Speaker Recognition in Door Access Control System

Enrico Manfron[1,2][0009−0002−6107−0857], João Paulo Teixeira[2][0000−0002−6679−5702], and Rodrigo Minetto[1][0000−0003−2277−4632]

[1] Universidade Tecnológica Federal do Paraná, Brasil
enricomanfron@alunos.utfpr.edu.br, rminetto@ufpr.edu.br
[2] CeDRI, Instituto Politécnico de Bragança, Portugal
joaopt@ipb.pt@ipb.pt

Abstract. In this paper, we explore the potential of speaker recognition technology as a biometric authentication method for access control systems. We focus on the development and evaluation of two machine learning models, the Gaussian Mixture Model (GMM) and Multilayer Perceptron (MLP), for speaker identification. Our research presents a review of speaker recognition literature, followed by a detailed methodology for constructing and training the GMM and MLP models on a specific dataset. Experimental results highlight the performance of these models in terms of accuracy and efficiency. This study contributes to the application of GMM and MLP models for speaker recognition-based access control systems, serving as a resource for future research and development in secure and effective access control solutions.

Keywords: Speaker recognition · Gaussian Mixture Model (GMM) · MLP (Multilayer Perceptron).

1 Introduction

Speaker recognition is a technology with potential applications in biometric authentication, which can be employed in access control systems for secure environments. In recent years, progress has been made in speaker recognition technology, including the development of sophisticated algorithms and machine learning models. By analyzing the unique characteristics of an individual's voice, this technology can accurately identify people and grant them access to restricted areas.

This paper presents the initial work on the development and evaluation of two speaker recognition models: the Gaussian Mixture Model (GMM) and the Multilayer Perceptron (MLP). We begin with a review of the literature on speaker recognition, followed by a detailed description of the methodology employed to develop the GMM and MLP models, as well as the dataset used for training and testing. Finally, we present and analyze the experimental results obtained from the evaluation of both models.

Our objective in this paper is to present our work on speaker recognition-based access control systems by exploring the application of GMM and MLP models. We believe that this work can serve as a resource for future research in this area, as well as aid in the development of more secure and efficient access control systems.

2 Related Work

In recent years, the field of speaker recognition (SR) has seen significant advancements, driven by the development of novel techniques and the growing availability of large datasets. As a result, a significant number of research articles have been published to explore various aspects of SR, from fundamental concepts and methodologies to the latest state-of-the-art models.

Among these early approaches, Gaussian Mixture Models (GMM) emerged as a popular and powerful technique in the pre-deep learning era of SR, from 1995 to 2006. The GMMs models were applied in numerous applications in computer vision, speech recognition, and speaker recognition, thanks to their ability to approximate complex distributions using a combination of simple Gaussian distributions [9]. Deep learning has been the dominant machine learning approach since around 2010 [3].

The concept shares similarities with Gaussian mixture models, using simple functions called neurons to approximate complex functions. In speech signal processing, recurrent neural networks have been particularly useful due to their ability to model sequence data effectively [4]. Deep learning models are more scalable and efficient when handling large datasets, with specialized hardware like GPUs and TPUs available for acceleration. From 2014 onwards, the field of speaker recognition has seen numerous advancements in deep learning models.

Hanifa et al. [3] provides a comprehensive survey of SR models, addressing major issues such as background noise, lack of data, and attacks on models. It presents a chronology of the field's development, highlighting the technologies created and the progress made. Researchers have explored various preprocessing techniques, common features extracted in the field, potential model types and classifiers, and application areas.

Examples include a 2010 study that employed 3 Discrete Wavelet Transform (DWT) with different coefficients, using a Multilayer Perceptron (MLP) and a Gaussian Mixture Model (GMM) as classifiers [6]. Both models achieved high accuracy (98% and 99%), but the MLP could be trained with audio samples half the duration of those used in the GMM model. In subsequent years, research utilized neural networks, such as the Fuzzy Min-Max Neural Network (FMMNN) [5], as well as variations of the GMM model [7], and comparisons with Hidden Markov Model (HMM), all using Mel Frequency Cepstral Coefficients (MFCC) vectors as input [11].

