task. To ensure consistency across different MFCC coefficients, we apply z-score normalization to each coefficient since MFCC0, which represents the log-energy of the signal, can have a significantly different scale compared to the higher-order MFCCs.

We visualized the distribution of the average MFCCs for each audio sample using t-SNE (see Figure 2). The results reveal distinct distributional differences between male and female speakers. Additionally,

individual speakers form separate clusters, suggesting that the average MFCCs encode both gender and speaker-specific characteristics. We observed that speaker IDs 2078, 2902, and 7976 are labeled as male, yet their distributions closely align with the female cluster. This suggests potential labeling errors. To verify this, we examined the dataset and discovered that speaker ID 2078 (Kathy Caver) and 7976 (Jennifer Rutters) were incorrectly labeled as male. Therefore, we removed these mislabeled samples (outliers) from the classification task.

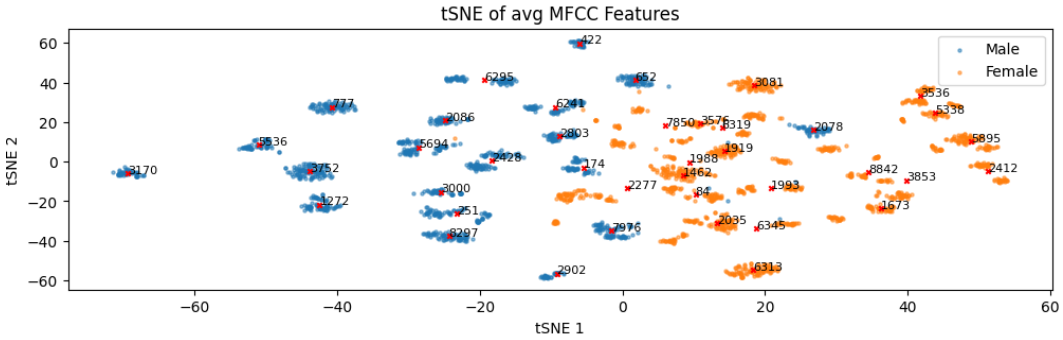

Figure 2: The t-SNE visualization depicts the distribution of the average MFCCs for each audio sample, where each number in the figure represents a unique speaker ID. The visualization reveals distinct clustering patterns between male and female speakers, indicating that the average MFCC is a meaningful feature for gender classification.

We also visualized the differences in average MFCCs between male and female speakers in Figure 3. The results show that female speakers tend to have higher MFCC0 values compared to male speakers, indicating that female voices generally exhibit higher energy levels. Additionally, male speakers have higher average values in the MFCC1–19 range, while female speakers exhibit higher values in the MFCC20–40 range. This observation aligns with the fact that male voices typically have lower formant frequencies than female voices.

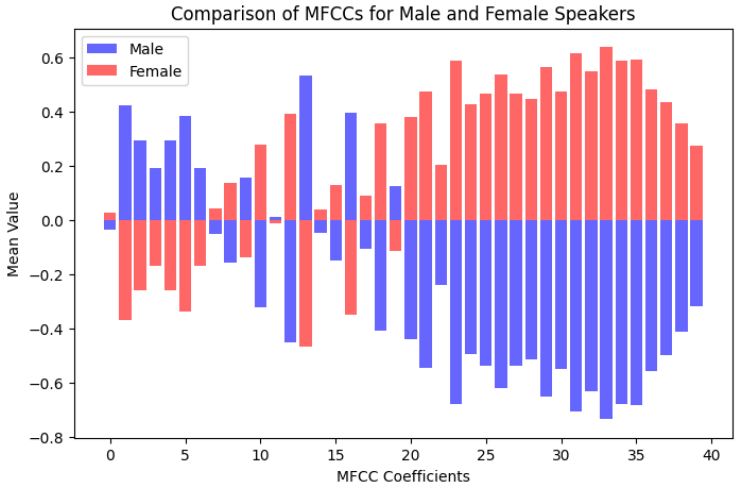

Figure 3: Difference between the normalized average MFCCs between male and female speakers.

1.2 Methodology

In this section, we evaluate the performance of four classifiers (Logistic Regression, SVM, MLP, and CNN) and compare their results. To ensure the robustness of our experiment and prevent data leakage, we employ 5-fold cross-subject stratified cross-validation. This approach maintains a consistent ratio of male and female samples between the training and test sets while ensuring that samples from the same subject appear exclusively in either the training or test set.

SVM Model For the SVM model, we use a RBF kernel with a regularization parameter of 1.

