Bence Pór Sunčica Rosić

Sound Signatures: Digit and Gender Recognition of Human Speech Using Machine Learning

Abstract

In the realm of content-based multimedia indexing gender and digit identification using speech signals represents an important task. Existing techniques for audio signal representation involve Mel-Frequency Cepstral Coefficients (MFCC) feature extraction, waveform signal representation and mel spectrogram. In this paper we present a gender and digit identification approach based on the Librosa library in Python, broadly utilized for music information retrieval (MIR) and different sound processing tasks. In this project, we explore how to employ Librosa and PyTorch to process sounds and develop different models for prediction, including installation, feature extraction, model fitting and visualization. The audio classifier we employ for gender classification models the audio based on the MFCC features and uses a feed-forward neural network to classify the gender of the speaker. The first digit recognition model we propose relies on a CNN trained with spectrograms to predict spoken digits from 0 to 9. We also utilise some transfer learning and experimented with a second bigger, pre-trained model for digit recognition, called Wav2Vec2 developed by Facebook. The presented techniques show robustness, with gender classification results that attain 99.68% test accuracy and digit recognition results with 97.42% for the CNN and 93.17% for the pre-trained model.

Introduction

The significance of sound recognition and classification across various domains has driven researchers to explore innovative solutions to these challenges. In this paper, we introduce an approach that integrates MFCC feature extraction, spectrogram analysis, and the Pytorch deep learning framework to enhance sound recognition and classification. This method relies on the abundant information within audio signals to develop robust and accurate models.

The feature extraction process is performed differently depending on the objective. In the first part of the study, where we predict gender based on audio, feature extraction is executed by taking MFCC from features, which becomes our input data. In the second part of the project, we perform digit recognition from audio, by analyzing the spectrograms to identify key attributes that facilitate sound recognition. Both methods are used widely in the field of audio machine learning, so we attempt working with both.

The gender classification step employs a feed-forward network using Pytorch, leveraging the pattern recognition strengths of modern neural networks. The model is trained and validated on a diverse dataset of audio samples, ensuring high accuracy in identifying and categorizing various sound signals.

For developing the CNN for digit recognition, we employ the Optuna library for hyperparameter tuning with a validation set and tuning several parameters of the model to attain the highest possible accuracy. Because of the limitations of computational power, we are constrained to get as much out of hyperparameter tuning as possible, but one can easily set the number of epochs higher in the notebook to get an even better model.

Background Literature

Gender recognition based on the voice of a speaker involves detecting if a speech signal is delivered by a male or a female. Automatic gender detection has numerous possible applications. Gender identification has been mostly used as a way to advance recognition performance and to decrease the computation required. Accurate gender identification involves various applications in spoken language systems, “where it can permit the synthesis module of a system to respond appropriately to an unknown speaker” (Khanum & Sora, 2015). The most important feature which can differentiate between speaker’s genders is the fundamental frequency F0 with usual values of 110 Hz for male speech and 200 Hz for female speech. Still, there exist Gaussian distributions of such ranges, implying wide dispersion and thus the difficulty to categorize the acoustic signal reliably by using solely this criterion only (Pednekar, Tiware, & Bhagwat, 2008). Hence, this reinforces the need for researching novel gender identification based on the speaker’s voice.

An effective representation of speech signals enables their application in various domains, including speech recognition, speech synthesis, speaker identification, and speech translation (Toufa, 2020). This paper focuses on speech recognition, a branch of computational linguistics dedicated to developing methods and systems capable of converting spoken language into text. Speech recognition is widely utilized in real-world scenarios, such as in-car systems, smartphone applications and military technologies, including fighter jets and helicopters.

