Project：

Voice Gender Classification Using Fourier Transform Features and Ensemble Machine Learning Models

Name: Renjie Dai

SID:490162237

1 Background

Voice gender classification is an increasingly important and popular area with a wide range of applications that profoundly change how humans interact with technology and with each other. The capacity to discern a speaker's gender in the context of human-computer interaction enables more responsive and customised voice-activated gadgets and virtual assistants. These systems can personalise services, recommendations, and responses based on the gender of the user, which improves the user experience in general. Voice gender classification gives an extra degree of biometric authentication to security systems. Gender verification can be a critical step in the verification process when voice is used as a key identifier, like in emergency response systems or secure access control. When visual identification is neither possible or trustworthy, this use is very important. Furthermore, knowing the characteristics of callers—including their gender—can help marketers and telecom professionals develop more focused marketing campaigns and improve customer service. With this information, businesses can better communicate with their clients, tailor their approach, and create marketing strategies that work better.

Gender-specific vocal traits and communication patterns can be studied with the use of voice gender classification in social and psychological research. Understanding gender dynamics in social interactions, language studies, and even health-related sectors where voice alterations may signal certain medical disorders are all affected by this research. Furthermore, gender classification helps the entertainment sector with automated dubbing, gaming, and content tailoring. Through the identification of user or character gender in games and media, companies may provide a more personalised and engaging experience. Due to the inherent variety of human voices and the requirement for advanced tools to effectively record and interpret these subtleties, voice gender classification is a challenging task. Gender classification based on voice is a challenging undertaking because vocal characteristics can be influenced by various factors, including age, ethnicity, and emotional state. But thanks to developments in machine learning and audio processing, gender classification from voice data is becoming more accurate and reliable, opening the door to more creative uses and improved user experiences.

In conclusion, voice gender classification is more than simply a technical difficulty; it opens up a wide range of applications that improve the usability, security, and intuitiveness of technology. As this area develops, it should make it possible for people to engage with technology in increasingly more efficient and customised ways.

2 Related Work

Badhon, Rahaman, and Rupon's (2019) study explores the automated gender classification of Bengali voices using machine learning. By extracting Mel-frequency cepstral coefficients (MFCCs) from voice signals and employing models like Logistic Regression, Random Forest, and Gradient Boosting, they achieved an impressive 99.13% accuracy in gender identification. Their dataset comprised over 1652 samples from more than 250 speakers, highlighting the importance of voice variety rather than quantity in gender classification. This research underlines the growing significance of gender detection in voice recognition systems, especially for languages like Bengali, as the world increasingly leans towards voice-based applications.

Tzanetakis (2005) investigated the efficacy of bootstrapping in audio-based gender identification, particularly for large datasets where extensive manual annotation is impractical. The study utilized a dataset comprising 10 news broadcast excerpts from the TREC 2003 video retrieval set, focusing on automatic classification with minimal user annotations. The findings demonstrated that Neural Networks, when employed in bootstrapping, yielded the best performance in gender identification, even with limited training data. This approach significantly reduces manual annotation time, making it a promising method for efficiently processing large audio datasets in multimedia information retrieval systems.

Lee, Kang, Kim, and Chang's (2008) study presents a novel approach to voice-based gender identification using Support Vector Machine (SVM). The SVM, known for its high classification performance in binary categorization, is compared to a Gaussian Mixture Model (GMM)-based method utilizing mel frequency cepstral coefficients (MFCC). Their innovative contribution lies in the fusion of MFCC with the fundamental frequency to enhance identification accuracy. The results revealed that the SVM outperforms the GMM-based method, and the performance is further improved with the proposed feature fusion scheme. This research signifies a step forward in the accuracy and efficiency of gender identification using speech signals.

Ali, Islam, and Hossain (2012) developed a gender recognition system based on speech signal processing, emphasizing the crucial role of feature selection in achieving high classification accuracy. They focused on power spectrum as the primary feature, using the First Fourier Transform (FFT) for extraction. The system was tested using speech signals from 10 individuals, achieving an average recognition accuracy of 80%. The paper highlights the effectiveness of statistical analysis and threshold techniques in pattern comparison, suggesting that dynamic thresholds could potentially enhance recognition accuracy. This study contributes to the field by demonstrating the potential of simple computational methods in gender recognition systems, while also acknowledging the challenges in maintaining efficiency with an increasing number of reference speakers.

