Multiclass Speech Identification

Lean Jeng Wen Joshua
Interdisciplinary Program of Electrical Engineering and Computer Science
National Tsing Hua University
Hsinchu, Taiwan
joshualeanjw@gmail.com

KAI-YU CHENG
Department of Computer Science
National Tsing Hua University
Hsinchu, Taiwan
kaiyucheng2003@gmail.com

Aurick Daniel Franciskus S.
Department of Computer Science
National Tsing Hua University
Hsinchu, Taiwan
aurickd@gmail.com

WANG DE FU
Department of Physics
National Tsing Hua University
Hsinchu, Taiwan
tofuwang128@gmail.com

RAY XIANG SU
Interdisciplinary Program of Science
National Tsing Hua University
Hsinchu, Taiwan
bsnoopyb0610@gmail.com

CHUN YAO TSENG
Department of Computer Science
National Tsing Hua University
Hsinchu, Taiwan
yao2494@gmail.com

***Abstract*—We present an approach to multiclass speech identification, focusing on the simultaneous identification of multiple speakers in an audio stream. Utilizing a deep neural network, our method demonstrates potential in differentiating and recognizing individual speakers from a composite audio signal. The system's effectiveness is evaluated through experiments, showcasing its applicability in diverse real-world scenarios where multiple-speaker recognition is crucial.**

Keywords—speech recognition, multi-speaker identification, deep neural networks, acoustic signal processing, audio feature extraction.

I. Introduction

Speech recognition technology has become a cornerstone in various applications, ranging from virtual assistants to security systems. However, the challenge intensifies when the task extends to identifying and recognizing multiple speakers within the same audio stream. The complexity of this task arises from the need to differentiate between individual voices, understand each speaker's unique speech patterns, and do so in a potentially noisy environment.

In this paper, we explore approaches to deciphering human speech with a focus on enhancing the accuracy and efficiency of speech recognition systems. Our research primarily addresses the multi-class speech identification and recognition challenge, where the objective is to identify and recognize several speakers concurrently. This is a significant step beyond traditional speech recognition, which typically deals with a single speaker at a time.

Our approach stands out for its simplicity and accessibility. We do not rely on advanced or state-of-the-art models, which often require extensive resources and expertise. This paper presents our methodology, experiments, and the results of our approach.

II. Methods

A. Feature extraction

We leverage the Fourier transform to convert time-based audio data into frequency-based data, capitalizing on the unique timbre and tone characteristics that differentiate individual voices. These characteristics are influenced by the mixture of frequencies present in each voice.

Our research utilizes features generated by the Librosa library, notably the Mel Spectrogram (mel_spec), Mel-Frequency Cepstral Coefficients (MFCCs), and Zero Crossing Rate (ZCR) since they are the most important part to classifying audio [1]. The Mel Spectrogram is crucial for voice recognition as it represents the energy distribution across Mel-frequency bins, aligning with the non-linear nature of human hearing. This feature is particularly adept at capturing elements relevant to the human voice, focusing on the 85 to 255 Hz frequency range and employing a logarithmic scale to mitigate the impact of high-frequency noise. The Mel Spectrogram manifests itself as a 2-D array.

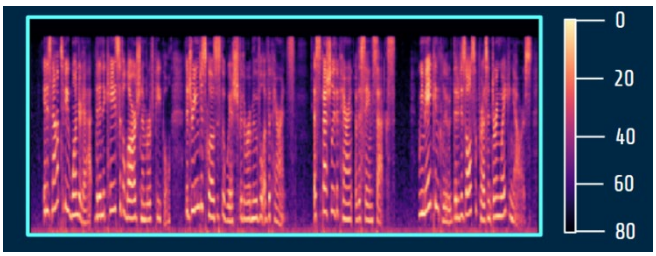

Fig. 1. The visulization of a mel-spectrogram. (*The X axis represents time, the Y axis represents frequency. Darks parts in the graph indicates no one is speaking.*)

MFCCs, another significant feature, are essential in audio signal processing and speech recognition. While they share similarities with the Mel Spectrogram, MFCCs undergo an additional Fourier transform. This process shifts the data onto a 'quefrency' basis, creating a cepstral representation. In practice, we typically extract the first 13 coefficients to capture the most informative voice features, as these represent the spectral characteristics of the audio signal, particularly those pertinent to human perception. MFCCs also form 2-D arrays.

