Blending Face and Voice Recognition for robust Biometric Systems

Biometric Systems (a/y 2024/2025), prof. Maria De Marsico
Master Degree in Computer Science

Marco Realacci
realacci.1938880@studenti.uniroma1.it

Gioele Maria Zoccoli
zoccoli.1850491@studenti.uniroma1.it

Federico Pizzari
pizzari.1936451@studenti.uniroma1.it

Abstract

This project explores a multimodal biometric system that integrates face and voice modalities to enhance the accuracy in both verification and identification tasks. The system utilizes deep architectures, leveraging the FaceNet architecture to extract face embeddings and ECAPA2 to extract robust speaker embeddings. The two subsystems are fused at feature level, to evaluate their combined performance. The evaluation includes multiple scenarios, for each of which we evaluated the system in both verification and identification. Results demonstrate the effectiveness of multimodal systems in achieving superior performance compared to unimodal systems.

Contents

1 Introduction

Multimodal biometric systems enhance security and reliability by combining physiological and behavioral traits, such as fingerprints, facial recognition, and voice patterns. This approach overcomes the limitations of unimodal systems, which rely on a single biometric trait, by improving accuracy and robustness against unauthorized access.[1]

The integration of multiple biometric modalities can occur at various stages of the recognition process, including feature extraction, matching scores, or decision-making. This adaptability enables the development of tailored solutions for specific security needs and operational contexts. [1]

In conclusion, multimodal biometric systems represent a major advancement in identity security. By leveraging multiple traits, they provide superior accuracy, resilience, and versatility, making them ideal for applications ranging from secure access control to identity management.

1.1 Summary of our work

The system leverages two state-of-the-art deep learning architectures for feature extraction, as described in Sections 2 and 3. Additionally, a face localization module was implemented to ensure accurate detection and alignment of facial regions prior to feature extraction (2.1.2).

To facilitate evaluation, we constructed a multimodal dataset by combining samples from two large-scale datasets tailored to each modality (as described in Section 4).

Section 5 presents the evaluation methodology and results, focusing on the system's performance across varying conditions. Finally, Section 6 outlines our conclusions and future directions.

2 Face Modality developement

2.1 Face Detection

2.1.1 MediaPipe: Open-Source Face Detection Framework

MediaPipe, an open-source framework developed by Google, facilitates the creation of multimodal machine learning pipelines, including areas such as computer vision, audio processing, and gesture recognition. [2] One of MediaPipe's key features is the Face Mesh solution, which estimates 468 three-dimensional facial landmarks in real time, even on mobile devices. These landmarks represent key facial points, including the eyes, nose, mouth, and jaw, and are expressed in normalized 3D coordinates (x, y, z).

2.1.2 Face Alignment: Cropping and Centering the Face

Face alignment is the process of normalizing facial images to achieve a consistent representation of the face. In our project, we utilized MediaPipe exclusively for detecting

[1] Mitek Systems, "A Comprehensive Overview of Multimodal Biometrics" link

[2] Google, "MediaPipe" repository

the face and obtaining its bounding box, while the cropping and further adjustments were implemented independently.

The key steps are:

- **Face Detection**: Using MediaPipe's face detection model to identify the face and obtain a bounding box around the facial region.
- **Cropping the Region of Interest**: Isolating the face from the surrounding context by cropping the image based on the obtained bounding box.
- **Adding Margins**: To include additional details and avoid cutting off parts of the face, margins were added above and below the bounding box. These margins are proportional to the height of the bounding box and ensure the cropped face remains within the image bounds.
- **Square Image Adjustment**: After adding the margins, the resulting image was adjusted to ensure the face is centered and the output image is squared, maintaining uniformity.

This process ensures that facial images are consistent and normalized, enhancing accuracy in applications such as facial recognition and expression analysis.

2.2 Feature extraction

To extract features from the cropped face, we leveraged a deep Convolutional Neural Network using the FaceNet architecture [2]. FaceNet directly maps face images to a compact Euclidean embedding space where distances correspond to a measure of face similarity. Tasks such as face recognition, verification, and clustering are efficiently implemented using embeddings derived from this space. In our implementation, we utilized the pre-trained weights of FaceNet, trained on the VGGFace2 dataset [3], which are available on the Hugging Face platform[3]. This approach allowed us to accelerate the development process while leveraging a robust feature extractor.

2.2.1 Implementation Details

The implemented model adopts the Inception-ResNet V1 architecture [4] as its backbone. This architecture combines the strengths of Inception modules and residual learning, achieving high representational efficiency while maintaining computational feasibility. Below, we describe key architectural components in detail. To define the architecture, we leveraged some code fragments from a GitHub repo[4]. We implemented the model using the PyTorch[5] library.

[3] Pre-trained FaceNet weights on HuggingFace: https://huggingface.co/py-feat/facenet

[4] PyTorch implementation of Inception-Resnet-V1: https://github.com/timesler/facenet-pytorch

[5] PyTorch documentation: https://pytorch.org

2.2.2 Basic Building Blocks

The architecture (diagrams in Figure 1) leverages the following modules:

- **BasicConv2d**: A convolutional block comprising a convolutional layer (without bias), batch normalization, and a ReLU activation. This block ensures efficient feature extraction while maintaining numerical stability.

- **Residual Blocks**: Three primary types of residual blocks are used. Each block implements parallel convolutional branches with varying kernel sizes, concatenating their outputs and introducing residual connections scaled by a factor. These residual connections enhance gradient flow during training.

2.2.3 Key Architectural Components

The architecture is divided into distinct stages:

- **Initial Feature Extraction**: The model begins with a series of convolutional layers and a max pooling layer (Stem block), extracting low-level features from the input image.

- **Intermediate Layers**:
  - **5x Inception-Resnet-A** employs 5x Inception-Resnet-A modules to capture intricate spatial hierarchies.
  - **Reduction-A** integrates three branches: a 3x3 convolution, a 1x1 followed by two 3x3 convolutions, and a max pooling layer. This stage reduces spatial resolution while increasing feature dimensionality.
  - **10x Inception-resnet-B** consists of stacked Block17 modules, refining features with a wider receptive field through asymmetric kernels.

- **Final Feature Extraction and Embedding**:
  - **Reduction-B** combines four parallel branches (3x3, 1x1-3x3, 1x1-3x3-3x3, and max pooling).
  - **5x Inception-resnet-C** to refine high-level features.
  - The final layers include global average pooling, dropout (to prevent overfitting), and a fully connected layer mapping to a 512-dimensional embedding space. Batch normalization ensures the embeddings lie on a unit hypersphere, enabling robust similarity measures.