Later attempts explored variations of MFCC features, such as Normalized Dynamic Spectral Features (NDSFs) and Linear Prediction Cepstral Coefficients (LPCCs), to determine if these sets could provide better feature representation

than MFCCs [1]. The field then shifted to using convolutional neural networks [2] and x-vectors for their robustness to noise [12]. Recent studies in the survey employed combinations of the aforementioned features, such as MFCCs + PNCC [8] and LDA + MFCCs [13], among others.

3 Methodology and Results

In this research, we developed a methodology to explore and understand Speaker Recognition (SR) techniques through the implementation of early SR approaches, including the Gaussian Mixture Model (GMM), GMM-UBM, and the Multilayer Perceptron (MLP) model. These implementations allowed us to compare different approaches in the SR domain and develop a deeper understanding of the underlying concepts and challenges.

For the GMM implementation, we began by extracting the first 20 MFCCs from each audio file, consisting of 32 speakers. This resulted in a matrix $m \times n$, where m is related to the audio duration and n is equal to 20, as we are using the first 20 MFCCs.

We created 32 GMM models, one for each speaker, and extracted the 20 MFCCs from each audio file. We concatenated the tables with the MFCCs for each speaker and divided the data into 80% for training and 20% for testing. A GMM model with 32 Gaussians was trained for each speaker. After training, we used a function that calculates the average log-likelihood per sample of the provided data. To identify a speaker, we took a speech sample and compared it to all GMM models, selecting the model with the highest score as the most probable representation of the speaker.

In this first approach, the model correctly recognized all speakers in the test set; however, since the GMM model assumes a closed set, it attempts to identify the speaker from the set by assigning a score to the most likely candidate. This makes it impossible to determine the presence of an imposter that is out of the set. Following this method, when processing an imposter's speech, we would have a score for each model, and the imposter would be classified with the model with the lowest score, incorrectly assigning the imposter instead of rejecting them.

This initial GMM model is a Speaker Identification model, where the task is to identify which speaker said a given phrase. However, we are now interested in verifying if a speech came from a specific speaker, which is called Speaker Verification.

To perform Speaker Verification (i.e., to determine if the speaker is who they claim to be), another approach called the Universal Background Model (UBM) is necessary [10]. This approach uses a speaker-independent GMM model to represent alternative speakers or imposters. We then reduce the problem to test two hypotheses:

H_0: The utterance is from the hypothetical speaker S.

H_1: The utterance is not from the hypothetical speaker S.

To calculate the model, we use the log-likelihood ratio, defined as:

$$L(X) = \ln p(X|H_0) - \ln p(X|H_1) \tag{1}$$

Where X is the feature vector extracted from the speech utterance.If $L(X) \geq 0$, we accept Hypothesis H_0; otherwise, we accept Hypothesis H_1.

To implement the above-described approach, two methods were applied. In the first method, for a single speaker S, a GMM model containing only S's training data and a GMM-UBM model containing data from all speakers excluding S were trained. In total, there were 32 GMM models and 32 GMM-UBM models for each speaker, resulting in two GMM models for each speaker. In this approach, the model recognized 30 out of the 32 speakers, however, the 2 unrecognized speakers' scores were really near zero.

We test a new approach based on the proposal of Reynolds [10], where the idea for creating the GMM-UBM is slightly different. First, we used all the training data to create the GMM-UBM. Then, using an algorithm called Bayesian Adaptation, the UBM model was adapted for each speaker's specific data. In the end, there were 32 GMMs, one for each speaker, and a single UBM. Although it is possible to re-adapt the weights, means, and covariances, the best approach is to adapt only the means. The training pipeline is illustrated in Figure 1.

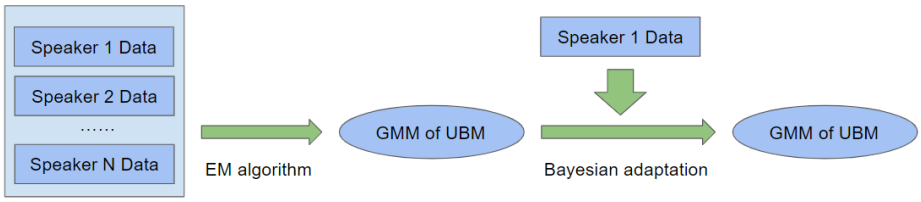

Fig. 1. GMM-UBM training pipeline using Bayesian Adaptation algorithm.

