Sound Signatures

Digit and Gender Recognition of Human Speech Using Machine Learning

How do machines ‘hear’ and understand sounds?

Outline

- Problem Summary
- Data
- Theory I
- Theory II
- Digit Recognition
- Gender Identification
- Conclusion

Problem Summary

Why is sound recognition important?

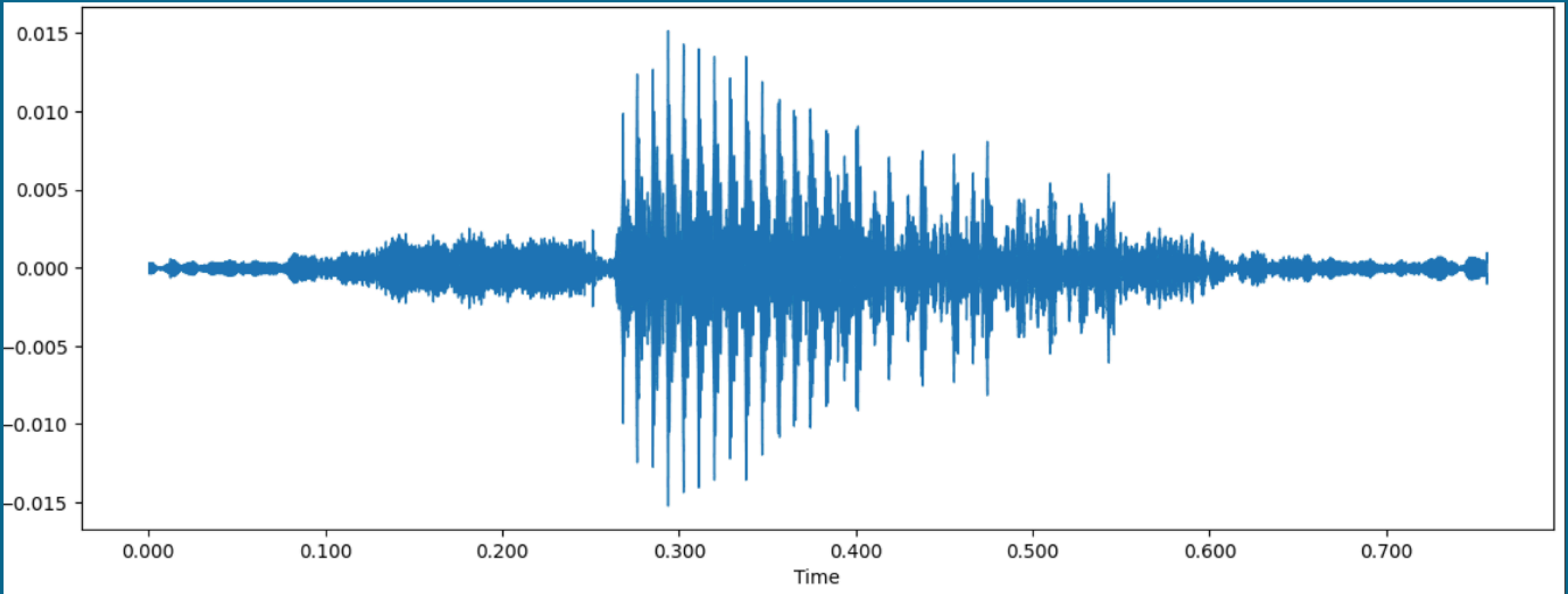
Applications in multimedia indexing, speech recognition and audio processing

