# Speaker Verification on Small Datasets with ResNet50

Enrico Manfron[1,4], Rodrigo Minetto[4], and João Paulo Teixeira[1,2,3]

[1] Research Centre in Digitalization and Intelligent Robotics (CeDRI), Instituto Politecnico de Bragança, Campus de Santa Apolónia, 5300-253 Bragança, Portugal
[2] Associate Laboratory for Sustainability and Technology (SusTEC), Instituto Politécnico de Bragança, Campus de Santa Apolónia, 5300-253 Bragança, Portugal
[3] Applied Management Research Unit (UNIAG), Instituto Politecnico de Braganca, 5300-253 Braganca, Portugal
[4] Federal University of Technology – Paraná (UTFPR), 80230-901 Curitiba, Brazil
enricomanfron@alunos.utfpr.edu.br, rminetto@utfpr.edu.br, joaopt@ipb.pt

**Abstract.** In this study, we explore the capabilities of speaker recognition technology for biometric authentication, developing speaker recognition-based access control systems, and serving as a resource for future research. We focused on developing and evaluating the ResNet50 model for speaker verification. The model was trained and tested on private datasets with 32 speakers and public datasets with 1251 to 6112 speakers. The model ResNet50 achieved a good result on our private dataset by achieving the best EER of 0.7%.

**Keywords:** Speaker Verification, Speaker Recognition, Siamese Convolutional Neural Networks, VoxCeleb1, VoxCeleb2

## 1 Introduction

The human voice is a powerful communication means of conveying emotions and intentions. Moreover, it can also be used as an identification tool since it is a unique signature loaded with individual characteristics that distinguish one speaker from another [1]. Voices vary among individuals due to several factors. Physiologically, differences in the size and shape of our vocal organs, such as vocal folds and the vocal tract, play a role. In addition, gender, age, language, and accents also contribute to this diversity [2, 3, 4, 5].

Speaker recognition technology can be a biometric authentication method in secure environments, such as access control systems. This technology has seen significant progress recently due to the development of advanced algorithms and machine learning models. By analyzing an individual's distinct vocal characteristics, speaker recognition technology can accurately identify them and grant access to restricted areas.

In this work, we discuss strategies for developing and testing speaker verification techniques. We examine the ResNet50 model discussed in the VoxCeleb2 [6]. The primary goal of this paper is to apply speaker recognition-based systems to small datasets to grant access to CeDRI's laboratory. Our future work will involve deploying the most effective model on a device. We explored the application of different models and compared the results specifically for the speaker verification task, using a small audio dataset of Portuguese speakers, including European, Brazilian, African, and foreign speakers.

In recent years, Speaker Recognition (SR) has made significant progress due to the development of new methods and the availability of large datasets. As a result, many academic papers have been published exploring various aspects of SR, ranging from fundamental concepts and methodologies to the latest cutting-edge models. Deep learning has led to impressive results in speaker recognition [7, 8]. Hanifa et al. [1] provides a comprehensive overview of the field, including evolution, technologies, and advancements. Researchers have explored preprocessing methods, standard features, model types, classifiers [9], and applications. Research on neural networks has included the Fuzzy Min-Max Neural Network (FMMNN) [10] and comparisons with the Hidden Markov Model (HMM) [11].

In 2017, Nagrani, Chung, and Zisserman [12] introduced VoxCeleb, a speaker identification (SI) dataset with hundreds of thousands of utterances from over 1,000 celebrities. They used a Convolutional Neural Network (CNN) for active speaker verification and compared various SI techniques on the dataset to establish a baseline performance. Chung, Nagrani, and Zisserman [6] presented an improved process for gathering a dataset of speech data for speaker identification. The resulting dataset included over a million utterances from more than 6,000 speakers and was multilingual. The study compared different CNN models and training strategies, showing that these models outperformed previous works. In their work, Nagrani et al. [13] addressed the issue of speaker recognition (SR) in noisy and uncontrolled environments. They proposed and compared different convolutional neural network (CNN) architectures, aggregation techniques, and training loss functions to accurately recognize speaker identities in various conditions. Their models, which were trained using the VoxCeleb dataset, significantly improved over previous works in this domain.