After training, there was a slight improvement in the number of correct identifications, now the model accepts 31 out of 32 speakers. However, the system using this approach resulted in more false positives.

Based on the work done by Kral et al. [6], our study focuses on implementing a Multilayer Perceptron (MLP) model for speaker recognition. We build upon the existing research that demonstrates the effectiveness of MFCCs in speaker recognition tasks.

For the initial testing phase, Model A, a simple Multilayer Perceptron (MLP) model was created to classify 32 speakers. The architecture of all MLP models is shown in Table 1. The model took the first 13 MFCCs, delta MFCCs, and delta-delta MFCCs as input features for a descriptor size of 39 features. The MLP architecture consisted of three hidden layers with 256, 128, and 64 neurons, respectively. The audio features were efficiently extracted and saved in a CSV file, allowing for easier feature selection. The initial tests achieved an accuracy

of 74% using CrossEntropy as the loss function, training for 50 epochs, in Figure 2 it is shown the loss function during the training.

Table 1. MPL architecture.

Layer (type)	Input Shape	Output Shape
Linear-1	N	256
Linear-3	256	128
Linear-5	128	64
Linear-7	64	32

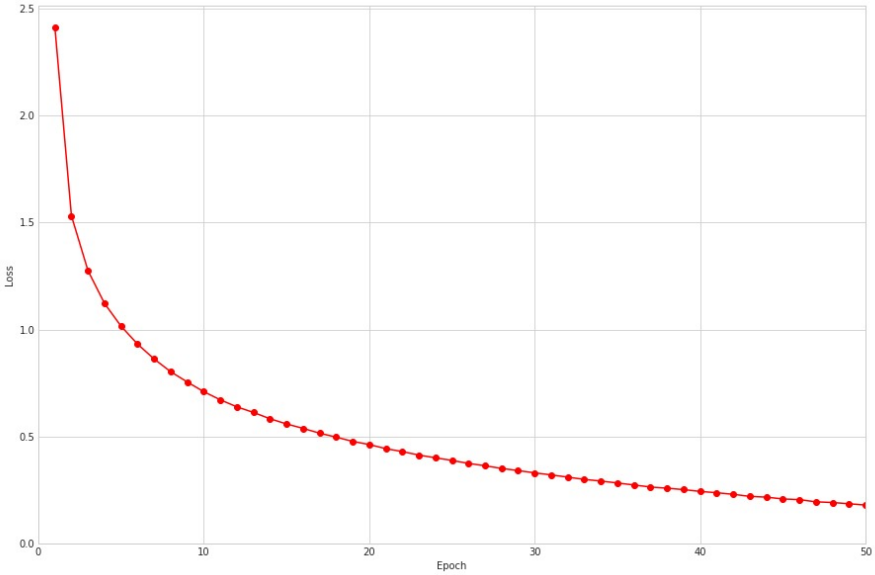

Fig. 2. Train Loss function for MLP model for 50 epochs.

Subsequently, the model could be trained for mode epochs and an early stopping mechanism was implemented in order to optimize time training. This attempt, which utilized the same MLP architecture for classification, will be referred to as Model B for future reference. Model B was used for classification with 70% of the data allocated for training, 10% for validation (early stopping), and 20% for testing. The early stopping criteria were initially based on the loss function but later switched to accuracy, as it was observed that accuracy could still improve even when the loss increased.

The highest accuracy achieved with this model was 91.12% on the test data. It was hypothesized that the increasing loss with improved accuracy might be due to class imbalance caused by varying audio lengths among speakers, which results in different numbers of MFCCs vectors extracted per speaker.

Additional features were incorporated into the model, which we will refer to as Model C, such as chroma stft, rmse, spectral centroid, spectral bandwidth, spectral roll-off, zero crossing rate, and more MFCCs (20 columns) for a total descriptor with size of 77 features. The input size increased, and the model was trained over 500 epochs. This resulted in a validation accuracy of 93.48% and a test accuracy of 92.19%. The training became more stable with the increased number of features, although the loss still increased.