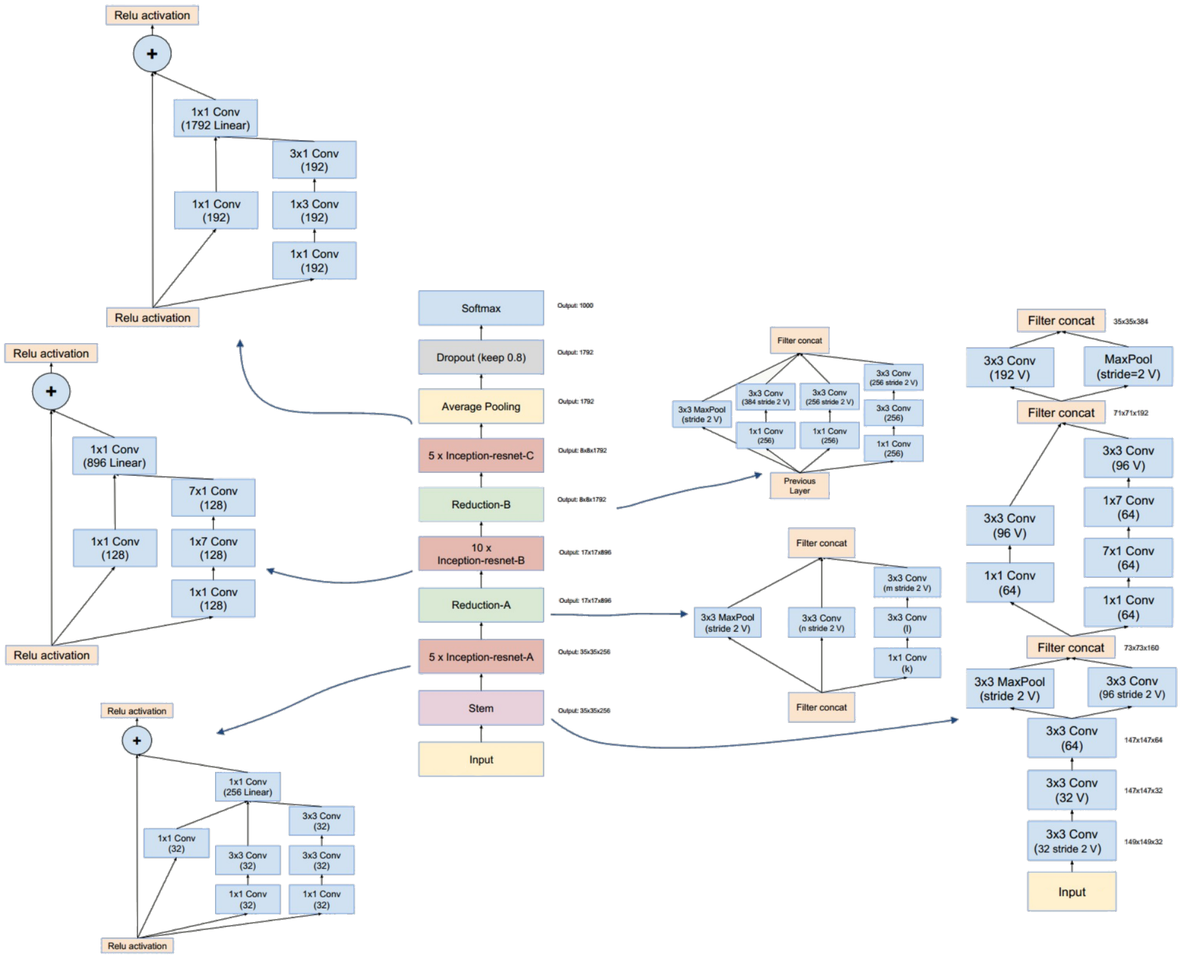

Figure 1: Inception-Resnet-V1 Architecture

Image source: https://www.aiuai.cn/aifarm465.html

3 Voice Modality developement

3.1 Implementation and Architecture

For the voice modality, we utilized an hybrid neural network based on the ECAPA2 architecture [5]. ECAPA2 is a hybrid neural network designed to generate robust speaker embeddings by combining the strengths of both 1D and 2D convolutional operations. This architecture excels in speaker recognition tasks by producing embeddings that are robust against overlapping speech and short utterances. For the purpose of the project, we leveraged an implementation of ECAPA2 pre-trained on the VoxCeleb2 dataset [6], which is publicly available on Hugging Face [6], which facilitated rapid development and integration into our system.

[6] Pre-trained ECAPA2 on Hugging Face: https://huggingface.co/Jenthe/ECAPA2

3.1.1 Implementation Details

ECAPA2's hybrid design leverages a combination of Local Feature Extractor (LFE) blocks and a Global Feature Extractor (GFE) module to address the limitations of traditional speaker verification models. Below, we detail the key components of this architecture.

3.1.2 Key Architectural Components

- **Local Feature Extractor (LFE)**: The LFE module consists of a cascade of 2D-convolutional layers followed by frequency-wise Squeeze-and-Excitation (3.1.3) modules (Figure 3). These components enable the network to learn spatially invariant features, improving robustness against input perturbations such as overlapping speakers or noise. The module also integrates learnable positional encodings to incorporate frequency positional information.

- **Global Feature Extractor (GFE)**: The GFE module is implemented using a lightweight TDNN subnetwork placed at the end of the architecture. This subnetwork aggregates frequency information from the LFE and ensures a uniform Effective Receptive Field (ERF) across the frequency dimension. The GFE enhances the model's capacity to exploit global frequency patterns while maintaining computational efficiency.

- **Channel-Dependent Attentive Statistics Pooling (CAS)**: CAS pooling is applied to integrate global context into the final embeddings. This module computes weighted mean and standard deviation statistics, which are then projected to a 192-dimensional speaker embedding.

3.1.3 Frequency-wise Squeeze-Excitation (fwSE) Block

The Frequency-wise Squeeze-Excitation (fwSE) block [7] is an enhancement of the standard Squeeze-Excitation (SE) mechanism [8]. While traditional SE blocks apply a single scalar weight per channel, fwSE introduces frequency-specific scaling, enabling the network to better capture frequency-dependent variations in speech data. This is particularly advantageous in speaker verification tasks, where frequency characteristics are critical.

Squeeze Operation In the fwSE block, the squeeze operation computes a mean descriptor vector $\mathbf{z} \in R^F$, where F is the number of frequency bins. For an input feature map $\mathbf{X} \in R^{C\times F\times T}$, where C, F, and T represent the channel, frequency, and temporal dimensions respectively, $\mathbf{z}$ is calculated as:

$$z_f = \frac{1}{C\cdot T}\sum_{i=1}^{C}\sum_{j=1}^{T} x_{f,i,j}, \quad \forall f \in \{1,\dots,F\}, \tag{1}$$

where $x_{f,i,j}$ denotes the value of the feature map at frequency f, channel i, and time j.