The paper begins with a detailed description of the methodologies employed to develop the models and the datasets in Section 2. Section 3 and 4 presents a comparative performance analysis of these models. The paper concludes in Section 5 with an analysis and evaluation of the experimental results obtained from evaluating the models.

## 2 Materials and Methods

### 2.1 Dataset

The ultimate objective of this project is to develop a speaker identification system that grants entry to the CeDRI laboratory. For this purpose, we utilized

the CeDRI dataset as our primary data source. This dataset includes a unique set of recorded speeches from members of the CeDRI community. It consists of 169 spoken utterances recorded by 32 Portuguese-speaking individuals through reading exercises. The length of these recordings ranges from 2.7 seconds to a maximum of 21.5 seconds, as shown in Figure 1.

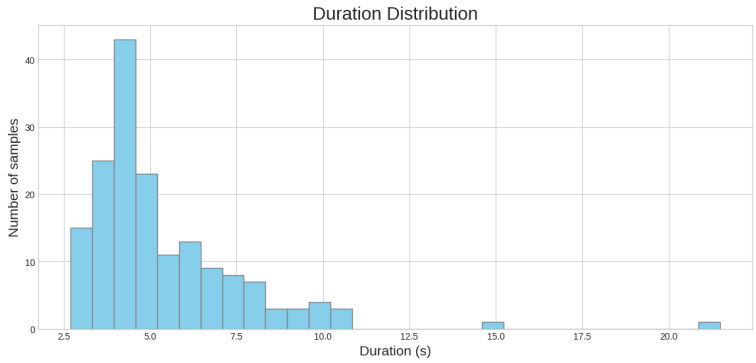


**Fig. 1.** Histogram showcasing the duration distribution of audios.

The speakers included in the dataset were from diverse backgrounds and dialects. Around 69% of the participants are male, whereas 31% are female. The dataset comprised various accents from different Portuguese-speaking regions, such as Portugal, Angola, São Tome, Cabo Verde, Mozambique, Brazil, and Spain. Most speakers were in their early twenties, making the dataset highly representative of young adult speech. See Figure 2 for a detailed overview.

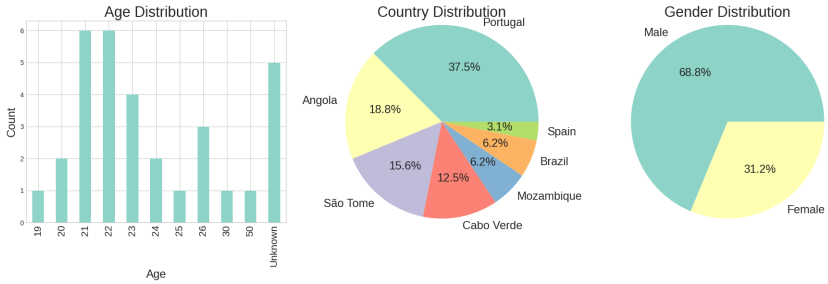


**Fig. 2.** Distribution Analysis (age, country, and gender) within the CeDRI dataset.

We implemented two strategies to standardize our dataset. The first strategy involved using a silence removal filter, which discarded any parts of the audio below a 30db threshold, resulting in audios with a range duration of 1.824s to

10.358s. The second strategy was the application of a trim filter. This filter removes silence from the beginning and end of the audio, using the 30db threshold. As a result, the lengths of the processed audio files ranged from 1.888 seconds to 12.608 seconds.

We standardized the audio files of both strategies by segmenting and resampling them to ensure a uniform duration and 16kHz sampling rate. This created two new datasets: Framed Silence Revomed (FrSR) from the first strategy and Framed Trim (FrT) from the second. For the silence removal strategy, we divided the audio durations into segments of 1.824 seconds, while for the trim strategy, we adjusted the audio durations to 0.944 seconds.

We needed more data than our existing datasets could provide to train our convolutional model. So, we added VoxCeleb1 and VoxCeleb2 datasets, comprising over 1.1 million utterances from more than 7,000 speakers. These datasets include speakers from diverse backgrounds, ethnicities, accents, professions, and age groups and contain real-world noise like background chatter, overlapping speech, laughter, and various room acoustics. The audio files are in a single-channel, 16-bit stream at a 16kHz sampling rate and may include English, German, and French.