To summarise, most of the evaluated research on audio signal-based gender classification makes use of intricate feature engineering and is significantly dependent on large amounts of sound data. These investigations frequently use highly developed machine learning models. Nonetheless, the approaches' complexity notwithstanding, the gender categorization accuracy is typically moderate. This suggests that both feature engineering and algorithmic techniques in this field have room for further development and optimisation.

3 Problem Formulation

The study uses the Fourier Transform for feature extraction in an effort to provide a fast and effective approach for gender classification from voice signals. This conversion captures the essential features of both male and female voices, emphasising elements such as pitch and frequency components. These features are then fed into a number of widely used machine learning models in order to train and forecast simultaneously. This approach is special because of its ensemble methodology, which aims to improve accuracy and robustness by determining the final classification using a vote system among the several models. This method addresses a significant difficulty in machine learning and audio signal processing by attempting to strike a compromise between classification accuracy and computational economy.

4 Methodology

4.1 Fourier Transform for extraction

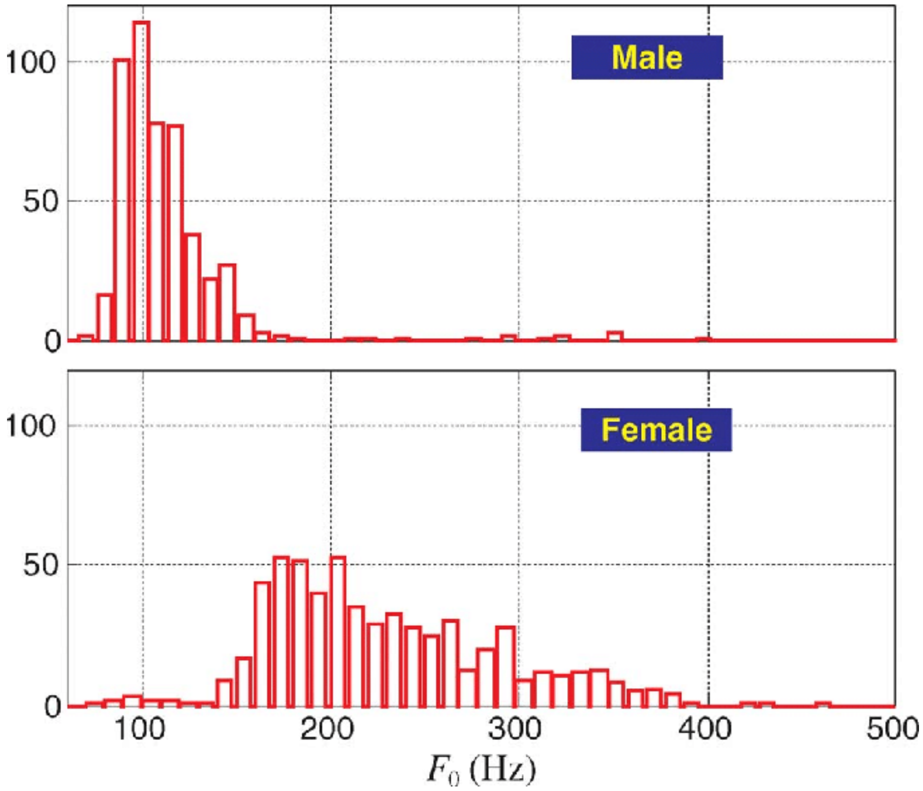


Figure 1. Frequency Distribution Patterns of Male and Female Voice Signals

Figure 1 illustrates the distinct frequency distribution patterns of male and female voice signals. This distinction in frequencies is why humans can effortlessly differentiate between male and female voices by merely listening. By digitally representing the sound's frequency, we can enhance our ability to discern voices based on their unique frequency characteristics. Male voices typically resonate within a 50 Hz to 250 Hz frequency range, as depicted by the upper histogram. On the other hand, female voices tend to have frequencies ranging from 100 Hz to 500 Hz, as showcased in the lower histogram. Utilizing the Fourier Transform, we can effectively extract these pivotal frequency features, offering a systematic approach to gender classification through voice analysis.