Excitation Operation The excitation operation generates scaling factors $\mathbf{s} \in R^F$ for each frequency bin based on the mean descriptor vector $\mathbf{z}$. The scaling factors are computed as:

$$\mathbf{s} = \sigma(\mathbf{W}_2 ReLU(\mathbf{W}_1 \mathbf{z} + \mathbf{b}_1) + \mathbf{b}_2), \tag{2}$$

where $\mathbf{W}_1$ and $\mathbf{W}_2$ are learnable weight matrices, $\mathbf{b}_1$ and $\mathbf{b}_2$ are biases, $f(\cdot)$ is a ReLU activation function, and $\sigma(\cdot)$ denotes the sigmoid function.

Rescaling The input feature map $\mathbf{X}$ is scaled along the frequency dimension using the computed factors $\mathbf{s}$. Specifically, for each frequency bin f:

$$\mathbf{X}_f \leftarrow \mathbf{s}_f \cdot \mathbf{X}_f, \tag{3}$$

where $\mathbf{X}_f \in R^{C \times T}$ is the slice of the feature map corresponding to frequency bin f.

By rescaling features in a frequency-specific manner, the fwSE block injects global frequency information into the network while preserving local details. This modification significantly enhances the model's ability to capture speaker-dependent characteristics in speech signals.

3.1.4 Training Strategy

The ECAPA2 architecture was trained using the development partition of the VoxCeleb2 dataset, employing a Subcenter Additive Angular Margin (AAM) softmax loss function to enhance intra-class compactness and inter-class separation. Notable training strategies include:

- **Margin-Mixup**: A technique that enhances robustness against overlapping speakers by predicting target classes from a mixture of two speakers with varying energy ratios.

- **Variable Length Training (VLT)**: To improve performance on short utterances, the training protocol alternated between standard utterance lengths and shorter random crops.

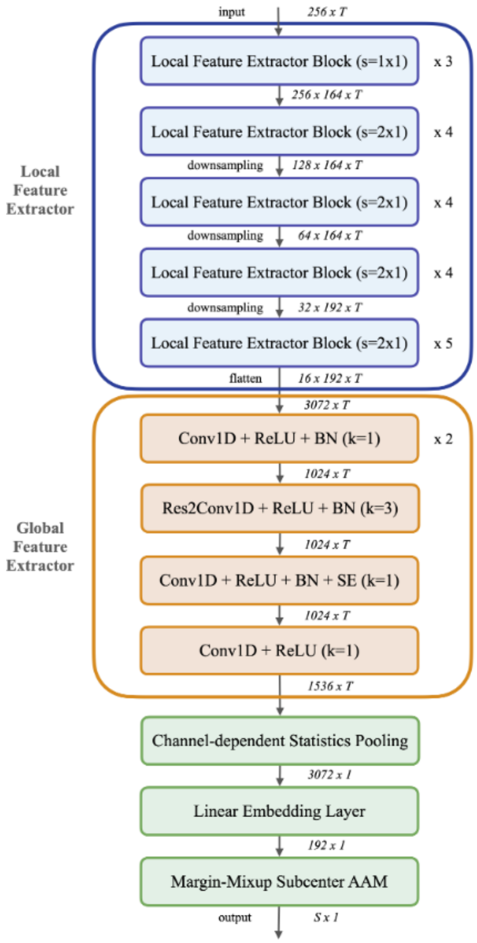

Figure 2: ECAPA2 Architecture

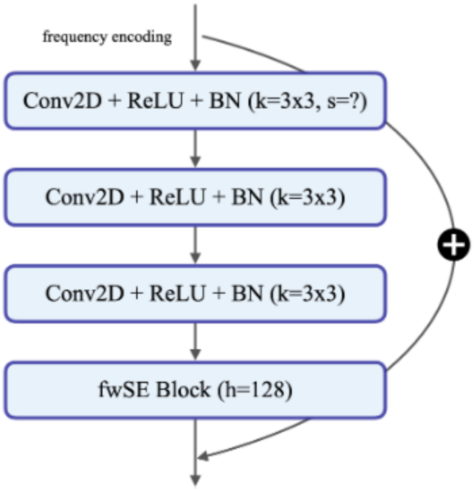
frequency encoding
Conv2D + ReLU + BN (k=3x3, s=?)
Conv2D + ReLU + BN (k=3x3)
Conv2D + ReLU + BN (k=3x3)
fwSE Block (h=128)

Figure 3: ECAPA2: Local Feature Extractor Block

4 Constructing a Multimodal dataset

Before proceeding with the evaluation phase, we needed to find a dataset that met our need of multiple faces and voice tracks of an individual. After a careful research, we lacked significant findings since a dataset this specific is not freely available on the web, and the few ones that exists require a registration and authorization phase in order to acquire the files.

In order to overcome this problem, our first and final solution, was to merge in a smart fashion two different datasets that provided respectively multiple instances of faces and voice registrations of the same individual. In this way, we achieved our goal of creating a multimodal dataset that enabled us to reliably evaluate our model.

4.1 Face dataset

For the faces dataset, we decided to use the CelebFaces Attributes Dataset (CelebA) from the University of Hong Kong [9]. The dataset provided to the public does not include real face identities, available only upon request, but was nonetheless considered fit for our use-case, since the images of the same individual were mapped with the same anonymous id.

Our choice provided us with 202.599 anonymously labeled images that represented 10.177 different identities, covering large pose variations and background clutter. We preferred this dataset over other ones for the high mean of images per identity and the high number of identities in total. We also had the impression that the images of this dataset looked like they were taken during a larger time span for several identities, but this is just an impression we noted when searching for datasets.

After downloading it, we performed some analysis in order to evaluate the distribution of face photos for each identity and we found out that it was a little bit inconsistent. In order to overcome this problem, we created our version of the CelebA dataset, with a more convenient file structure and in which we included only the identities with more than 25 occurrences, in order to discard the identities with less representation. After this operation was successfully done, a total of 1487 identities fit for our purpose were found.

4.2 Voice dataset

For the voice dataset, our choice fell upon LibriSpeech [10]. This dataset is a corpus of approximately 1000 hours of read English speech. The data is derived from read audio books from the LibriVox project, and has been carefully segmented and aligned. We chose this dataset because it contains lots of short voice registrations (~ 10 seconds), perfect for our use-case.

The authors present two versions of the LibriSpeech dataset: *clean* and *other*, which differ in the quality of their recordings and transcriptions. The *clean* subset features higher-quality audio with lower word error rates (WER), facilitating its use in less challenging ASR scenarios. In contrast, the other subset includes recordings with greater variability and higher WER, offering a more demanding benchmark for ASR systems. This distinction allows for comprehensive evaluation across diverse conditions.

Each dataset is divided into different downloadable archives, differing in the size. Some comprehensive files are also provided in order to give more information about the speakers (like name, book read and gender). The data is divided in folders, with each folder representing the unique id of the speaker. Inside this folder, there are other folders with the ids of the various audio books and inside those we finally find the voice registrations.