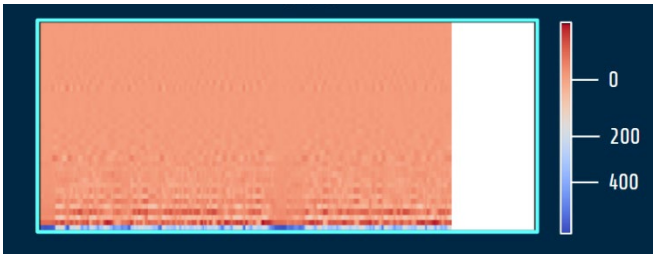

Fig. 2. The visulization of a MFCCs. (*The X axis represents time, the Y axis represents MFCCs. The white part is due to a cropping action.*)

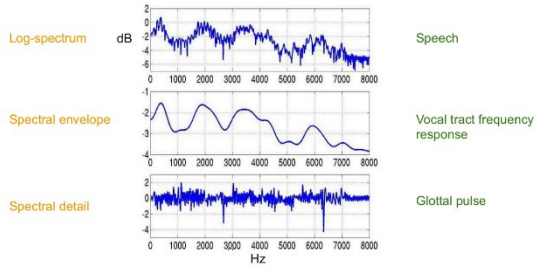

Fig. 3. The contents of an example cepstrum. (*The log-spectrum is similar to what is done in mel-spectrogram. The goal for MFCCs is to collect features in the spectral envelope. Every peak represents different vocal features and avoids glottal pulse, which are uninterested details.*)

Lastly, we use the Zero Crossing Rate (ZCR), a measure of the rate at which the audio signal crosses zero amplitude. ZCR is instrumental in differentiating between speech and non-speech segments of an audio file by identifying rapid signal fluctuations.

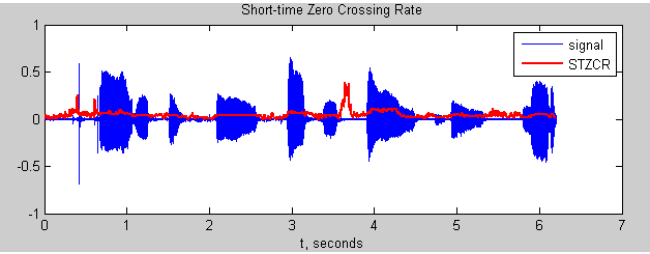

Fig. 4. The visualization of a ZCR

B. Testing data

For our model, which is designed to recognize individual voices, we initially utilized our own recorded audio files as the primary training dataset. However, to compensate for the limited volume of these recordings, we supplemented them with an external dataset from 'http://www.openslr.org/12/'. This additional dataset is not only larger but also of higher clarity, making it ideal for training purposes. By integrating our recordings with this expansive dataset, we aim to both improve our model's performance and ensure a comprehensive evaluation process.

C. Data preprocessing

Our dataset comprises speech samples from 50 distinct speakers, each contributing 10 unique speaking recordings. This approach not only provides a rich variety of vocal characteristics but also aids in creating a balanced dataset, crucial for effective model training.

To standardize the dataset, we processed each speech recording to maintain a consistent duration of 600 seconds. This normalization is vital to ensure uniform array sizes across all data samples, thereby facilitating smoother processing and analysis by our model. The uniform length also aids in mitigating any bias that might arise from varying durations of speech samples [2].

D. Data Augmentation

In addition to standardizing the length of the recordings, we employed data augmentation techniques to enhance the robustness of our model against real-world variables. This augmentation included introducing various types of background noise to the dataset, which simulates real-life scenarios where ambient sounds are present. By training our model with these noise-augmented data, we aim to improve its performance in diverse and potentially noisy environments [3].

We also implemented random shifting of the audio data. This process involves slightly altering the time alignment of the speech within the recordings, thereby creating variations that the model might encounter in practical applications. Such alterations help in training the model to be resilient to timing discrepancies in speech [3].