4.2 Data Preprocess

Given the task at hand, our preprocessing involves multiple steps to ensure that the data is suitable for model training. Here's a comprehensive breakdown of the preprocessing steps:

- Dataset Overview: We have a dataset comprising over 5,000 m4a files, each containing voice recordings of individuals speaking in various languages. This rich variety increases the robustness of our model as it's exposed to a plethora of accents and intonations.
- Segment Extraction: From each audio file, we randomly extract a 2-second segment. This duration is chosen based on prior knowledge that a 2-second snippet contains adequate voice patterns for gender identification. This also ensures uniformity in our data, as each sample fed into the model is of the same length.
- Bandpass Filtering: Next, a bandpass filter is applied to each 2-second segment, retaining frequencies in the range of 50-500Hz. This range is critical as it typically encompasses the fundamental frequencies of human voices. By filtering out frequencies outside this range, we remove unnecessary noise and other components that aren't beneficial for our task.
- Feature Extraction: For every filtered voice segment, we sample 200 amplitude values at equidistant frequency points within the 50-500Hz range. These amplitude values serve as the features that represent each voice sample. Sampling 200 points

ensures that we capture the essence of the voice frequency spectrum without making the dataset too dense or computationally heavy.

- Label Assignment: Finally, each voice sample is labeled. Male voices are assigned a label of '0', while female voices are assigned a label of '1'. This binary classification simplifies our task, enabling us to utilize a wide array of machine learning algorithms optimized for binary tasks.

Upon completion of these steps, we obtain a well-structured dataset with uniform voice samples, each represented by 200 features and a binary label. This dataset is now primed for splitting into training and testing sets and is ready to be fed into our machine learning model for gender identification.

4.3 Model selection

In our endeavor to create a robust gender identification system using voice frequency features, choosing the right machine learning model is paramount. Here's a brief overview of each proposed model and its potential suitability:

- Random Forest: Random Forest is an ensemble method that builds multiple decision trees and combines their outputs. Given its ability to handle large datasets, capture complex relationships, and provide feature importance, it is an excellent candidate for our voice classification task.
- Support Vector Machine (SVM):SVM excels in binary classification problems and can efficiently manage high-dimensional data. Using the kernel trick, SVM can even handle non-linear boundaries, which might be present in our voice data.
- Adaboost: As a boosting algorithm, Adaboost focuses on instances that are harder to classify, thereby enhancing model performance. It's especially effective for binary classification problems and could offer a high accuracy rate for our dataset.
- Logistic Regression (LR):A simple yet effective model for binary classification. While it assumes a linear relationship between features and the log odds of the outcome, with proper feature engineering, it can still be a strong contender.
- Deep Neural Network (DNN):DNNs have the capacity to capture intricate patterns

in large datasets, making them suitable for our task. With multiple layers and activation functions, DNN can model the complex relationships that simpler models might miss.

- K-Nearest Neighbors (KNN):KNN classifies based on the majority class of 'k' nearest data points. Given the nature of voice data and potential overlapping frequencies between genders, KNN can be effective, especially with an optimized value for 'k'.
- Naive Bayesian: Based on Bayes theorem, this probabilistic classifier can be highly effective for binary tasks. Given its assumption of feature independence, feature selection and engineering will play a pivotal role in its success for our dataset.

After individually training the seven models, a common approach to enhance the robustness and accuracy of the prediction is to employ ensemble methods. One such method is 'Voting', which aggregates the predictions of each model to arrive at a final decision. Here's how this can be implemented for our voice gender identification system. Here, we use Hard Voting: Each of the seven models predicts a class label (Male or Female) for a given voice sample. The class label that receives the majority vote becomes the final prediction.