Like with the images dataset, we proceeded with some data skimming and organization before the integration of the two datasets. In particular, we extrapolated the files of each person in the various datasets and merged all the voice recordings into one folder per identity, removing in this way the layer of folders regarding the audio book read. We then removed the identities with less than 25 registrations, leaving us with a total of 5874 different speakers.

4.3 Integrating Face and Voice data

For the integration part, we wanted to create our dataset by combining the identities using an approach that will guarantee us to have a near 1:1 rate of images and voice registrations for each identity. In order to achieve this result, we iterated on each identity in the faces dataset and selected a correspondent in the voices dataset with the same gender and the most similar number of samples. Then, a new unique id was generated and the files were randomly paired one another.

After this operation, we were left with a total of 1167 identities, 487 males and 680 females.

5 Evaluation

5.1 Fusion of Biometric Traits

To identify the most effective method for combining the two biometric traits we conducted an ALL-against-ALL (explained in the section below) evaluation on our dataset. For the voice modality, we utilized the LibriSpeech-other dataset, which presents a more challenging setting compared to LibriSpeech-clean. Our dataset comprises 1167 identities, each with 5 samples, where the voice recordings were truncated to 3 seconds per utterance. Our goal was to explore various feature-level fusion strategies and evaluate their performance across multiple metrics.

5.1.1 All-against-All evaluation

In the All-against-All evaluation, each template plays in turn the role of either genuine/impostor or enrolled/not enrolled more than once according to the recognition application.

We evaluated the systems for the following applications considering the *multiple-template* scenario:

- **Verification:** For each probe, we perform one genuine attempt (comparing the probe against the gallery templates of the same identity) and $N-1$ impostor attempts (comparing the probe against the gallery templates of different identities).
- **Identification Open Set:** For each probe, we perform one genuine attempt (we try to identify the user against the gallery) and one impostor attempt (we test how the user behaves as an impostor).
- **Identification Closed Set:** In this case, for each probe, we perform only one genuine attempt. The system is constrained to identify the probe within the known identities of the gallery set.

5.1.2 Fusion Strategies

We considered several feature-level fusion strategies, described as follows:

- **Face only**: Using only the face embeddings, represented as a feature vector of length 512.
- **Voice only**: Using only the voice embeddings, represented as a feature vector of length 192.

- **Concatenation (concat)**: Direct concatenation of face and voice embeddings:

$$\mathbf{v} = [\mathbf{v}_{\text{face}}, \mathbf{v}_{\text{voice}}]$$

  resulting in a combined feature vector of length 704.

- **L2 norm + concat**: Normalizing both face and voice embeddings to unit vectors before concatenation:

$$\mathbf{v} = \left[\frac{\mathbf{v}_{\text{face}}}{\|\mathbf{v}_{\text{face}}\|}, \frac{\mathbf{v}_{\text{voice}}}{\|\mathbf{v}_{\text{voice}}\|}\right]$$

- **L2 norm + sum**: Normalizing both embeddings to unit vectors and summing them, padding the voice embeddings with zeros to match the face embedding length:

$$\mathbf{v} = \frac{\mathbf{v}_{\text{face}}}{\|\mathbf{v}_{\text{face}}\|} + \text{pad}\left(\frac{\mathbf{v}_{\text{voice}}}{\|\mathbf{v}_{\text{voice}}\|}\right)$$

- **Scaling + concat**: Standardizing each embedding (mean subtraction and division by standard deviation) before concatenation.

$$\mathbf{v} = \left[\frac{\mathbf{v}_{\text{face}} - \mu_{\text{face}}}{\sigma_{\text{face}}}, \frac{\mathbf{v}_{\text{voice}} - \mu_{\text{voice}}}{\sigma_{\text{voice}}}\right]$$

- **Scaling + sum**: Standardizing both embeddings and summing them, with zero-padding for the voice embeddings.

$$\mathbf{v} = \frac{\mathbf{v}_{\text{face}} - \mu_{\text{face}}}{\sigma_{\text{face}}} + \text{pad}\left(\frac{\mathbf{v}_{\text{voice}} - \mu_{\text{voice}}}{\sigma_{\text{voice}}}\right)$$

5.1.3 Matching function

We used cosine similarity as the matching function because it effectively measures the similarity between two feature vectors by calculating the cosine of the angle between them. Cosine similarity is widely used in biometric and other pattern recognition tasks, making it an ideal choice for our evaluation. It is preferred over Euclidean distance because it focuses on the angle between vectors, making it insensitive to their magnitudes. This is important when comparing embeddings, as it captures the relative orientation between vectors rather than their scale. Additionally, cosine similarity is more robust in high-dimensional spaces, where Euclidean distance can become less informative.

As the codomain of the cosine similarity is $[-1, 1]$, to normalize the output in the range $[0, 1]$, we leveraged the min/max function:

$$F(x) = \frac{x+1}{2}$$

Thus, the similarity between two vectors A and B is computed as:

$$s(A, B) = F\left(\frac{A \cdot B}{\|A\|\|B\|}\right)$$

5.1.4 Results

The results of the evaluation are summarized in Table 1. We report metrics for verification (EER, AUROC, ZeroFAR, ZeroFRR), open-set identification (EER, AUROC), and closed-set identification (CMS@1).

From the results, it is evident that the **L2 norm + concat** strategy outperformed other fusion methods across all metrics. This approach achieved the lowest EER for verification (0.199%), the highest AUROC in both verification (0.9998) and open-set identification (0.9965), and the best closed-set identification accuracy (99.61% CMS@1). Normalizing embeddings prior to fusion appears to enhance performance, as seen in the comparison between raw concatenation and normalization-based approaches.

Methods using summation generally underperformed compared to concatenation, likely due to the inherent differences in dimensionality and distribution between face and voice embeddings. Scaling without normalization showed a significant degradation in performance, emphasizing the importance of feature normalization when combining embeddings.

These findings highlight the effectiveness of normalization in feature-level fusion and suggest that concatenation, when paired with proper preprocessing, is the optimal strategy for combining face and voice biometrics.

Fusion rule	**Verification**				**Open Set**		**Closed Set**
	EER (%)	**AUROC**	**ZeroFAR** (%)	**ZeroFRR** (%)	**EER** (%)	**AUROC**	**CMS@1** (%)
Voice only	0.252	0.9998	91.94	68.23	4.55	0.9844	99.30
Face only	2.778	0.9949	91.58	99.42	26.6	0.8092	81.92
L2 norm + concat	**0.199**	**0.9998**	**10.84**	**10.56**	**1.80**	**0.9965**	**99.61**
L2 norm + sum	0.272	0.9997	12.98	18.50	1.95	0.9957	99.58
Scaling + sum	0.908	0.9993	41.61	86.37	6.36	0.9696	97.19
Scaling + concat	0.821	0.9995	21.73	19.70	5.24	0.9763	97.76

Table 1: Impact of audio quality on performance.

