YZV303E Deep Learning
Project Proposal
Voice-based Gender and Identity Detection Using Deep Learning Model

Aybike Battal
Artificial Intelligence and Data Engineering
Istanbul Technical University
battal20@itu.edu.tr
150200343

Nazrin Abdinli
Artificial Intelligence and Data Engineering
Istanbul Technical University
abdinlin22@itu.edu.tr
150220925

I. Project Description

This project is about creating a deep learning model that can recognize a student's gender and identity from their voice data. Voice recordings will be collected from 50 students. Then the data will be prepared and the deep learning model will be trained for making these identifications.

The goal is to see how well voice features can be used for identifying gender and identity. This project could provide better understanding about use of voice features in identifying individual properties.

II. Problem Definition

In the project, the aim is to predict both the gender and identity of students from their voice recordings. As each student's voice is unique, it reflects characteristics about their identity and gender. Using deep learning, a model will be trained for identifying patterns of voice of each student and differences between characteristics of male and female voices.

While voice features like tone and accent help identification process easier, additional factors such as background noise or recording quality make it more challenging. The goal is to create a model that is accurate and reliable, able to handle these variations and work successfully with new voice data.

III. Dataset

It is planned to create a dataset with voice recordings from 50 participants. Half of the participants will be male and the other half will be female individuals, and it will help to work on balanced dataset for gender prediction. Each participant will record 6 voice samples and this process will result in 300 recordings in total, with 200 samples and 100 samples used for training and testing respectively.

Participants will mostly be ITU students, and all recordings will be collected in a quiet environment. The data will be anonymized and the permission will be received from each student for ethical purposes.

After recording process, the audio will be preprocessed for maintaining consistency by standardizing volume levels and removing background noise. This preparation will provide a clean and reliable dataset for analysis of this project.

IV. Literature Review

The use of voice for gender and identity recognition has attracted attention in speech processing owing to the relative ease of extracting key patterns from audio data using deep learning algorithms. There has been plenty of research conducted into the voice traits including pitch, tone and frequency that reveals the gender and identity of a person.

Convolutional neural networks (CNNs) have become dominant in audio recognition tasks, particularly with spectrograms, as they effectively capture the spatial relationships between frequency and time. Mohamed et al (2012) talked about performing acoustic modeling with deep belief networks and underlined usefulness of neural networks when working on an audio signal analysis task. Also, Amodei et al., (2016) made a contribution towards deep learning models by developing Deep Speech 2 framework which was used for end to end speech recognition for English and Chinese languages, strengthening the case for neural models as robust against noise and variable input data. These findings were extended by Hershey et al. (2017) who also focused on the strengths of CNN architectures in large-scale audio classification.

It has been emphasized through these studies how realistic it is to apply CNNs for the analysis of spectrograms in terms of gender recognition and individual recognition. There still remain challenges such as noise disturbance, and the quality of the recording, which require efficient preprocessing approaches like pitch shifting and time stretching. These augmentations enhance the model's ability to generalize to variations in voice characteristics, improving its classification accuracy and reliability.

Building on insights from these foundational studies, this project aims to leverage CNNs applied to spectrograms for

recognizing speaker gender and identity from voice, utilizing advanced preprocessing methods and reliable evaluation metrics.

V. Methodology

In order to analyze and classify voice data, the first step is converting the audio recordings into spectrograms using the Short-Time Fourier Transform (STFT) methods. These spectrograms represent the audio signals visually, and will be normalized and augmented using techniques such as pitch shifting and time stretching. Pitch shifting helps the model to generalize better across individuals with similar tones or vocal patterns, and time stretching helps the model learn to focus on important features in the audio signal rather than being sensitive to speed variations.

Additionally, a Convolutional Neural Network (CNN) model will be developed for learning features from the spectrograms. CNN is chosen due to its ability to effectively process 2D spectrograms and capture frequency-time patterns. CNNs are robust to pitch and timing variations, so they are ideal for gender and identity identification. Fully connected layers will also be a part of the model's architecture for determining the gender and identity of the participant.

The Adam optimizer and cross entropy loss function will be used to train the model and the most suitable hyperparameters will be selected for achieving the best performance. Finally, the evaluation of the model performance will be done using the metrics like accuracy, precision, recall, and confusion matrices.

VI. Pipeline

The project's pipeline consists of the following steps:

A. Data Collection

- Collecting voice recordings from 50 participants (25 male and 25 female)
- Gathering 6 voice samples from each student in a quiet environment for ensuring quality
- Obtaining ethical approval and anonymizing data for potential privacy compliance

B. Preprocessing

- Normalizing audio signals maintaining consistency in volume levels across all recordings
- Applying noise reduction techniques for removing background noise
- Converting audio signals into spectrograms using Short-Time Fourier Transform (STFT)

C. Data Augmentation

- Performing pitch shifting for simulating variation in voice tones and improving model generalization
- Applying time stretching for training the model on speed variations

D. Model Development

- Designing a Convolutional Neural Network (CNN) for processing 2D spectrogram inputs
- Including fully connected layers for predicting both gender and identity of a individual
- Utilizing the Adam optimizer and cross-entropy loss for model training

E. Hyperparameter Tuning

- Experimenting with different learning rates, batch sizes, and layer configurations for achieving the best model performance

F. Model Training and Validation

- Splitting the dataset into 200 training samples and 100 testing samples
- Training the model and evaluating it on unseen data using metrics like accuracy, precision, recall, and confusion matrices

G. Evaluation and Testing

- Analyzing performance using confusion matrices to understand strengths and weaknesses in classification
- Evaluating the model's reliability against noise and variations in recording quality

H. Deployment and Future Testing

- Preparing the model for deployment in real-world scenarios by further testing it with new datasets.
- Investigating potential applications for automated voice-based identification systems

References

[1] Mohamed, A., Dahl, G. E., & Hinton, G. (2012). "Acoustic Modeling Using Deep Belief Networks." IEEE Transactions on Audio, Speech, and Language Processing.

[2] Amodei, D., et al. (2016). "Deep Speech 2: End-to-End Speech Recognition in English and Mandarin." Proceedings of the International Conference on Machine Learning (ICML).

[3] Hershey, S., et al. (2017). "CNN Architectures for Large-Scale Audio Classification." Proceedings of IEEE ICASSP.