Digit recognition in reconstructed signals relies on pre-trained models trained on original audio recordings that exhibit high classification accuracy (Toufa, 2020). The key assumption is that if a classifier effectively learns to recognize digits from original recordings, it should also be capable of classifying reconstructed digits, even if they contain distortions. Various classifiers have been evaluated using different input types. First, the raw audio classifier processes unaltered audio recordings of length n and

utilizes a convolutional neural network (CNN) with 1D convolution. This network consists of five layers, incorporating batch normalization and dropout. Additionally, two other classifiers, also CNN-based, use spectrograms or gammatone-like spectrograms as input. Since both representations are two-dimensional (2D), they can be processed similarly to images.

Data

The dataset contains recordings of 60 individuals pronouncing numbers from 0 to 9. There are 500 recordings of each speaker, resulting in a total of 30,000 WAV files (we might end up using only a subsection of this because of computational reasons). The data is structured into 60 folders, with each folder representing one speaker and containing their respective 500 audio recordings. Alongside the audio files, there is a text file that provides metadata about each speaker. This metadata includes various details such as accent, age, gender, whether they are a native speaker, their country and city of origin, the date and time of the recording, and the specific room where the recording took place. For example, the first speaker in the dataset has a German accent, is 30 years old, male, not a native speaker, and recorded their speech in Wuerzburg, Germany, in a room labelled “Kino” on June 17, 2022. While most of these additional details could potentially help improve the performance of a machine learning model (maybe except for the date and time of the recording and the room), incorporating all of them may require more time and expertise than we currently have.

Methodology

We utilize Machine Learning to extract useful features from training data which improves the accuracy of the final models as well as their ability to adjust to a broad wide range of applications.

Upon feature extraction, preprocessing steps including normalization and scaling are performed to ensure alignment with machine learning models. The preprocessed features are then used to train a machine learning model; feed-forward network for gender classification and CNN for digit recognition. After training, the model is evaluated and validated using a separate dataset to assess its performance based on accuracy. This step ensures the model generalizes well to unseen data. Once validated, the trained models are applied for gender and digit prediction, where it classifies the gender and recognizes digits of new audio samples based on their extracted features. Lastly, we implement transfer learning in order to compare a pretrained model to our CNN for digit recognition.

The proposed approach follows a three-stage process: data upload, feature extraction and classification and digit recognition. The captured audio signals are preprocessed and transformed into spectrograms, which visually represent the frequency, amplitude and temporal characteristics of the sound data.

Analysis and Results

Gender Identification

The gender identification from the speaker's voice may be executed through two basic functions, feature extraction and classification. Feature extraction pertains to the process of transforming a speech signal into a set of parameters necessary for speaker gender identification. This transformation yields a vector called a feature vector. Classification refers to the action of classifying the extracted feature vectors into a suitable category i.e. either male or female. To achieve this function neural network architecture entitled multi-layered perceptron (MLP) with feed-forward algorithm is used.

For gender classification, we process metadata to extract gender labels associated with each audio sample. The dataset consists of male and female voice samples, which are mapped to their respective labels using metadata parsing techniques. The extracted labels were then used to form a data frame together with audio filenames to form a structured dataset.

The dataset was further processed by extracting relevant audio features. Each audio file is transformed into feature representations suitable for machine learning, by extracting MFCC coefficients and subsequently normalizing the feature space. MFCC librosa function is applied to features to extract the MFCC sequence, consisting of frame-wise features, cepstral coefficients, energy and acceleration coefficients. These features are stored in a structured format and subsequently split into training and testing sets, ensuring a balanced distribution of male and female samples.

We implement a feed-forward neural network for gender classification using PyTorch. An Artificial Neural Network (ANN) is a powerful pattern recognition system that mimics the human brain's neural processing capabilities. It consists of numerous artificial neurons interconnected based on a specific network structure. The primary goal of an ANN is to process input data and generate meaningful outputs.

A feedforward neural network consists of three layers:

- **Input Layer**: Receives the input vector - self.fc1
- **Hidden Layers**: Process and learn patterns through weighted connections, with the usage of ReLU to introduce linearity. In this project, the hidden layer is represented as self.fc2 and self.fc3, and it receives 64 inputs and yields 32 outputs, whereas the second hidden layer receives 32 inputs and produces 2 outputs.
- **Output Layer**: Produces the final predictions with a softmax function. Since `fc3` outputs **2 neurons**, they represent the two classes of male and female.