5.2 Performance Metrics and Visualization

The following plots relate to the best-performing scenario identified in this first evaluation. To provide a comprehensive analysis of its performance, we present the following visualizations.

5.2.1 Verification

In the verification task, the system is asked to determine whether a given pair of biometric traits, face and voice, belong to the same identity or not. To evaluate this, we present several key performance curves. The **FAR vs FRR curve** displays the relationship between the False Acceptance Rate (FAR) and the False Rejection Rate (FRR), highlighting the Equal Error Rate (EER) where these two rates are equal. The **DET curve** shows the tradeoff between detection errors (False Acceptance and False Rejection) at different thresholds. Additionally, the **ROC curve** provides an overall performance view by plotting the Genuine Acceptance Rate (GAR) against the False Acceptance Rate (FAR) for various thresholds.

These curves allow us to assess the system's accuracy and robustness in matching face and voice biometrics for individual identity verification tasks.

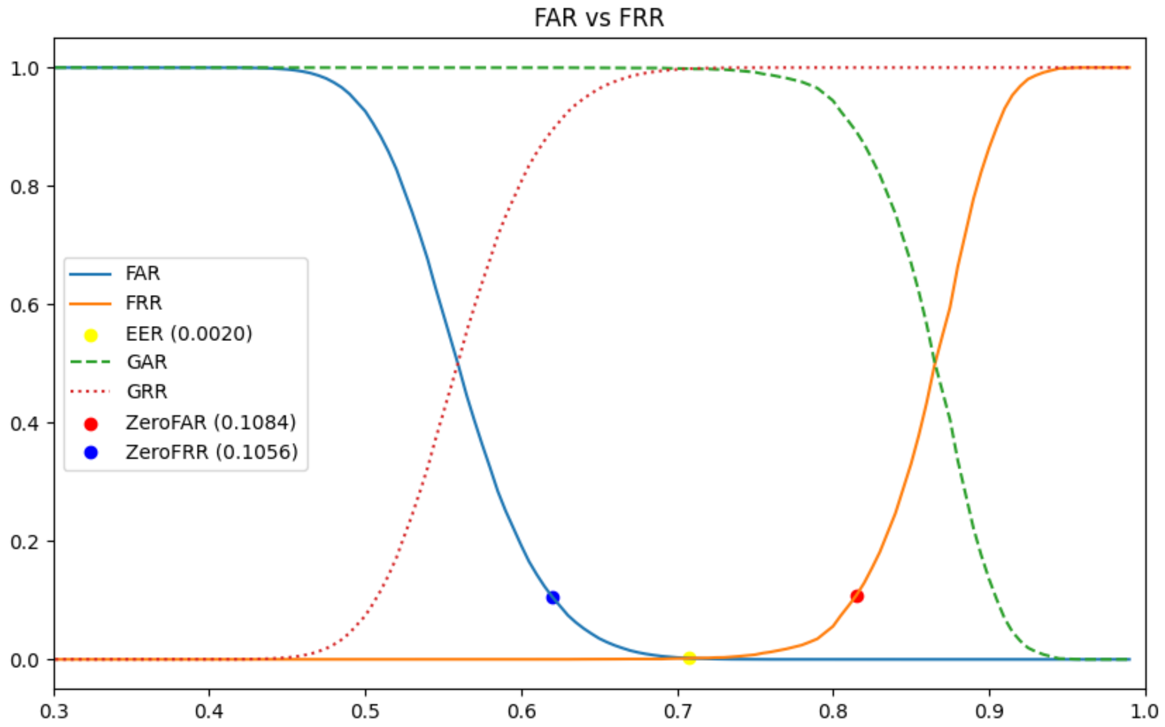

Figure 4: Verification: FAR vs FRR curves

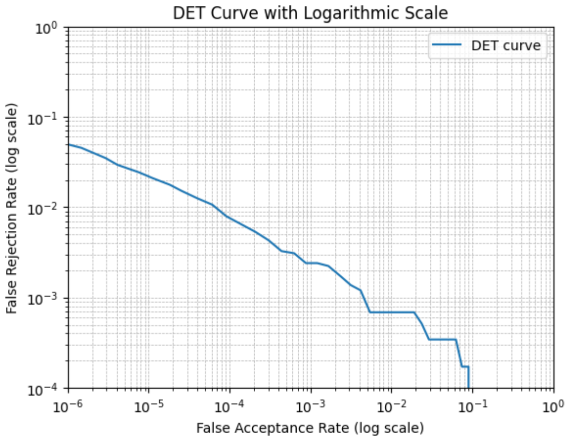

Figure 5: Verification: DET curve

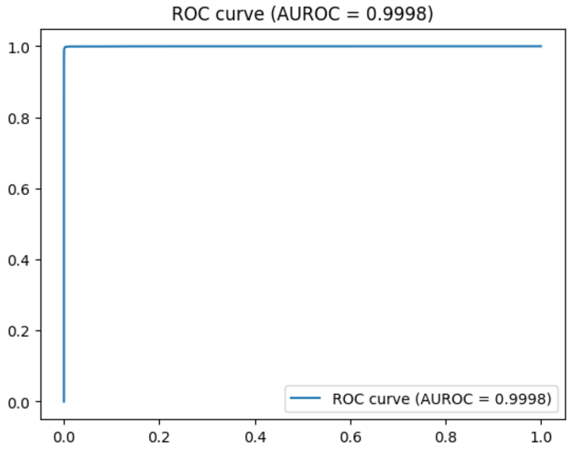

Figure 6: Verification: ROC curve

5.2.2 Identification Open Set

In the open-set identification task, the system is required to recognize whether a given template matches one of the known identities or if it belongs to an unknown person. This scenario is more challenging than closed-set identification, as it involves distinguishing between known and unknown individuals. For this task, we present the **FAR vs FRR curve**, which illustrates the tradeoff between false acceptances and false rejections, providing insight into the system's performance across various thresholds. Additionally, the **ROC curve** is used to evaluate the system's overall discrimination ability by plotting the Detection and Identification Rate at rank 1 (DIR@1) against the False Acceptance Rate (FAR) at different thresholds.

These visualizations help in understanding how well the system can perform in a more realistic setting, where some biometric samples may belong to unknown individuals.