### 2.2 Model

The ResNet architecture adapts a standard CNN with skip connections, which allows the layers to add residuals to an identity mapping on the channel outputs. The model is characterized by high efficiency and good audio classification performance. The VoxCeleb2 [6] explores ResNet-34 and ResNet-50, adapting their layers to accommodate the spectrogram input.

In our study, we adopted the ResNet-50 architecture and trained it on the VoxCeleb2 data. Following the VoxCeleb papers [6, 12, 13] and other works [14], we first trained the model for the Speaker Identification task using the VoxCeleb datasets and then took the best models to train for a Speaker Verification task. The architecture of the ResNet-50 model is in Table 1.

In Speaker Verification tasks, the model is given two inputs, and its job is to predict a binary outcome. If the inputs are from different speakers, the model returns a 0; if they are from the same speaker, it returns a 1. The model needs at least two input pairs for training, and it no longer requires globally meaningful labels. This is because, in binary decision problems, the label is not typically calculated explicitly. This strategy is called Siamese Networks.

To illustrate this process, let us look at the flowchart in Image 3. Given two inputs $x_1$ and $x_2$ to a Siamese Network, we obtain two embeddings $e_1$ and $e_2$. These embeddings are then used to calculate a distance score $s_{12}$. The score and the known speaker labels are then used to define our binary parameters. As a result, our loss becomes a function of $s_{12}$ and $y_{12}$. The network parameters are updated based on the Loss Function. Training with this approach involves selecting balanced input trials as either a positive trial (same speaker) or a negative trial (different speakers).

**Table 1.** Modified ResNet-50 architecture. Each row specifies the number of convolutional filters and their sizes as $size \times size$, # filters.

| Layer (type) | ResNet-50 |
|---|---|
| conv1 | $7 \times 7$, 64, stride 2 |
| pool1 | $3 \times 3$ max pool, stride 2 |
| conv2_x | $\begin{bmatrix} 1 \times 1 & 64 \\ 3 \times 3 & 64 \\ 1 \times 1 & 256 \end{bmatrix} \times 3$ |
| conv3_x | $\begin{bmatrix} 1 \times 1 & 128 \\ 3 \times 3 & 128 \\ 1 \times 1 & 512 \end{bmatrix} \times 4$ |
| conv4_x | $\begin{bmatrix} 1 \times 1 & 256 \\ 3 \times 3 & 256 \\ 1 \times 1 & 1024 \end{bmatrix} \times 6$ |
| conv5_x | $\begin{bmatrix} 1 \times 1 & 512 \\ 3 \times 3 & 512 \\ 1 \times 1 & 2048 \end{bmatrix} \times 3$ |
| fc1 | $9 \times 1$ 2048 stride 1 |
| pool_time | $1 \times N$ *avgpool* stride 1 |
| fc2 | $1 \times 1$ 5994 |

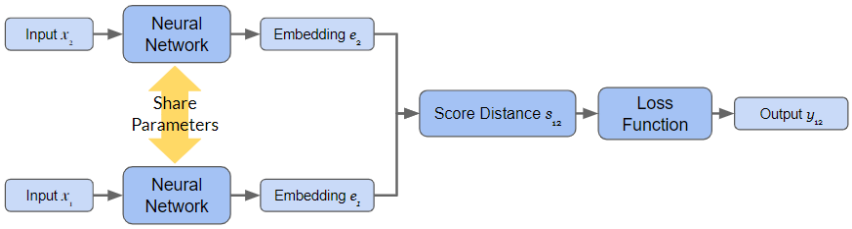


**Fig. 3.** Siamese Workflow.

## 3 Experiments

We began our experiments by downloading the VoxCeleb1 and VoxCeleb2 datasets. We aimed to use the ResNet to train a model for the Speaker Identification task as proposed in the VoxCeleb works [12, 6, 13]. We modified the model to accept Spectrogram data instead of an image. To extract the data, we segmented the audio data into 3-second intervals, sampled at 16kHz, using a window of 25ms and a step of 10ms. We then normalized the data and passed it to the model for classification. In this step, our best model achieved 95.09% accuracy on the VoxCeleb 2 dataset with 5,994 speakers.