After training, the model is evaluated on the test dataset. The evaluation prioritizes accuracy as the primary metric. The final test accuracy is 99.68 %, with a total of 48 instances being misclassified. The model demonstrates a solid ability to differentiate between male and female voices. However, potential misclassifications could be attributed to overlapping voice characteristics or noise in audio recordings. Future improvements could include incorporating more advanced feature extraction techniques and exploring deeper neural network architectures for enhanced performance.

Digit Recognition

For the digit classification task, we convert each audio file into a **Mel spectrogram**—a 2D visual representation of sound where time is shown on the x-axis, frequency on the y-axis, and amplitude is encoded by brightness (Figure 1). Mel spectrograms are commonly used in speech recognition as they reflect how humans perceive pitch more accurately than standard spectrograms. To ensure compatibility with our model, we pad or truncate all spectrograms to maintain a consistent shape.

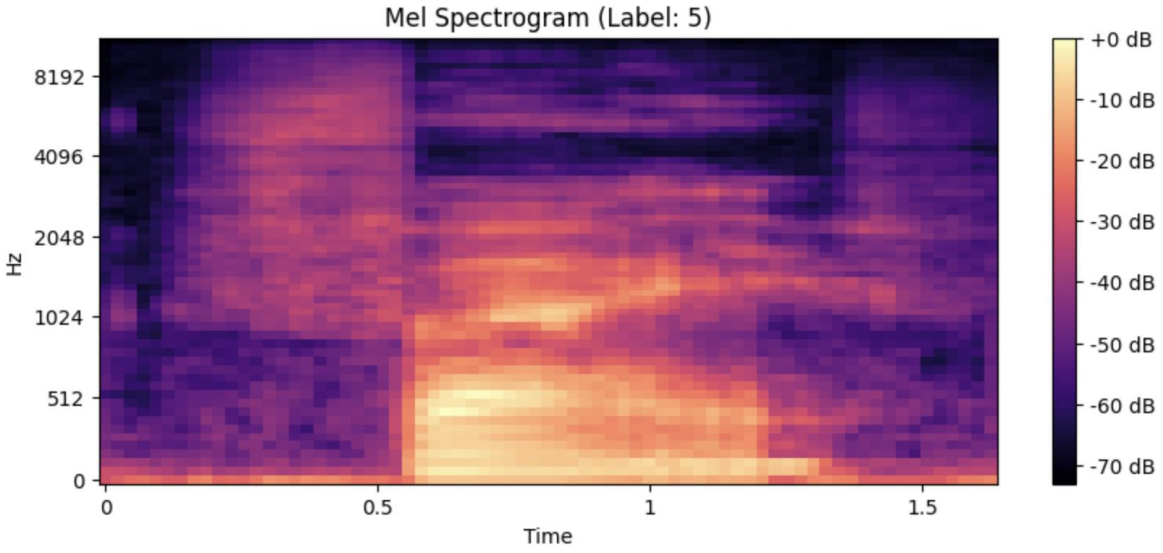

Figure 1: Mel Spectrogram

CNN

We conduct four different training runs using Convolutional Neural Networks (CNNs) to classify spoken digits from 0 to 9. In the first training, we use only 600 low-resolution Mel spectrograms and a simple CNN architecture, without any hyperparameter tuning. This model meets the required performance goals, achieving an accuracy of roughly 89%. For the second training, we introduce Optuna for hyperparameter tuning while still using only 600 spectrograms. However, test accuracy drops significantly to 59%, likely due to overfitting or the insufficient dataset size for robust tuning.