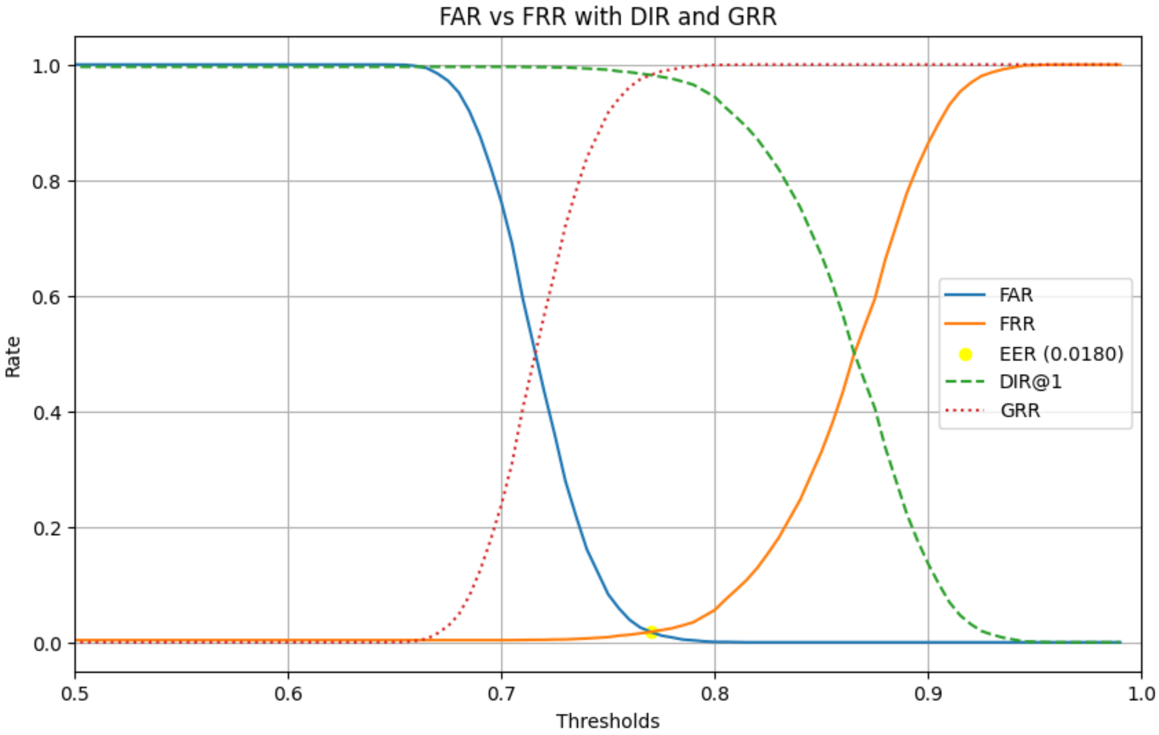

Figure 7: Open Set Identification: FAR vs FRR curves

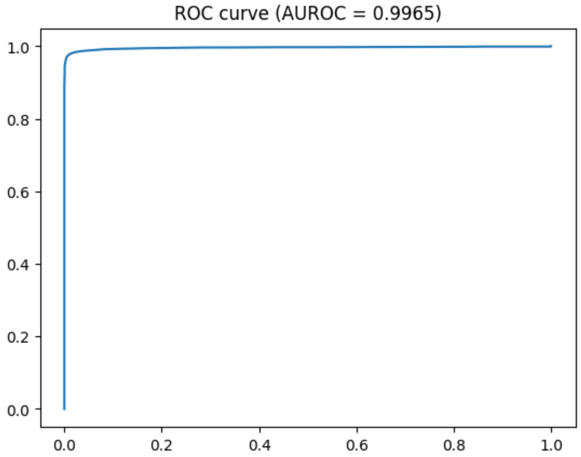

Figure 8: Open Set Identification: ROC curve

5.2.3 Identification Closed Set

In the closed-set identification task, the system is given a template and must correctly identify the person from a predefined set of known identities. This is a standard identification scenario, where all possible identities are part of the system's database. To evaluate performance in this context, we present the **CMC (Cumulative Match Characteristic) curve**, which shows the probability of a correct match as a function of the rank in the ordered list of identities. The CMC curve helps to assess how quickly the system can find the correct identity within the set, providing valuable insight into the ranking quality and the effectiveness of the fusion method.

This metric is crucial for understanding the system's ability to correctly identify individuals from a closed set of known identities.

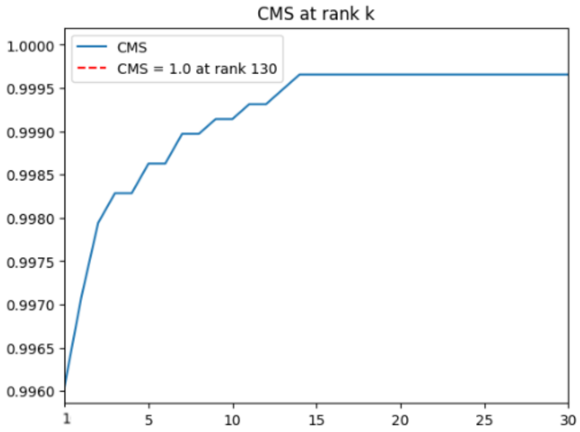

Figure 9: Closed Set Identification: CMC curve

These plots provide insights into the system's performance across different biometric evaluation tasks and metrics.

5.3 Using LibriSpeech-clean

In this evaluation, we utilize the LibriSpeech-clean dataset for the voice modality, in contrast to the previously used LibriSpeech-other dataset. The 'clean' subset features high-quality recordings with minimal background noise. For the face modality, we continue to use the CelebA dataset.

The test scenario remains unchanged from the previous evaluation. We performed an ALL-against-ALL evaluation with a total of 1167 identities, using 5 samples per identity. For the fusion strategy, we applied the L2 norm + concat method, where both the face and voice embeddings were normalized and then concatenated, allowing us to combine the features from both modalities for improved performance.

