Deep Learning for Speaker Identification: Architectural Insights from AB-1 Corpus Analysis and Performance Evaluation

Matthias Bartolo[1]

[1]University of Malta

matthias.bartolo.21@um.edu.mt

Abstract

In the fields of security systems, forensic investigations, and personalized services, the importance of speech as a fundamental human input outweighs text-based interactions. This research delves deeply into the complex field of Speaker Identification (SID), examining its essential components and emphasising Mel Spectrogram and Mel Frequency Cepstral Coefficients (MFCC) for feature extraction. Moreover, this study evaluates six slightly distinct model architectures using extensive analysis to evaluate their performance, with hyperparameter tuning applied to the best-performing model. This work performs a linguistic analysis to verify accent and gender accuracy, in addition to bias evaluation within the AB-1 Corpus dataset.

Index Terms: speaker identification, deep learning, feature extraction, classification problem

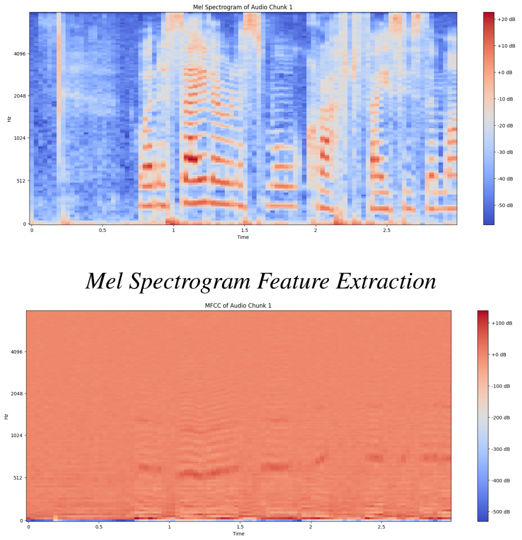
1. Introduction

Speaker identification (SID) is the task of determining a speaker's identity from a specific audio sample chosen from a pool of known speakers. With applications in forensics, security, and customization [1], SID may be expressed as a pattern recognition problem. The SID pipeline, according to [2], is dependent on two critical components: *feature extraction* and *feature classification*. These factors work together to classify an input speech segment as belonging to one of N known enrolled speakers.

In feature extraction, certain characteristics required for an individual's identification are taken from the voice data. During the feature classification process, features extracted from an unidentified individual are compared to those of various speakers in order to detect and match the unique qualities that will eventually lead to the identification of the specific speaker [1]. A perfect speaker identification system should have minimal intra- and inter-speaker variance, readily quantifiable attributes, be less susceptible to noise, respond correctly to imitation, and not rely on other qualities [2].

2. Feature Extraction

Feature extraction is a critical step in speech analysis, stemming from the fact that it allows raw audio data to be transformed into useful features. To this extent, *Mel Spectrogram* and *Mel Frequency Cepstral Coefficients (MFCC)* are two widely used techniques in this field of study [1]. By translating frequencies into the Mel scale, the Mel Spectrogram visualises the frequency content of an audio source across time, emphasising human auditory perception. Meanwhile, MFCCs extract compact representations by capturing the spectral features of the audio signal [2]. In this study, both of the aforementioned feature extractors were tested as inputs to the developed system architectures, with the goal of producing a thorough analysis to evaluate their performance and appropriateness for the current context. Figure 1 depicts the output of both extractors.



Mel Spectrogram Feature Extraction

MFCC Feature Extraction

Figure 1: *Comparison of different audio feature extraction methods for the same filtered three-second audio chunk.*

3. Model Architecture

This study investigated a total of six slightly distinct model architectures. The first model architecture, as shown in Table 1 incorporates elements from both architectures listed in [3] and [4]. The model incorporates early feature extraction convolution layers modified from the structure suggested in [3], followed by an altered single lightweight LSTM layer suggested by [4]. The addition of batch normalization and dropout layers was intended to regularize the flattened LSTM output, hence improving the model's robustness and avoiding overfitting. Finally, classification was performed by activating softmax in the final fully connected dense layer. This function generated probability scores for each of the 285 speakers in the AB-1 corpus dataset, indicating the likelihood that the input data corresponds to one of the aforementioned speakers.

The succeeding model architectures in this research are all derived from the above architecture. Furthermore, the second architecture focused on raising the convolutional depth in the first feature extractors (CNNs) to improve feature extraction capabilities. Meanwhile, in the third model architecture, the

No.	Layer Type	Details
1	Conv2D	(3, 3), 32 filters
2	ReLU	–
3	Conv2D	(3, 3), 64 filters
4	ReLU	–
5	MaxPooling2D	(2, 2)
6	Conv2D	(3, 3), 64 filters
7	ReLU	–
8	MaxPooling2D	(2, 2)
9	Reshape	–
10	LSTM	64 units
11	Flatten	–
12	BatchNormalization	–
13	Dropout	30%
14	Dense	285 neurons
15	Softmax	285 outputs

Table 1: *Model 1 Architecture*

LSTM was modified by increasing the number of LSTM units and incorporating an extra layer to increase sequence comprehension. The fourth architecture included an extra dense layer following the CNN-LSTM configuration to further enhance the output sequence. In contrast to the second model architecture, the fifth architecture was intended to reduce model complexity by recommending fewer convolution filters. Finally, the sixth architecture used batch normalisation across all CNN blocks in the model with the aim of reducing the effects of overfitting.