To address these issues, we expand the dataset to 6000 spectrograms and add a dropout layer to the CNN to reduce overfitting. In this third training, Optuna tunes hyperparameters such as the number of filters, fully connected layer size, learning rate, dropout rate, and batch size. This version achieves 97% test accuracy, showing that a larger dataset combined with regularization and tuning significantly improves performance.

Finally, we re-run the third training setup using the same model, parameters, and data, obtaining a slightly higher but consistent accuracy of approximately 97%. This consistency reinforces the stability and reliability of the model under well-balanced conditions.

Overall, our experiments confirm that CNNs effectively classify short spoken digits from audio, especially when trained on Mel spectrograms with sufficient data, regularization, and carefully tuned hyperparameters.

Transfer Learning

In addition to the CNN-based approach, we experiment with transfer learning using the pre-trained facebook/wav2vec2-base-960h model from Hugging Face. This model, originally trained for large-scale speech recognition, provides a powerful foundation for audio-related tasks. To adapt it to our digit classification problem, we apply Parameter-Efficient Fine-Tuning (PEFT) using LoRA (Low-Rank Adaptation), which significantly reduces the number of trainable parameters by injecting small, trainable matrices into the model's attention layers.

We implement a custom PyTorch Dataset class that loads raw audio files, pads them to a fixed duration, and extracts features using the Wav2Vec2FeatureExtractor. A new classification head is added to the model and trained using cross-entropy loss. Initially, we train the model for only one epoch due to time and hardware constraints, without any data augmentation or hyperparameter tuning (e.g., LoRA rank, dropout, or layer

configuration). Even under these limited conditions, the model achieves a test accuracy of 66.9%, showing early promise. In a follow-up run, without changing any parameters or increasing training time, the model unexpectedly reaches 93.17% test accuracy. While the exact reason for this performance jump remains unclear (possibly due to better weight initialization or slight variations in train/validation splits), it demonstrates the potential of Wav2Vec2 even with minimal fine-tuning.

Despite its strong results, we conclude that for a relatively simple task like digit recognition from short audio clips, CNNs trained on Mel spectrograms are not only sufficient but also more efficient. However, experimenting with Wav2Vec2 provides valuable insights into how modern transformer-based audio models can be adapted for downstream tasks with limited data and computing resources.

Conclusion

This project demonstrates the effectiveness of machine learning techniques in classifying human speech for both gender and digit recognition. We rely on Librosa for feature extraction and PyTorch for model training, successfully developing and evaluating models capable of accurately interpreting audio signals. Our feed-forward neural network using MFCC features achieves 99.68% accuracy in gender classification, validating the robustness of traditional feature-based models for speaker-level tasks. For digit recognition, a CNN trained on Mel spectrograms reaches a peak test accuracy of 97.42%, illustrating the potential of image-like audio representations for classification problems.

We also explore transfer learning with the Wav2Vec2 model, which achieves up to 93.17% accuracy despite minimal fine-tuning and limited computational effort. While CNNs ultimately prove more efficient and stable for our specific use case, Wav2Vec2 showcases the flexibility and promise of pre-trained models in speech-related tasks.

Overall, our experiments highlight that reliable results can be obtained even with moderate computational resources by using well-chosen features, appropriate architectures, and regularization techniques. Future work could focus on multi-language audio classification, real-time deployment, or the integration of metadata to further improve classification performance.

Bibliography

Khanum, S., & Sora, M. (2015). Speech based gender identification using feed forward neural networks. Int. J. Comput. Appl, 0975-8887.

Madhavi S. Pednekar, KavitaTiware and SachinBhagwat, “Gender Distinction Using Short Segments of Speech Signal”, IJCSNS International Journal of Computer Science and Network Security, VOL.8, No.10, October 2008.

Toufa, A.-S. (2020). DIGIT RECOGNITION APPLIED TO RECONSTRUCTED AUDIO SIGNALS USING DEEP LEARNING. In Aristotle University of Thessaloniki, MASTER THESIS [Thesis]. https://ikee.lib.auth.gr/record/322737/files/GRI-2020-28878.pdf