5.3.1 Verification

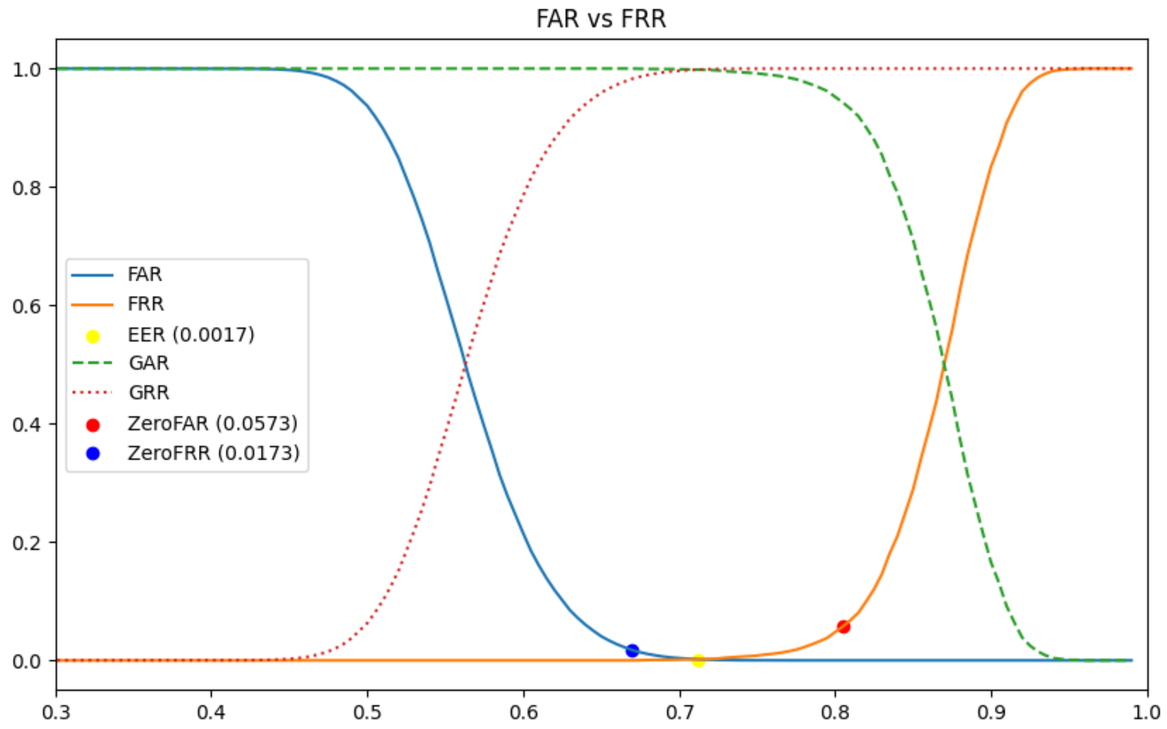

Figure 10: Verification: FAR vs FRR curves

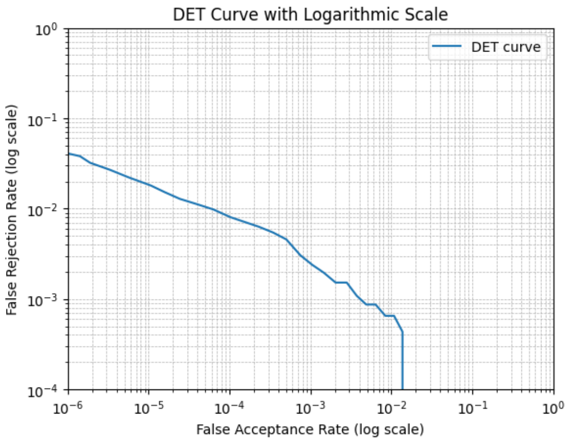

Figure 11: Verification: DET curve

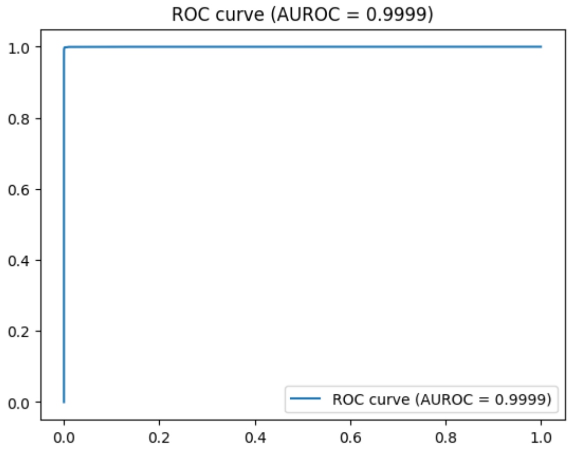

Figure 12: Verification: ROC curve

5.3.2 Identification Open Set

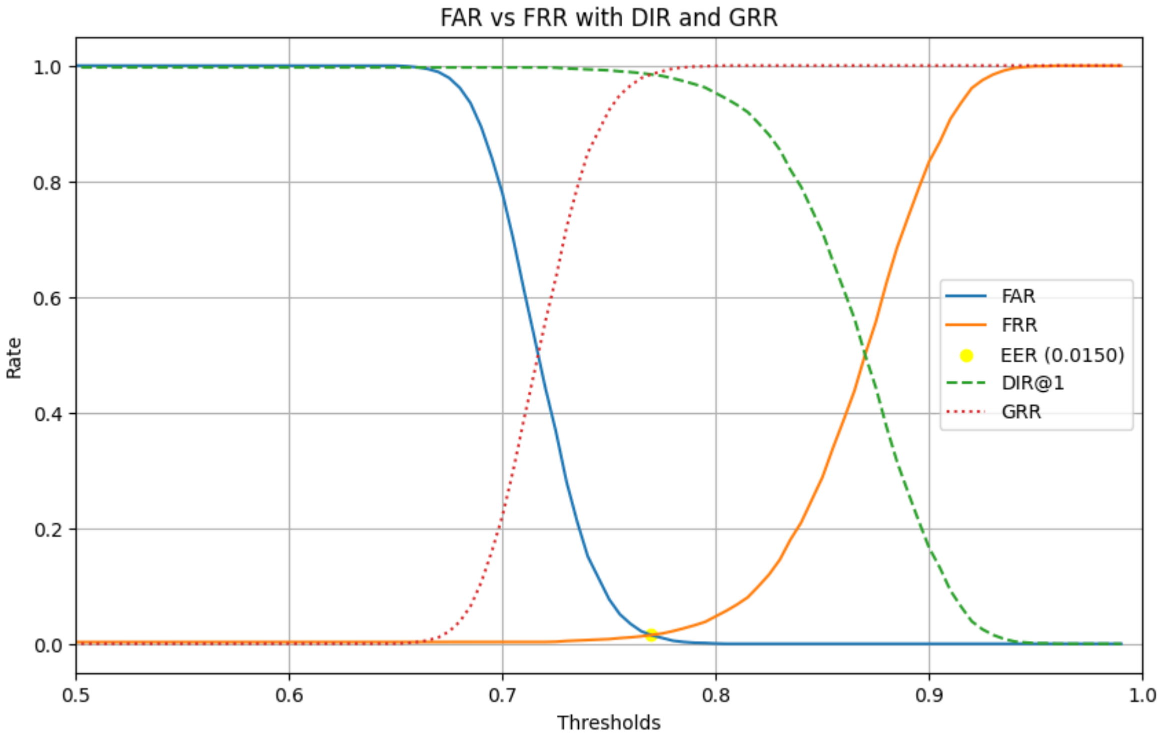

Figure 13: Open Set Identification: FAR vs FRR curves

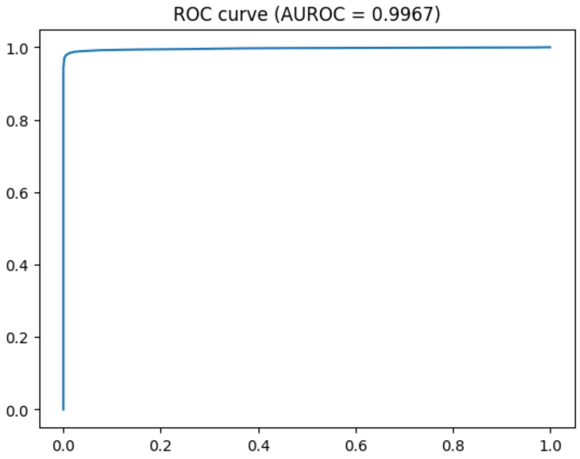

Figure 14: Open Set Identification: ROC curve

5.3.3 Identification Closed Set

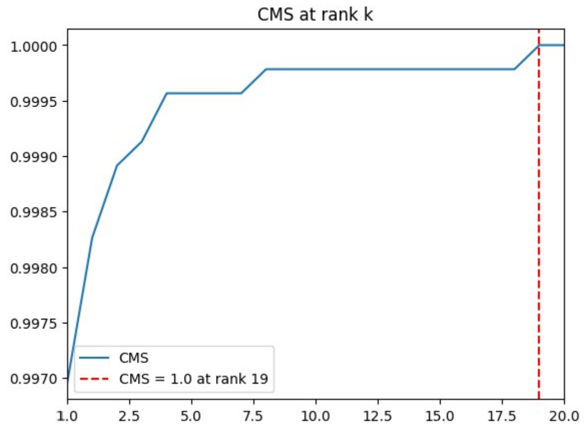

Figure 15: Closed Set Identification: CMC curve

5.3.4 Performance comparison with LibriSpeech-clean vs LibriSpeech-other

As expected, the performance of the fusion strategy improves when using the LibriSpeech-clean dataset compared to the LibriSpeech-other dataset. The results in Table 2 show that the LibriSpeech-clean dataset achieves lower error rates across all the metrics, with surprising low values for **ZeroFAR** and **ZeroFRR**. These differences highlight the impact of higher-quality, cleaner data. The gain in performance was expected, given the reduced noise and clearer speech in the LibriSpeech-clean dataset.

Voice dataset	**Verification**				**Open Set**		**Closed Set**
	EER	**AUROC**	**ZeroFAR**	**ZeroFRR**	**EER**	**AUROC**	**CMS@1**
	(%)		(%)	(%)	(%)		(%)
LibriSpeech-other	0.199	0.9998	10.84	10.56	1.80	0.9965	99.61
LibriSpeech-clean	**0.171**	**0.9999**	**5.73**	**1.73**	**1.50**	**0.9967**	**99.70**

Table 2: LibriSpeech-clean vs LibriSpeech-other

5.4 Impact of the Number of Samples per Identity in the Gallery

For the purpose of this analysis, we performed an All Probe-against-All Gallery evaluation to assess the impact of the number of samples per identity in the **template gallery**. In this setup, there are two distinct sets: the **probe set** and the **gallery set**. We again performed the evaluation under the **multiple-template** scenario, as described in Section 5.1.1.

We tested three different scenarios where the number of samples per identity in the gallery varied, while the total number of identities in the gallery set remained constant at 300. The probe set consisted of 300 genuine users (who also appear in the gallery) and 500 impostor users. Below, we provide a detailed analysis of the results for each of these three tasks.

5.4.1 Results

The Table 3 presents the performance results for varying numbers of gallery samples per identity, evaluated across three different tasks: Verification, Open Set Identification, and Closed Set Identification. It can be observed that increasing the number of samples per identity in the gallery leads to improved performance. For Verification, the Error Rate (EER) decreases as the number of gallery samples increases, from 0.68% with 1 sample to 0.16% with 5 samples. The EER in the Identification Open Set scenario improved by a factor of *3.9*. Similarly, the Correct Match at Rank 1 (CMS@1) improves with more gallery samples, reaching 99.80% with 5 samples. Overall, these results demonstrate the benefits of having more samples per identity, especially in the Open Set Identification task, where the system benefits from enhanced recognition accuracy.

#gallery templates	**Verification**				**Open Set**		**Closed Set**