4. Evaluation

The aforementioned model architectures were trained with the *TensorFlow Keras framework*, which was chosen for its ease of use, efficacy in creating neural network models, and facilitation of efficient training methods. During training, the Adam optimizer, which is noted for its versatility and extensive use, was used [5]. Additionally, an early stopping callback with a patience score of 5 was also included to prevent overfitting.

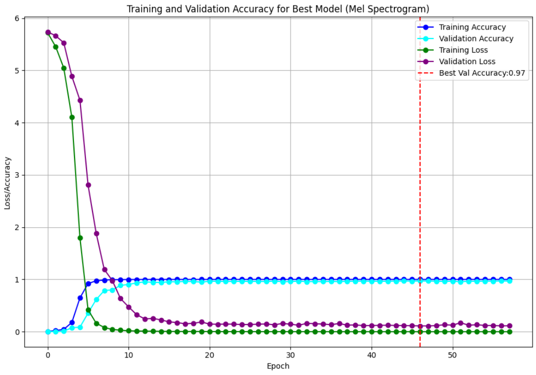


Figure 2: *Curves demonstrating the best model's training and validation loss, as well as accuracy.*

The findings shown in Table 2 reveal that models 1 and 5 outperformed all others in terms of test accuracy, precision, recall, and f-score. Overall, all the presented architectures worked relatively well, with the exception of model 2 using the MFCC feature extractor, which consistently yielded comparable poor results over several runs. Furthermore, the results also illustrate that models using the Mel Spectrogram feature extractor outperformed those using the MFCC feature extractor. Additionally, the models consistently produced test accuracy, precision, recall, and f-score metrics in the 0.8 to 0.97 range, indicating a noteworthy degree of performance. Furthermore, the test loss observed among the models ranged between 0.14 and 0.65, indicating that the models were not overfitting.

Following the discovery of model 1's superior efficiency, a thorough hyperparameter tuning procedure comprising fifteen trials was carried out to fine-tune its parameters. From the test, the best hyperparameters found were a learning rate of **0.001**, a dropout rate of **0.4**, and the activation functions **tanh** applied specifically to the second layer in the model architecture presented in Table 1, whilst the remaining layers utilised **relu**. The findings shown in Table 2 demonstrate a significant performance improvement attained by the best model refined with these optimised parameters over the original model 1.

The visible improvements in the curves in Figure 2, which reflect the training and validation loss as well as accuracy for the best model, illustrate the efficiency of the adjusted parameters in optimising the model's performance. Furthermore, the curves exhibit a consistent and continuous evolution, demonstrating the model's incremental improvement during the training period, which is characterised by a decreasing loss and an increasing accuracy.

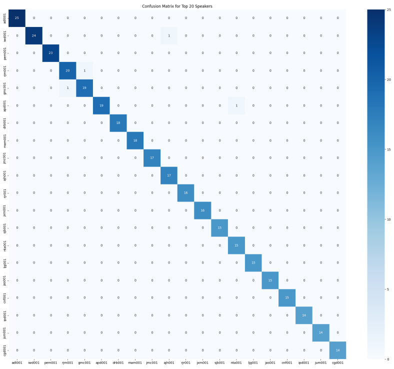


Figure 3: *Confusion matrix showcasing the top 20 projected speakers for the best model.*

The confusion matrix for the top 20 projected speakers in Figure 3 exhibits distinct diagonal components. Furthermore, these diagonal components, which particularly demonstrate the model's capacity to precisely anticipate speakers, are important indicators of the classification performance of the developed model.

5. Analysis

A linguistic analysis was performed on the top-performing model to assess its gender and accent correctness while assuring the model's neutrality and lack of bias towards certain accents or genders. Figure 4 demonstrated a slight variance in gender accuracy. Surprisingly, the model performed somewhat better in predicting female speakers than male speakers, by a margin of 0.02. Furthermore, these findings imply that the model's predictions exhibit gender equality, indicating that there is equal