In the speaker identification approach, the ResNet50 could accurately identify the speaker of a given utterance. However, this approach assumes a closed set, which means an impostor would be identified as one of our set of speakers. In a Speaker Verification task, we must determine whether a speech belongs to a specific speaker. To accomplish this, we require a new approach that utilizes siamese neural networks, as represented in Figure 3. Verifying a single speaker's identity changes the model's task to a binary decision problem.

In this approach, we take the pre-trained parameters from the Speaker Identification task and then employ the model to learn a distance metric to determine

whether two utterances are spoken by the same person. To train the model, we use the Contrastive Loss function [15, 16]. Given a pair of inputs, the model creates two embeddings and calculates their distance using either the Euclidean Distance or the Cosine Similarity. If the Euclidean distance is chosen, we want embeddings from the same speaker to have a smaller distance than those from different speakers. Alternatively, if the cosine similarity is chosen, we want the score for embedding the same speaker to be higher than those from different speakers.

It can be challenging to balance positive and negative trials during training. An efficient training process for Contrastive Loss depends on selecting the most challenging trials. However, this becomes complicated due to the neural network's parameters changing during training. To address this challenge, we can use hard samples. In this work, we have used offline hard mining, where the trials are computed before the epoch starts using the entire dataset. For our experiments with ResNet50, we adjusted the size of the last layer to 512 and used the trial pairs list from VoxCeleb 1 to evaluate the models using the entire audio as input, as this list is commonly used to compare Speaker Verification models using the Equal Error Rate (EER) metric.

Before performing the training, we evaluated the model using the embeddings from the Speaker Identification step with no retraining to determine a baseline to compare how the training is improving the Verification tasks. We took the embeddings from the penultimate layer. Then, we calculated the ERR by measuring the score distance between each pair of embeddings.

We tested Contrastive Loss by generating $2^{18}$ random pairs without mining. Both models used Euclidean distance and cosine similarity. Furthermore, we aimed to determine the effectiveness of training only the new layer as VoxCeleb, the entire classification module, or even the whole model. The results of these tests are presented in Table 2. In the table, True values indicate that the weights of the modules were retrained, while False values indicate that the weights were frozen. After generating random pairs, we did offline hard mining. We randomly selected only 1 percent of the generated pairs and mined for the hardest negative samples, similar to what was done in VoxCeleb works. This can be understood as the pairs from different speakers with a minimum distance for Euclidean distance. We also generated $2^{18}$ pairs for this strategy.

For each experiment, namely Without Retrain, Random Pairs, and Hard Pairs, we selected the models that gave the best results for each Distance function. After that, we tested these models on our FrSR and FrT datasets. We used the training set of our private datasets to create an embedding for each speaker. Then, we compared them using the data in the test set and calculated the Equal Error Rate (EER). As the duration of audio files is less than 3 seconds, we used the entire audio for evaluation. The results are presented in Table 3.

**Table 2.** EER results for ResNet-50 through experiments *(lower is better)*.

| Experiment | Score Function | CNN Module | FC Module | EER(%) |
|---|---|---|---|---|
| Without Retrain | Euclidean Distance | - | - | 30.40 |
| | Cosine Similarity | - | - | 22.95 |
| Random Pairs | Euclidean Distance | False | False | 32.87 |
| | | False | True | 34.39 |
| | | True | True | 49.24 |
| | Cosine Similarity | False | False | 8.51 |
| | | False | True | 8.81 |
| | | True | True | 7.78 |
| Hard Pairs | Euclidean Distance | False | False | 24.18 |
| | | False | True | 37.35 |
| | | True | True | 49.43 |
| | Cosine Similarity | False | False | **5.96** |
| | | False | True | 8.12 |
| | | True | True | 9.35 |

**Table 3.** EER results applied in CeDRI's datasets *(lower is better)*.

| Experiment | Score Function | FrSR EER(%) | FrT EER(%) |
|---|---|---|---|
| Without Retrain | Euclidean | 7.37 | 15.66 |
| | Cosine | 6.72 | 15.17 |
| Random Pairs | Euclidean | 5 | 15.17 |
| | Cosine | 5.38 | 12.35 |
| Hard Pairs | Euclidean | 1.67 | 13.1 |
| | Cosine | **0.7** | 9.39 |

## 4 Discussion

For our first experiment, we assessed the models using only the weights from the Speaker Identification task without any retraining. These initial results can serve as a baseline to help us understand if the model is being improved. We found that the verification performance was better when using cosine similarity.