per identity	**EER**	**AUROC**	**ZeroFAR**	**ZeroFRR**	**EER**	**AUROC**	**CMS@1**
	(%)		(%)	(%)	(%)		(%)
1	0.68	0.9998	**13.96**	18.26	4.18	0.9861	99.0
2	0.25	**0.9998**	14.23	7.78	1.50	0.9960	99.75
3	0.19	**0.9998**	17.57	9.70	1.47	0.9977	**99.82**
5	**0.16**	**0.9999**	21.53	**2.33**	**1.07**	**0.9984**	99.80

Table 3: Impact of varying gallery sample sizes on performance

5.4.2 Trade-off between FAR and FRR

Generally speaking, increasing the number of samples in the gallery tends to decrease the FRR while leading to an increase in the FAR. This trade-off is a common observation in biometric systems, where a larger gallery provides more data for accurate matching, but also raises the likelihood of misidentifying impostors. However, it is worth noting that the EER decreased as well, indicating that the system's overall performance improved with a larger gallery. This suggests that, while there is a trade-off between FAR and FRR, the overall system's efficiency in correctly identifying users may benefit from adding more samples to the gallery.

6 Conclusions

The evaluation work clearly shows the effectiveness of combining voice and video modalities for biometric person recognition. The integration of these two modalities leverages the strengths of both, resulting in a robust and reliable system with improved accuracy compared to single-modal approaches, offering enhanced resilience to variations in environmental conditions, occlusions, and noise.

6.1 Future work

One potential area for future improvement is the incorporation of liveness detection techniques to enhance the system's resilience against spoofing attacks. Spoofing, such as using pre-recorded voice samples or photos, poses a significant threat. To address this, a binary classifier could be implemented to distinguish between genuine and spoofed inputs for both voice and video modalities.

The scores generated by the liveness detection classifier could then be fused with the biometric recognition scores, creating a unified decision-making framework. This fusion would allow the system to mitigate the impact of spoofing attempts while maintaining high recognition accuracy.

References

[1] A. Ross and A. K. Jain, "Multimodal biometrics: An overview," in *2004 12th European Signal Processing Conference*, pp. 1221–1224, 2004.

[2] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 815–823, IEEE, June 2015.

[3] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," 2018.

[4] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, 2015.

[5] J. Thienpondt and K. Demuynck, "Ecapa2: A hybrid neural network architecture and training strategy for robust speaker embeddings," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–8, 2023.

[6] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *Interspeech 2018*, interspeech_2018, ISCA, Sept. 2018.

[7] J. Thienpondt, B. Desplanques, and K. Demuynck, "Integrating frequency translational invariance in tdnns and frequency positional information in 2d resnets to enhance speaker verification," in *Interspeech 2021*, interspeech_2021, p. 2302–2306, ISCA, Aug. 2021.

[8] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[9] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[10] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015.