Gender and digit identification using machine learning

Data

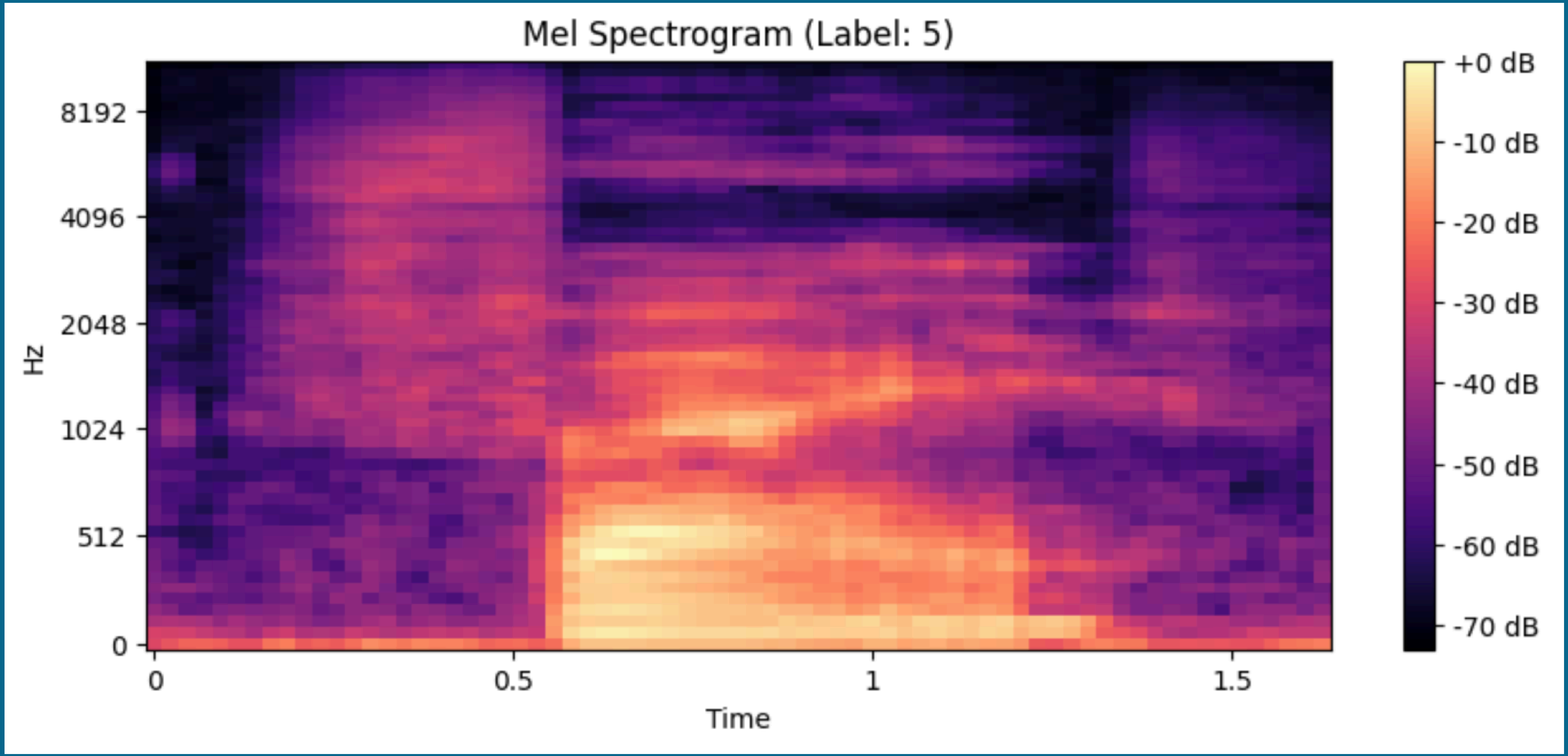
- 60 individuals recorded pronouncing numbers from 0 to 9.
- Each speaker has 500 recordings, totaling 30,000 WAV files.
- Structured into 60 folders, each representing one speaker.
- Each folder contains 500 audio recordings per speaker.
- Metadata Included: Accent, age, gender

Theory I



Wafeform

Spectograms



Mel Spectogram

Digit recognition

CNN

Data augmentation

- Simple CNN, 600 samples: 89%
- CNN with Optuna, 600 samples: 59%
- CNN (+dropout) with Optuna, 6000 samples

CNN

Architecture:

- convolutional layer (ReLU)
- pooling layer

x2

- flatten
- fully connected layer
- dropout
- fully connected layer
- output

Hyperparameter tuning with Optuna:

- Number of convolutional filters
- Fully connected layer size
- Batch size
- Dropout rate
- Learning rate

Test Accuracy: 97.42%
(second similar)

Transfer learning

Wav2Vec 2.0

(Wav2Vec2FeatureExtractor)

Fine-tuning with LoRa

1st Test results: 66.92%
2nd Test results: 93.17%
(only 1 epoch)

Gender Recogntion

- Extracted gender labels from files
- Feature extraction through MFCC
- Audio Classification using PyTorch
- Feedforward Neural Network (MLP)

Overview of neural network layers

- Input Layer: Audio features
- Hidden Layers: Feature learning with ReLU activation
- Output Layer: Softmax classifier for prediction

Training and Evaluation

- Fitted the model to extracted features
- Gender Recognition Accuracy: 99.68% (only 48 misclassified instances)

Conclusion

Key Takeaways

- Machine learning models effectively classify gender and digits from speech
- MFCC and spectrogram-based feature extraction improve classification accuracy
- Potential improvements with deeper neural networks and more diverse datasets

Applications

- Voice-controlled systems (e.g., smart assistants)
- Security & authentication (gender-based user identification)
- Speech-based accessibility tools
- Forensic analysis & speaker profiling

THANK YOU

Bence Pór

Suncica Rosic