5 Results and Analysis

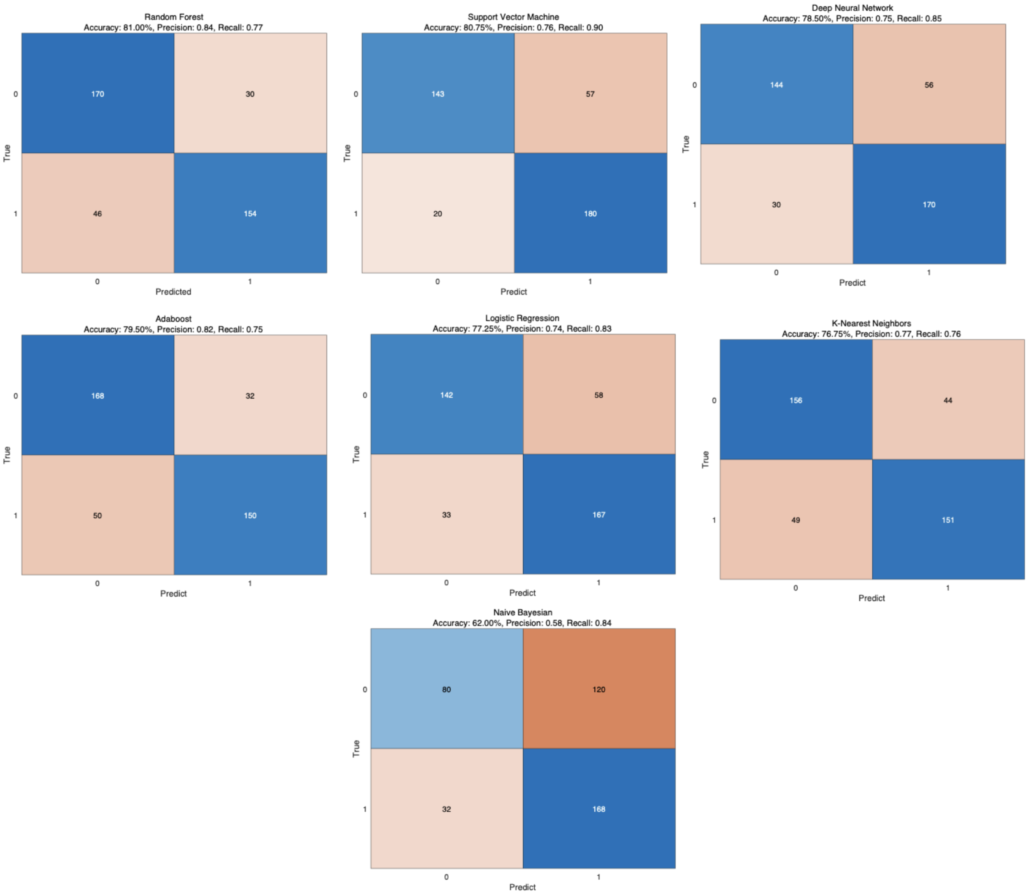


Figure 2 Comparative Analysis of Machine Learning Models for Voice Classification

Figure 2 depicts the classification performance of various machine learning models on a given dataset. The Random Forest model yields the highest accuracy at 81.00% with a precision of 0.84 and recall of 0.77. This is closely followed by the Deep Neural Network model with an accuracy of 75.98%. The Logistic Regression and K-Nearest Neighbors models also perform comparably with accuracies around 77%. The Support Vector Machine and AdaBoost models have accuracies of 80.76% and 79.38%, respectively. The Naive Bayesian model, however, lags behind slightly with an accuracy of 69.67%. The distinction between the 'True' and 'Predicted' labels across the matrices provides insights into the true positive, true negative, false positive, and false negative predictions of each model. Overall, while the Random Forest stands out as the top performer in this specific context, most of the models showcase a commendable classification capability, emphasizing the importance of choosing the right algorithm based on the nature of the data and the problem at hand.