Finally, the learning rate was reduced from 0.001 to 0.0001, leading to more stable training and a further improvement in accuracy. The model, which we will refer to as Model D, achieved 93.4% validation accuracy and 93.33% test accuracy. The loss and accuracy during the training could be seen in Figure 3 and Figure 4.

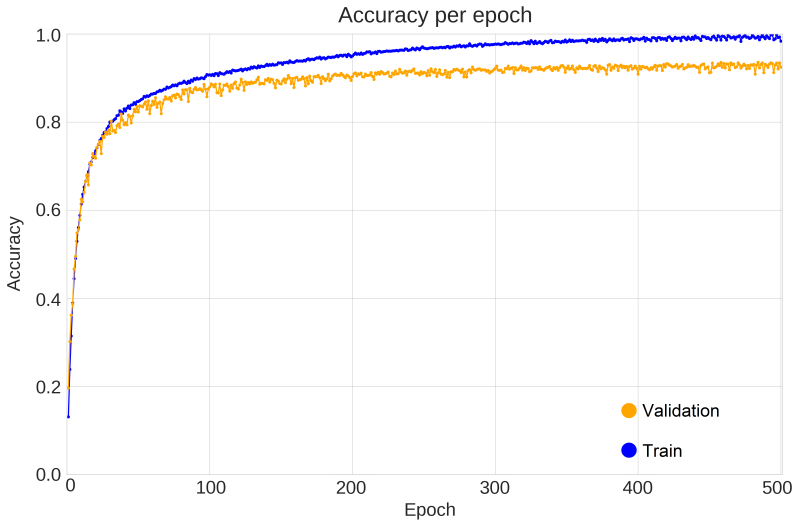

Fig. 3. Train Accuracy function for MLP model for 500 epochs.

A comparison of the MLP network's results can be found in Table 2.

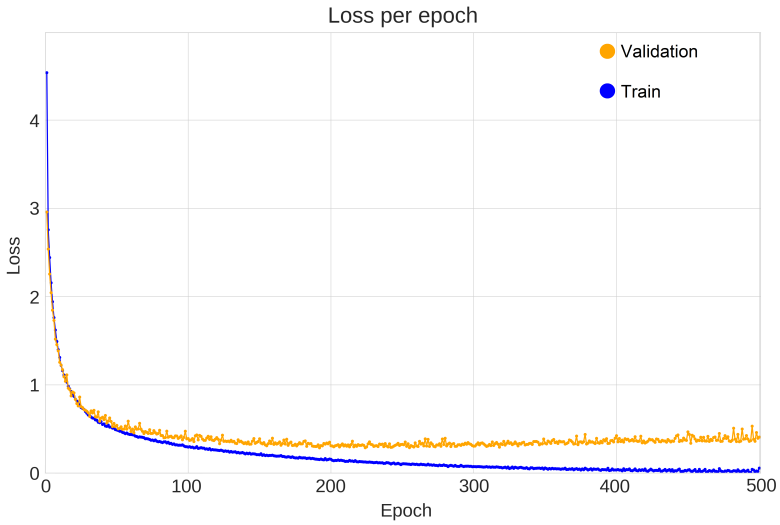

Fig. 4. Train Loss function for MLP model for 500 epochs.

Table 2. Performance of MLP Model Across Attempts

model	Input Features	Validation Accuracy	Test Accuracy
Model A	MFCC(13), ΔMFCC(13), Δ^2MFCC(13)	–	0.74
Model B	MFCC(13), ΔMFCC(13), Δ^2MFCC(13)	0.9176	0.9112
Model C	MFCC(20), ΔMFCC(20), Δ^2MFCC(20), chroma stft(12), rmse(1), spectral centroid(1), spectral bandwidth(1), spectral rolloff(1), zero crossing rate(1)	0.9348	0.9219
Model D	MFCC(20), ΔMFCC(20), Δ^2MFCC20, chroma stft12, rmse1, spectral centroid1, spectral bandwidth1, spectral rolloff1, zero crossing rate1	0.934	0.9333

4 Conclusion

This study demonstrates the effectiveness of a Multi-Layer Perceptron (MLP) model for speaker identification using various audio features. The integration of additional features has all contributed to a more robust and accurate model.