Model	Best Val Accuracy	Test Accuracy	Test Loss	Precision	Recall	F1-Score	Epoch Converged	Time Taken (minutes)
Model-1-Mel	**0.97**	**0.968**	0.148	**0.971**	**0.968**	**0.968**	22	30
Model-1-MFCC	0.93	0.928	0.301	0.937	0.928	0.928	**13**	20
Model-2-Mel	0.01	0.009	5.619	0.000	0.009	0.000	19	58
Model-2-MFCC	0.95	0.928	0.301	0.937	0.928	0.928	25	78
Model-3-Mel	0.93	0.934	0.262	0.947	0.934	0.934	**13**	18
Model-3-MFCC	0.91	0.917	0.343	0.930	0.917	0.918	21	42
Model-4-Mel	0.96	0.952	0.252	0.958	0.952	0.952	24	28
Model-4-MFCC	0.86	0.841	0.640	0.862	0.841	0.838	25	31
Model-5-Mel	**0.97**	0.965	**0.141**	0.968	0.965	0.964	21	13
Model-5-MFCC	0.93	0.920	0.313	0.929	0.920	0.919	19	**12**
Model-6-Mel	0.96	0.965	**0.141**	0.968	0.965	0.964	19	28
Model-6-MFCC	0.91	0.920	0.313	0.929	0.920	0.919	25	38
Best Model	**0.97**	**0.971**	**0.136**	**0.974**	**0.971**	**0.971**	**N/A**	**N/A**

Table 2: *Table of Results*

representation within the AB-1 corpus dataset. This equality in predicted accuracy indicates that there is no bias towards any one gender, emphasizing a fair and impartial training dataset.

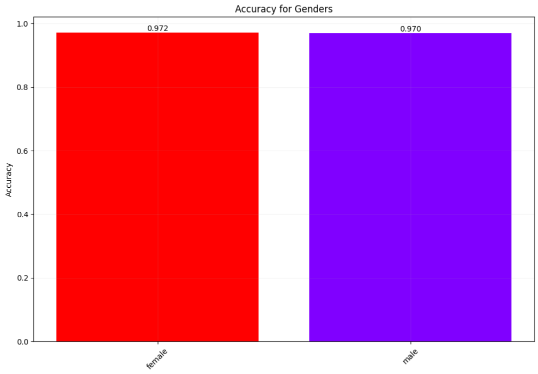


Figure 4: *Gender accuracy and bias evaluation for the best model's performance on the test set.*

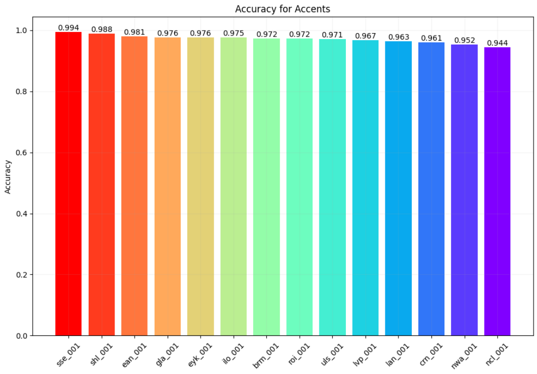


Figure 5: *Accent accuracy and bias evaluation for the best model's performance on the test set.*

As observed in Figure 5, the accent accuracy scores were quite similar across accents. However, when compared to the gender accuracy distribution, a more obvious disparity between the best and lowest accuracy, approximately 0.05, emerged. From Figure 5 it can be deduced that the three easiest accents to predict (in order) were the following:

1. Standard Southern English
2. Scottish Highlands
3. East Anglia

In contrast, the three hardest accents to predict (in order) were the following:

1. Newcastle
2. Northern Wales
3. Cornwall

6. Conclusion

This paper investigates the efficacy of Mel Spectrogram and MFCC as feature extraction approaches for speaker identification (SID) while proposing robust model architectures designed particularly for SID. The study's best model accuracy, precision, recall, and F-score of 0.97 demonstrate the usefulness of these techniques. Gender accuracy analysis revealed minimal variation, with slightly greater accuracy in identifying female speakers, confirming dataset balance and the lack of gender bias in predictions. Nonetheless, accent accuracy resulted in a more noticeable discrepancy, with Standard Southern English being the most predicted and Newcastle being the least predictable. The results of this study highlight the significance of future analysis and model development, notably in the case of accent-related challenges in SID.

7. References

[1] S. Sremath Tirumala and S. R. Shahamiri, "A review on deep learning approaches in speaker identification," 11 2016, pp. 142–147.

[2] A. Antony and R. Gopikakumari, "Speaker identification based on combination of mfcc and umrt based features," *Procedia Computer Science*, vol. 143, pp. 250–257, 2018, 8th International Conference on Advances in Computing & Communications (ICACC-2018). [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050918320908

[3] Y. Lukic, C. Vogt, O. Dürr, and T. Stadelmann, "Speaker identification and clustering using convolutional neural networks," in *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, 2016, pp. 1–6.

[4] Z. Zhao, H. Duan, G. Min, Y. Wu, Z. Huang, X. Zhuang, H. Xi, and M. Fu, "A lighten cnn-lstm model for speaker verification on embedded devices," *Future Generation Computer Systems*, vol. 100, pp. 751–758, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167739X18330620

[5] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017.