Enhancing Authentication Security: A Fusion of Face and Voice Recognition

Leila Hellal

Enhancing Authentication Security: A Fusion of Face and Voice Recognition

Hellal Leila[1], Boukezzoula Naceur-Eddine[2], Cheniti Mohamed, Chenni Kenza
LTS laboratory *University Ferhat Abbas*, Setif, Algeria
leila.hellal@univ-setif.dz[1], nasrbou@yahoo.fr[2], kenza.chenni06@gmail.com

***Abstract*— The XM2VTS dataset is a challenging dataset in biometric authentication due to variations in lighting, facial expressions, and voice characteristics. Conventional methods encounter difficulties in handling such uncertainties. To overcome these challenges, the study proposes the utilization of T-norms, a fuzzy logic-based approach that provides a flexible framework for modeling uncertainty. By employing T-norms, the study demonstrates the feasibility of combining multiple biometric modalities, specifically facial and voice recognition. This integration enhances authentication accuracy by necessitating a consensus between modalities for access. The application of T-norms not only improves security measures but also enhances adaptability, making it a promising solution for addressing the intricate nature of the XM2VTS dataset in biometric authentication.**

Keywords—biometric, score level, T-Norms, XM2VTS, DCT, fusion, face, voice

I. INTRODUCTION

Biometric identification uses unique physiological or behavioral traits like fingerprints, irises, faces, and voices to verify individuals' identities [1]. Single-modal biometric systems, which rely on a single trait, face challenges such as variations within categories and vulnerability to spoofing. On the other hand, multimodal biometrics, which combines multiple traits, consistently outperform single-modal systems, providing improved accuracy, resistance to noise, universality, anti-spoofing capabilities, and greater resilience.

Within multimodal biometrics, data fusion from different modalities is essential, occurring at feature, score, or decision levels. While feature-level fusion can lead to compatibility problems and high-dimensional redundancy, decision-level fusion is considered inflexible [2]. Score-level fusion, leveraging discriminative power between genuine and imposter scores, strikes a balance between data combination ease and information content, making it a favored approach for integrating biometric data [3].

Facial and speech biometrics are two key branches of biometric authentication, used in security, access control, and identity verification systems. Facial biometrics, known as facial recognition, relies on unique facial features like the arrangement of facial landmarks to establish identity [5]. It is non-intrusive, requiring no physical contact, and is employed in various applications, from security systems to smartphone unlocking. Advancements in deep learning and artificial intelligence have significantly improved its accuracy and performance [7]. However, challenges in facial recognition include variations in lighting, pose, expressions, and potential privacy concerns [8].

Speech biometrics, or voice biometrics, identifies and verifies individuals based on unique voice patterns, including pitch, tone, cadence, and spectral features [4]. These characteristics remain stable over time, captured during speech interactions. Voice recordings create voiceprints, used in applications like call center authentication and forensic voice analysis. Challenges include the need for high-quality voice recordings, susceptibility to environmental noise, and potential voice mimicry [9] [10]. Both facial and speech biometrics offer user-friendly and secure authentication methods, yet they also raise ethical and privacy concerns due to the sensitivity of biometric data. Consequently, the development and deployment of these technologies should be accompanied by safeguards and regulations to ensure the protection of individuals' privacy and rights.

This study aims to address challenges by reevaluating the underutilized XM2VTS biometrics database, known for its valuable features and classifications in facial recognition and related fields [11]. This publicly available database contains a wide range of facial images and metadata, making it a crucial resource for benchmarking facial recognition algorithms. Originally an extension of the M2VTS database, it was developed by the University of Surrey, UK, and is widely recognized in the research community [12].

The XM2VTS (Multi-Modal Verification for Teleservices and Security) database includes a diverse collection of facial images with variations in lighting, expressions, and poses. This diversity allows researchers to assess the performance of facial recognition algorithms in real-world scenarios. Each image and video sequence in the database is meticulously annotated with subject information, facial landmarks, and other relevant metadata [13]. This rich dataset provides an opportunity for experimenting with merging methods to enhance results and reduce fraud rates in biometric authentication.

This work is divided into three main sections. Section 1 explains the XM2VTS database and the Lausanne Protocols. Section 2 presents the proposed fusion protocols based on score-level fusion using T-norms algorithms, along with comparisons. In Section 3, we discuss the results and future perspectives of this research.

II. THE XM2VTS DATABASE AND PROTOCOLS

The XM2VTS dataset, as described in reference [14], encompasses synchronized video and speech data derived from 295 subjects. These recordings were conducted over four sessions separated by one-month intervals. During each session, two recordings were produced, consisting of both speech and headshot footage. The dataset is organized into three distinct sets:

- Training set (LP Train): Served as the foundation for constructing client models.
- Evaluation set (LP Eval): Played a pivotal role in determining decision thresholds and various hyper-parameters employed by classifiers.
- Test set (LP Test): Was utilized to assess performance levels.