For the second experiment, we sampled random trials and retrained for both Euclidean and Cosine Similarity functions. We performed three tests for each function by retraining only the last layer, fully connected layers, or the entire model by retraining the convolutional layers. We observed that when using Cosine Similarity, the EER improved, while using Euclidean Distance had the opposite effect, increasing the EER.

In a subsequent experiment, we tried mining hard samples for the model. We observed that using Euclidean Distance, the model slightly improved its rates when we did not retrain the convolutional layers. However, the improvement was more prominent when using Cosine Similarity. The ERR improved compared to the previous experiment when we did not retrain the convolutional layers.

The experiment indicates that retraining the convolutional layers with hard samples worsens the model, whereas random samples do not. Although the ex-

periments used the same hyperparameters, a possible explanation for this is the small number of samples compared to the dataset. While $2^{18}$ trials were sampled a couple of times for training, we have around $10^6$ utterances, which amounts to approximately $10^{12}$ pairs. Therefore, the set of selected pairs is tiny, and increasing the number of pairs can lead to better results.

The experiments indicated that the models performed well even when trained on a small number of pairs. To put this in perspective, there were approximately 1 trillion possible pairs, but the random training with contrastive loss only utilized around 8 million pairs. This means that the model only saw 0.0008% of the total possible number of pairs, yet the ResNet50 model achieved an EER of 5.96%. For comparison, the ResNet50 model in the VoxCeleb2 paper achieved an EER of 3.95% on the same test set.

For our final experiment, we evaluated the top-performing models for each score function and experiment. We then created an embedding to represent each speaker using the test set from FrT and FrSD datasets for each model. This enabled us to compare the same test set with the GMM model. We have presented each experiment's results in Table 3. The table shows that the models performed well. The ResNet50 model achieved an EER of 0.7% for the FrSR dataset and 9.39% for the FrT dataset using hard pairs and Cosine Similarity.

It is possible that the CNN model did not perform well on the FrT dataset because the audio samples were too short. The FrT audios have a length of 0.944 seconds, shorter than the 3-second length of the audios used to train the model. This could mean that the data from the Trim dataset may not have been sufficient. Therefore, we recommend capturing audio at least 3 seconds long in real-life applications. On the other hand, the FrSR dataset has audio samples that are almost 2 seconds long. This length was enough for the ResNet model to achieve an EER of 0.7%, which is better than the performance of the other models.

To implement a speaker recognition system, we can use the ResNet50 model. Initially, we need to record and store an embedding for each new speaker during the enrolment phase. This embedding will represent the unique characteristics of the speaker's voice. In the authentication phase, new audio will be captured and compared with the stored embeddings to generate a score. The score will then be compared to a threshold to determine if the speaker can access the laboratory. We can create different access levels by using multiple thresholds. Building the system in this manner will allow us to grant laboratory access with high accuracy.

## 5 Conclusion

This study investigates speaker verification methods for future biometric authentication systems. The speaker verification experiments showed that the ResNet50 model has excellent generalization capability, achieving an EER of 0.7% on the Framed Silence Removed dataset. Regarding the Framed Trim dataset, the models did not perform so well, probably because of the duration of the audio.

During the experiments, using cosine similarity led to better results. It is also important to mention that the Framed Silence Removed dataset is smaller, provides audio with a more extensive duration, and is cleaner than the Framed Trim dataset. These characteristics contribute to the model achieving better results in this dataset than in the Framed Trim dataset. However, as the model was retrained in a tiny subset of data for the speaker verification task, there is a margin to explore by training this model in a large set of pairs and consequently improving the models.