The results show that an MLP model can achieve high accuracy in speaker classification, with the best model reaching 93.33% on test data. These findings highlight the importance of using a diverse set of features and optimizing hyperparameters, such as the learning rate, to achieve better performance. By decreasing the learning rate value, the model showed more stable behavior, enabling a slow and steady adjustment of the neural network weights throughout training. This resulted in a better convergence of the training process. In the end, these modifications allowed the model to learn better and find a superior solution, which improved its performance on the given task.

Despite the promising results, there are still some challenges that need to be addressed. The increase in the loss during training, probably due to class imbalance, warrants further investigation to improve the model's stability and generalization. Future work could explore techniques such as data augmentation, resampling, or other mechanisms to handle class imbalance more effectively.

Additionally, other architectures, such as Transformer Networks Convolutional Neural Networks (CNNs), and Recurrent Neural Networks (RNNs), could be considered for speaker classification tasks to evaluate their performance in comparison to the MLP model.

In summary, this paper presents a foundation for further research and development in speaker classification using neural networks. The results achieved by the MLP model in this study demonstrate its potential for real-world applications, while the challenges identified provide valuable insights for future improvements.

References

1. Chougule, S.V., Chavan, M.S.: Robust spectral features for automatic speaker recognition in mismatch condition. Procedia Computer Science **58**, 272–279 (2015)
2. Chung, J.S., Nagrani, A., Zisserman, A.: Voxceleb2: Deep speaker recognition (2018)
3. Hanifa, R.M., Isa, K., Mohamad, S.: A review on speaker recognition: Technology and challenges. Computers &amp Electrical Engineering **90**, 107005 (Mar 2021)
4. Hori, T., Cho, J., Watanabe, S.: End-to-end speech recognition with word-based rnn language models. In: 2018 IEEE Spoken Language Technology Workshop (SLT). IEEE (Dec 2018)
5. Jawarkar, N.P., Holambe, R.S., Basu, T.K.: Use of fuzzy min-max neural network for speaker identification. In: 2011 International Conference on Recent Trends in Information Technology (ICRTIT). IEEE (Jun 2011)
6. Kral, P.: Discrete wavelet transform for automatic speaker recognition. In: 2010 3rd International Congress on Image and Signal Processing. IEEE (Oct 2010)

7. Krishnamoorthy, P., Jayanna, H., Prasanna, S.: Speaker recognition under limited data condition by noise addition. Expert Systems with Applications **38**(10), 13487–13490 (Sep 2011)
8. P, B.K., M, R.K.: ELM speaker identification for limited dataset using multitaper based MFCC and PNCC features with fusion score. Multimedia Tools and Applications **79**(39-40), 28859–28883 (Aug 2020)
9. Reynolds, D., Rose, R.: Robust text-independent speaker identification using gaussian mixture speaker models. IEEE Transactions on Speech and Audio Processing **3**(1), 72–83 (1995)
10. Reynolds, D.A., Quatieri, T.F., Dunn, R.B.: Speaker verification using adapted gaussian mixture models. Digital Signal Processing **10**(1-3), 19–41 (Jan 2000)
11. Tolba, H.: A high-performance text-independent speaker identification of arabic speakers using a CHMM-based approach. Alexandria Engineering Journal **50**(1), 43–47 (Mar 2011)
12. Villalba, J., Chen, N., Snyder, D., Garcia-Romero, D., McCree, A., Sell, G., Borgstrom, J., Richardson, F., Shon, S., Grondin, F., Dehak, R., García-Perera, L.P., Povey, D., Torres-Carrasquillo, P.A., Khudanpur, S., Dehak, N.: State-of-the-art speaker recognition for telephone and video speech: The JHU-MIT submission for NIST SRE18. In: Interspeech 2019. ISCA (Sep 2019)
13. Zergat, K., Selouani, S., Amrouche, A.: Feature selection applied to g.729 synthesized speech for automatic speaker recognition. In: 2018 IEEE 5th International Congress on Information Science and Technology (CiSt). IEEE (Oct 2018)