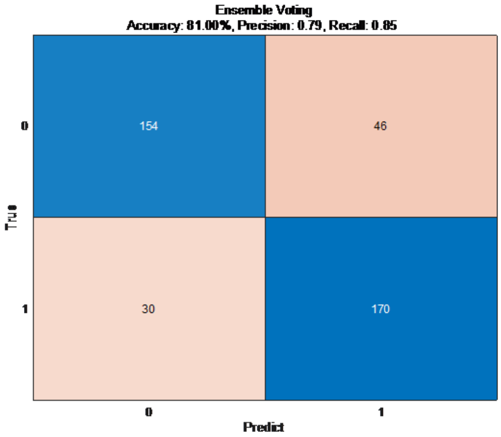
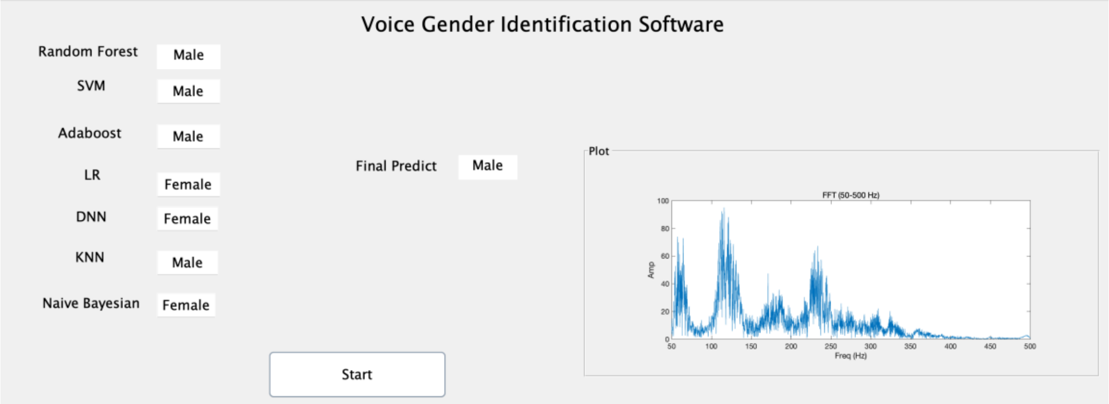


Figure 3 Comparative Analysis of Ensemble Voting Mechanism

The adoption of an ensemble strategy, amalgamating the predictions from seven distinct models via a voting mechanism, has manifested notable results as displayed in the confusion matrix. The ensemble approach leverages the strengths of individual models, minimizing their inherent biases and vulnerabilities, leading to enhanced precision. With 154 and 170 instances accurately classified for class 0 and class 1 respectively, the model's commendable accuracy becomes evident. Nevertheless, a few misclassifications are still present, with 46 instances incorrectly categorized as class 1 and 30 as class 0. While these discrepancies highlight areas for refinement, the ensemble's collective wisdom largely mitigates individual model errors, enhancing its robustness and generalization capability over diverse data scenarios.


Voice Gender Identification Software

Random Forest	Male
SVM	Male
Adaboost	Male
LR	Female
DNN	Female
KNN	Male
Naive Bayesian	Female

Final Predict Male

Figure 4 Voice Gender Identification Interface using MATLAB

I have created a sophisticated interface using MATLAB GUI. This interface is designed to allow users the capability to upload and process any given audio file. Once an audio file is uploaded, the software harnesses our pre-trained machine learning models to analyze the audio data. One of the primary features of this software is its ability to generate a spectral plot, providing a visual representation of the audio's frequency components. Following this, the software runs the audio data through each of our individual trained models, which then outputs their respective predictions. These individual model predictions are then integrated, and the software presents a final, consolidated prediction based on the collective analysis.

6 Conclusion

In conclusion, the development of the Comprehensive Voice Gender Identification Interface using MATLAB showcases the integration of various machine learning models for real-time audio analysis. By leveraging a combination of Random Forest, SVM, Adaboost, LR, DNN, KNN, and Naive Bayesian, the interface delivers robust and accurate gender prediction from audio input. The GUI not only streamlines the process of audio file uploading and spectral visualization but also provides instantaneous feedback on the predictions of each individual model and a consolidated final result. This innovation is indicative of the potential of combining multiple algorithms for enhanced performance in audio-based machine learning tasks.

Reference

Ali, M. S., Islam, M. S., & Hossain, M. A. (2012). Gender recognition system using speech signal. International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), 2(1), 1-9.

Badhon, S. S. I., Rahaman, M. H., & Rupon, F. R. (2019, November). A machine learning approach to automating Bengali voice based gender classification. In 2019 8th International Conference System Modeling and Advancement in Research Trends (SMART) (pp. 55-61). IEEE.

Lee, K. H., Kang, S. I., Kim, D. H., & Chang, J. H. (2008). A support vector machine-based gender identification using speech signal. IEICE transactions on communications, 91(10), 3326-3329.

Tzanetakis, G. (2005, August). Audio-based gender identification using bootstrapping. In PACRIM. 2005 IEEE Pacific Rim Conference on Communications, Computers and signal Processing, 2005. (pp. 432-433). IEEE.