MLP Model The MLP model consists of three fully connected layers with output dimensions of 128, 64, and 1, respectively. Each layer is activated using the ReLU function. The model is trained for 20 epochs using the SGD optimizer with the following hyperparameters:

- Learning rate: 1e-3
- Weight decay: 1e-3
- Momentum: 0.9

CNN Model For the CNN model, we experiment with two 1D convolutional layers with output filter sizes of 16 and 32, respectively. Each convolutional block consists of:

1. A 1D convolution with a kernel size of 3 and padding of 1
2. Batch normalization
3. ReLU activation
4. Max pooling with a pool size of 2

The output is then flattened and passed to a classifier consisting of two fully connected layers with output sizes of 32 and 1, respectively, with ReLU activation between them. The model is trained for 50 epochs using the SGD optimizer with the same hyperparameters as the MLP model.

1.3 Result

Across all models, the performance remains relatively consistent, with no significant differences observed (Figure 4). The highest accuracy was achieved by Logistic Regression (93.44%), followed closely by CNN (93.23%), SVM (92.95%), and MLP (90.75%). These results suggest that all four models perform effectively for the given classification task. The similarity in accuracy across different architectures indicates that the extracted average MFCC features contain sufficient information such that simpler models like Logistic Regression and SVM can achieve performance comparable to deep learning-based approaches like CNN and MLP. A more detailed result can be found in the provided Jupyter Notebook.

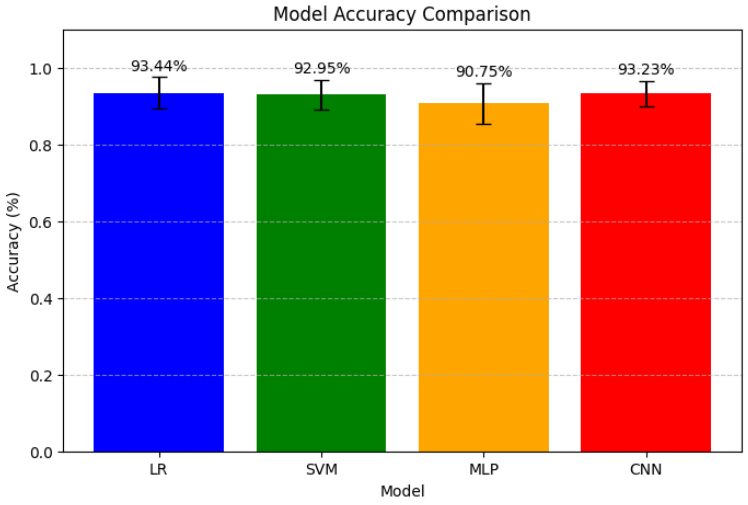

Figure 4: The mean classification accuracies across 5 folds for the four models: Logistic Regression (LR), Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), and Convolutional Neural Network (CNN).

2 Speaker Classification

In this section, we perform speaker classification based on audio recordings. Instead of using average MFCCs, we utilize the feature extractor from wav2vec (1), a self-supervised model designed to learn high-level representations from raw audio waveforms. These extracted features are then used as input to an LSTM-based model for the classification task.

2.1 Dataset

Since audio recordings vary in length, we split each audio into 2-second chunks, discarding the last segment if it is shorter than 2 seconds. We use wav2vec as a feature extractor to obtain representations for each chunk, resulting in a feature matrix of size $(T, D) = (99, 768)$, where $T = 99$ represents the time steps and $D = 768$ is the feature dimension.

Similar to the previous section, we analyze the feature distribution using t-SNE on the time-averaged features (see Figure 5). The t-SNE visualization reveals that features extracted from wav2vec tend to form distinct clusters for the same speaker. However, for some speakers, the data points appear more dispersed, suggesting that a two-dimensional projection may not be sufficient for clear separation.

To address this, in the following section, we develop machine learning models that leverage the full high-dimensional feature space to improve speaker classification performance.

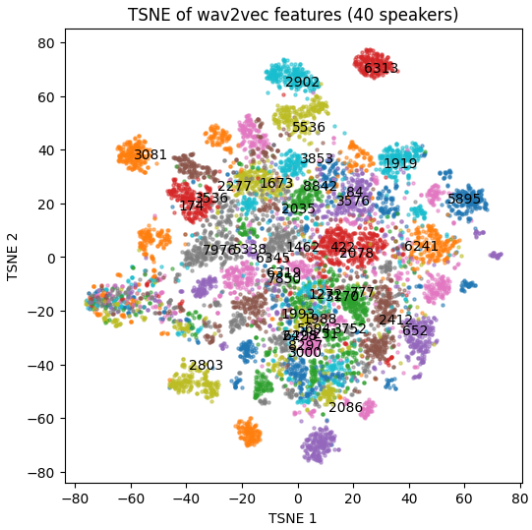

Figure 5: The t-SNE visualization of the average wav2vec features for each audio chunk. Each color represents a different speaker, and the numbers indicate the corresponding speaker IDs.