Creating a system for the CeDRI laboratory using the ResNet50 model is now possible. Firstly, we can collect an audio sample of each speaker in the laboratory, which will be used to create an "embedding" to represent them. This process is known as the enrollment phase. Once someone requests access, we use the model to generate an embedding of their voice and then compare it to the stored embedding to verify their identity. Additionally, we can enhance the same model by exploring hard mining techniques and replacing just the model weights when necessary. It is important to note that the model does not need to be retrained if a new speaker is introduced. In this case, only the enrollment phase will suffice.

## Acknowledgments

The authors are grateful to the Foundation for Science and Technology (FCT, Portugal) for financial support through national funds FCT/MCTES (PIDDAC) to CeDRI, UIDB/05757/2020 (DOI: 10.54499/UIDB/05757/2020) and UIDP/05757/2020 (DOI: 10.54499/UIDB/05757/2020) and SusTEC, LA/P/0007/2020 (DOI: 10.54499/LA/P/0007/2020).

## References

1. R. Mohd Hanifa, K. Isa, and S. Mohamad. "A review on speaker recognition: Technology and challenges". In: *Computers & Electrical Engineering* 90 (2021), p. 107005.
2. Z. Zhang. "Mechanics of human voice production and control". In: *The Journal of the Acoustical Society of America* 140.4 (Oct. 2016), pp. 2614–2635.
3. J. P. Teixeira and D. S. Freitas. "Segmental durations predicted with a neural network". In: *European Conference on Speech Communication and Technology, Eurospeech/Interspeech 2003*. ISCA. 2003, pp. 169–172.
4. J. P. Teixeira, D. S. Freitas, D. Braga, M. J. Barros, and V. Latsch. "Phonetic events from the labeling the european portuguese database for speech synthesis, FEUP/IPB-DB". In: *European Conference on Speech Communication and Technology, Eurospeech/Interspeech 2001*. ISCA. 2001, pp. 1707–1710.

5. J. P. Teixeira, J. F. T. Fernandes, F. L. Teixeira, and P. O. Fernandes. "Acoustic analysis of chronic laryngitis-statistical analysis of sustained speech parameters". In: *11th International Joint Conference on Biomedical Engineering Systems and Technologies*. 2018, pp. 168–175.
6. J. S. Chung, A. Nagrani, and A. Zisserman. "VoxCeleb2: Deep Speaker Recognition". In: *Interspeech 2018*. ISCA, 2018.
7. D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur. "X-Vectors: Robust DNN Embeddings for Speaker Recognition". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2018, pp. 5329–5333.
8. Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren. "A novel scheme for speaker recognition using a phonetically-aware deep neural network". In: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2014, pp. 1695–1699.
9. V. Guedes, A. Junior, J. Fernandes, F. Teixeira, and J. P. Teixeira. "Long Short Term Memory on Chronic Laryngitis Classification". In: *Procedia Computer Science* 138 (2018), pp. 250–257. ISSN: 1877-0509.
10. N. P. Jawarkar, R. S. Holambe, and T. K. Basu. "Use of fuzzy min-max neural network for speaker identification". In: *2011 International Conference on Recent Trends in Information Technology (ICRTIT)*. IEEE, 2011.
11. H. Tolba. "A high-performance text-independent speaker identification of Arabic speakers using a CHMM-based approach". In: *Alexandria Engineering Journal* 50.1 (2011), pp. 43–47.
12. A. Nagrani, J. S. Chung, and A. Zisserman. "VoxCeleb: A Large-Scale Speaker Identification Dataset". In: *Interspeech 2017*. ISCA, 2017.
13. A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman. "Voxceleb: Large-scale speaker verification in the wild". In: *Computer Speech & Language* 60 (2020), p. 101027.
14. E. Manfron. "Speaker Recognition for Door Opening Systems". Oriented pby João Paulo Teixeira and Rodrigo Minetto. MA thesis. Instituto Politecnico de Braganca, Dec. 2023.
15. S. Chopra, R. Hadsell, and Y. LeCun. "Learning a similarity metric discriminatively, with application to face verification". In: *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*. Vol. 1. IEEE. 2005, pp. 539–546.
16. R. Hadsell, S. Chopra, and Y. LeCun. "Dimensionality reduction by learning an invariant mapping". In: *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*. Vol. 2. IEEE. 2006, pp. 1735–1742.