Lastly, we applied value augmentation to the dataset. This technique involves subtly modifying the amplitude and frequency characteristics of the audio samples. By doing so, we introduce a range of variabilities in the speech signals, further challenging the model to recognize voices under different acoustic conditions [3].

Through these comprehensive preprocessing steps, we ensured that our dataset is not only uniform and balanced but also enriched with variations that mimic real-world challenges. This preparation is crucial for developing a speech recognition model that is both accurate and versatile in diverse conditions.

E. Model training

We harnessed the power of TensorFlow's deep learning capabilities to construct and train our speech recognition model. This state-of-the-art framework provided the foundation for building a robust and efficient neural network.

One of the key considerations in our model design was the handling of multi-class speaker identification. To address this, we employed one-hot encoding for the labels. This encoding strategy ensured that for each input sample, multiple true values could exist in the label array, effectively indicating the presence of multiple speakers in the audio segment [4]. This approach aligns perfectly with our objective of recognizing and identifying multiple speakers within the same audio stream.

Our dataset presented a unique challenge: varying input shapes due to the different durations of speech recordings. To accommodate this variability, we designed our model with multiple input shapes in mind. We achieved this by creating separate input layers for different input shapes and then concatenating the resulting tensors from these input layers. This innovative approach allowed our model to seamlessly handle input data of varying lengths, ensuring that it could process audio segments of different durations effectively.

The model training process involved optimizing various hyperparameters, such as the network architecture, batch size, and learning rate, to achieve the best possible performance [5]. We employed appropriate loss functions and evaluation metrics tailored to our multi-class speaker identification task.

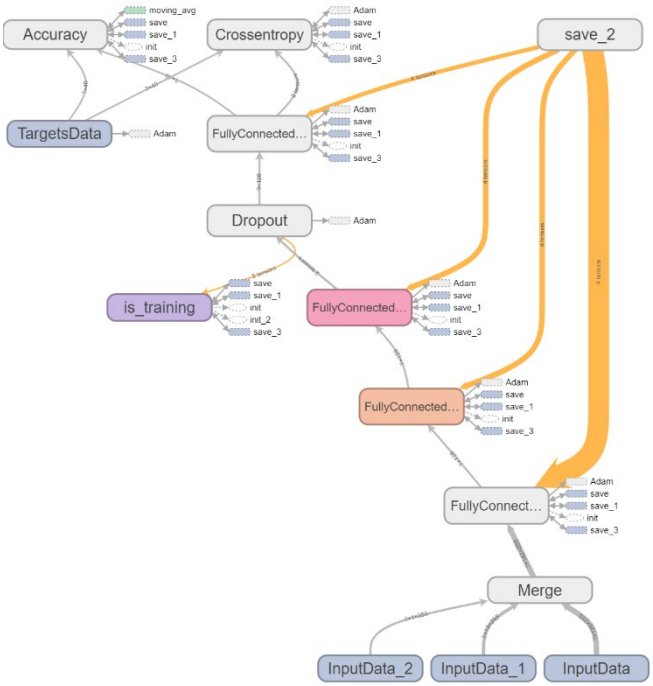

Fig. 5. Model topology, save_2 is the model name.

III. Results

After training our model on the enhanced dataset, which included background noise and time-shifted audio, we used recordings from a basic device, introducing slight noise, to further train the model. Subsequent evaluations were conducted to gauge its effectiveness in multi-class speaker identification. We used accuracy as the key metric to assess the model's performance comprehensively.

The model achieved an impressive 92% accuracy on the test dataset, effectively identifying multiple speakers in an audio stream. This performance is particularly relevant for real-world scenarios, where ambient noise can greatly influence the efficacy of speech recognition systems.