2.2 Methodology

In this section, we first split the data into training and testing sets. Unlike Task 1, which uses a cross-subject split, we apply a cross-audio split, ensuring that the same audio file does not appear in both the training and testing sets. This approach prevents data leakage, ensuring that the model generalizes to unseen audio rather than memorizing specific recordings. We also use stratified 5-fold cross-validation to split the data, ensuring that the proportion of each speaker remains consistent between the training and testing sets.

Since the wav2vec features are sequential data, we employ an LSTM-based model for speaker classification. The model consists of two bidirectional LSTM layers with a hidden dimension of 256. The output from the final LSTM layer is then passed to a classifier consisting of two fully connected layers with sizes 128 and $n_{classes}$, where $n_{classes} = 40$ (corresponding to the 40 speakers in the dataset). The model is trained for 20 epochs using the Adam optimizer with a learning rate of 1e-3 and a weight decay of 1e-3. For the loss function, we use cross-entropy loss to optimize multi-class classification.

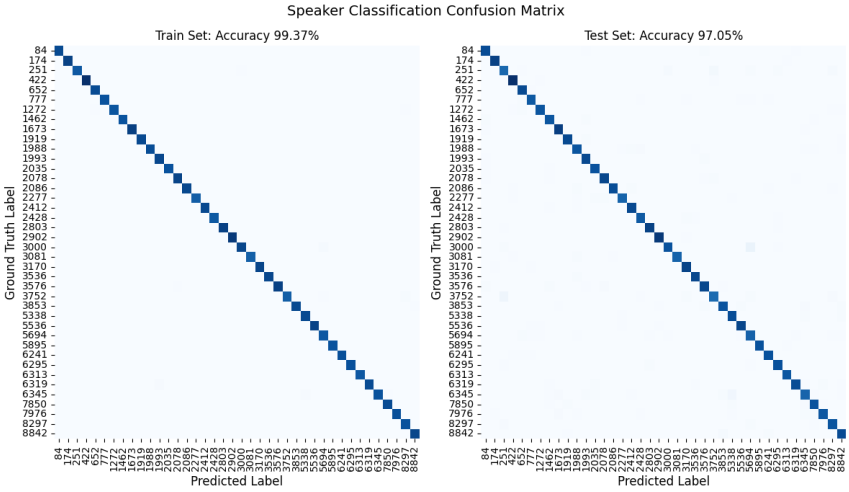

Figure 6: Confusion matrices for speaker classification on the train set (left) and test set (right) using the LSTM-based model with wav2vec features.

2.3 Result

We analyze the results using both quantitative and qualitative approaches. For the quantitative result, we present the average confusion matrix computed from the 5-fold cross-validation experiments to evaluate classification performance across different speakers. For the qualitative result, we extract the embedding outputs from the final bi-LSTM layer as speaker embeddings and visualize their distribution using t-SNE. This provides insights into how well the model separates different speakers in the learned feature space.

From the quantitative results (Figure 6), the model achieves 99.37% accuracy on the training set and 97.05% accuracy on the test set, demonstrating strong classification performance with minimal misclassification errors.

The t-SNE visualization of speaker embeddings (Figure 7) reveals that most speakers form distinct clusters, confirming the effectiveness of the learned feature representations.

Overall, these results indicate that the LSTM-based model with wav2vec features is highly effective in speaker classification, leveraging both temporal and spectral information to achieve high accuracy.

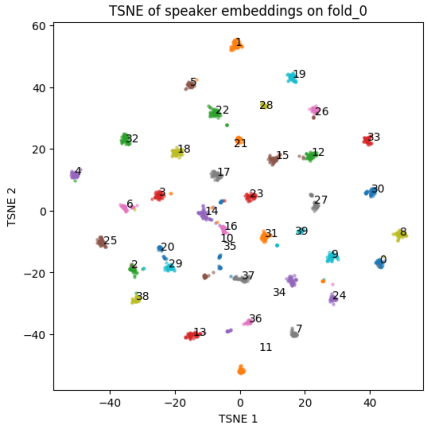

Figure 7: The t-SNE visualization of speaker embeddings extracted from the bi-LSTM model on fold 0. Each color represents a different speaker, and the numbers indicate speaker IDs.

3 Conclusion

This project demonstrates the effectiveness of MFCC and wav2vec features for gender and speaker classification tasks. MFCC-based features provide a compact spectral representation that captures formant characteristics, making them useful for gender classification. In contrast, wav2vec, a self-supervised representation learning model, extracts richer contextualized embeddings from raw audio, enabling more robust speaker classification. The code and dataset used in this study are available at https://github.com/Morris88826/gender_speaker, enabling reproducibility and further research.

References

[1] Baevski, A., Zhou, H., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations (2020), https://arxiv.org/abs/2006.11477