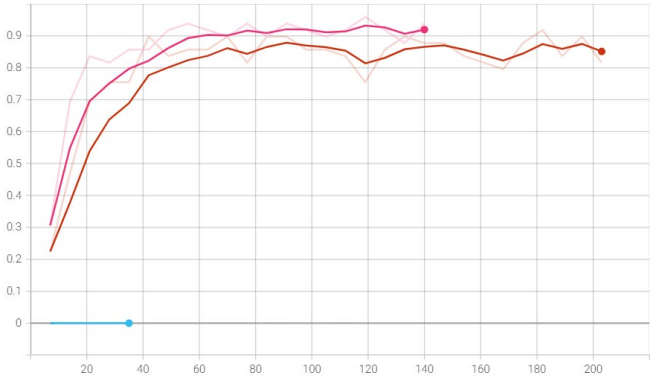

Fig. 6. Accuracy on validation while training.

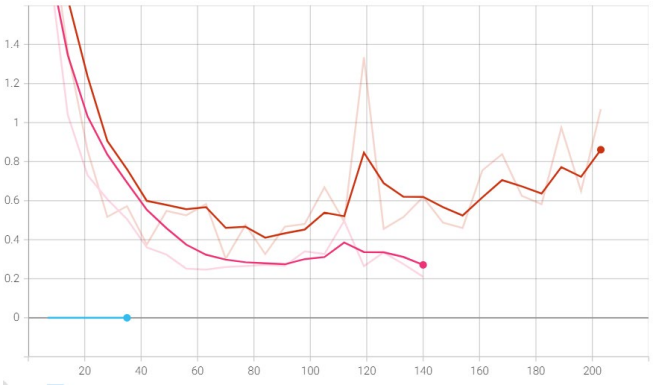

Fig. 7. Loss on validation while training.

IV. Conclusion

Our initial foray into multiclass speech identification demonstrates potential, yet it underscores our novice status in this complex field. The approach, utilizing basic deep neural network techniques and audio signal processing, offers insights but also reveals substantial areas for enhancement. Acknowledging the limitations of our dataset and the simplicity of our model, our study serves more as a learning experience rather than a contribution to the field. Future work would benefit from more advanced methodologies, diverse datasets, and collaboration with experts, aiming towards more sophisticated and accurate multi-speaker recognition systems.

Author Contribution Statements

- Lean Jeng Wen Joshua (16.7%): Leader, model architecture, model testing, project finalizing.
- Aurick Daniel Franciskus S. (16.7%): Model architecture.
- WANG DE FU (16.7%): Data interpretation, data analysis, learning data features.
- KAI-YU CHENG (16.7%): Data collection, data analysis.
- RAY XIANG SU (16.7%): Provide opinion on topic, data collection, data interpretation.
- CHUN YAO TSENG (16.7%): Data collection, data interpretation, data analysis.

Data and Code Availability

- Code is available at https://github.com/Joshimello/SpeechIdentification
- Dataset used for testing is available at http://www.openslr.org/12/

References

[1] M. Turab, T. Kumar, M. Bendechache, and T. Saber, "Investigating Multi-Feature Selection and Ensembling for Audio Classification," arXiv preprint arXiv:2206.07511, June 2022. DOI: 10.48550/arXiv.2206.07511. License: CC BY 4.0. Available: https://arxiv.org/pdf/2206.07511.pdfJ. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] Butko, T.; Nadeu, C. Audio segmentation of broadcast news in the Albayzin-2010 evaluation: Overview, results, and discussion. EURASIP J. Audio Speech Music Process. 2011, 2011, 1.

[3] Salamon, Justin, and Juan Pablo Bello. "Deep convolutional neural networks and data augmentation for environmental sound classification." IEEE Signal Processing Letters 24.3 (2017): 279-283.

[4] J . T. Hancock and T. M. Khoshgoftaar, "Survey on Categorical Data for Neural Networks," Journal of Big Data, vol. 7, no. 1, pp. 1-41, April 2020. DOI: 10.1186/s40537-020-00305-w.

[5] L. Yang and A. Shami, "On hyperparameter optimization of machine learning algorithms: Theory and practice," Neurocomputing, vol. 415, pp. 295-316, Nov. 20, 2020. Received: Dec. 13, 2019; Revised: May 14, 2020; Accepted: Jul. 16, 2020. Available online: Jul. 25, 2020. doi: 10.1016/j.neucom.2020.07.061.