Out of the total 295 individuals in the study, they were categorized into three different groups. One group consisted

of 200 clients, 25 evaluation impostors, and the third had 70 impostors meant for the testing phase. The study used two distinct approaches for dividing subjects into training and evaluation groups, known as Lausanne Protocol I and II.

This paper focuses on explaining and analyzing the Lausanne protocols I and II [14]. The following section will provide detailed information about these platforms, including their features, classification methods, and the entire system, which involves a combination of a feature type and a classifier.

A. Face and Speech Analysis Features

The study employs two sets of baseline features for biometric analysis. For facial analysis, these features consist of the Facial Histogram (FH), Discrete Cosine Transform for small images (DCTs), and Discrete Cosine Transform for large images (DCT_b). In the realm of speech analysis, the selected features include Linear Frequency Cepstral Coefficients (LFCC) and Perceptual Linear Predictive Analysis Cepstral Coefficients (PAC-MFCC). These features play a crucial role in the biometric identification process and are used to analyze both facial and speech data for the purpose of identity verification.

B. Classifiers

Two different types of classifiers were employed in these experiments: Multi-Layer Perceptron's (MLPs) and a Bayesian Classifier utilizing Gaussian Mixture Models (GMMs). In theory, both classifiers could be trained using any of the previously defined feature sets; however, in practice, MLPs excel at matching feature vectors of fixed size, while GMMs are better suited for matching sequences (feature vectors of varying sizes). Regardless of the chosen classifier, the hyper parameters (e.g., the number of hidden units for MLPs or the number of Gaussian components for GMMs) were fine-tuned on the evaluation set LP1 Eval.

When these experts are combined, they yield a total of 15 distinct biometric systems in Protocol I and 9 in Protocol II, each employing a unique approach to recognizing individuals. These systems leverage various modalities, indicating that they may rely on different types of information encoded within the feature-classifier pair.

III. Proposed Score-Level Fusion Method

In this study, our objective was to build upon the well-established XM2VTS database. Our approach involved a thorough exploration of this database, followed by the application of techniques aimed at combining information at the score level, inspired by concepts like T-Norms [14]. After implementing these methods, we proceeded to compare our results with those from previous research [15].

A. Preliminaries of t-norms

T-norms are kinds of binary functions, which generalize the intersection at the fuzzy sets. A t-norm is a function: $T[0,1] \times [0,1] \rightarrow [0,1]$ that satisfies the following properties:

- Commutativity: $T(x,y) = T(y,x)$
- Associativity: $T(x(y,z)) = T((x,y)z$
- Monotonicity: $T(x,y) \leq T(x,z)$ if$y \leq z$.
- Identity element: $T(x,1) = x$

B. Score-level fusion method based on

Before entering the fusion phase, the scores from each system were normalized using the Min-Max normalization, this method normalizes the raw scores (T_i) while maintaining their distributions to a scale factor close and transforms all scores in the range [0,1] according to:

$$T_{iNom} = \frac{T_i - T_{min}}{T_{max} - T_{min}} \quad (1)$$

In this paper, we have used some t-norms examples represented in table 1.

TABLE I
EXAMPLES OF T-NORMS USED

$T_i(x,y)$	***Property***
$T_1(x,y) = xy$	associative
$T_2(x,y) = x + y - xy$	non-associative
$T_3(x,y) = (\frac{xy}{x+y-xy})$	non-associative
$T_4(x,y) = \min(x,y)$	mean
$T_5(x,y) = \max(x,y)$	mean

IV. Evaluation Criteria

Due to the accept-reject outcomes, the biometric system may make two types of errors: false acceptance (FA) and false rejection (FR). These errors are defined as follows: (FAR=FA/NI), and (FRR=FR/NC).

Where FA and FR count the number of FA and FR accesses, respectively and NI and NC are the total number of impostor and client accesses, respectively. HTER stands for Half Total Error Rate, which is another performance metric commonly used in biometric systems. It represents the average of the FAR and FRR at the EER threshold.

$$HTER = \frac{FAR+FRR}{2} \quad (2)$$

When score distributions overlap, the Equal Error Rate (EER) is the point where False Acceptance Rate (FAR) equals False Rejection Rate (FRR). The choice of the threshold value in applications like biometric recognition or data classification is crucial, as it acts as a decision boundary. The threshold value depends on the application's goals, such as application type, FAR, FRR, performance metrics, and data distribution. There is no universal threshold value; it must be determined through experimentation and evaluation to strike the right balance between false positives and false negatives based on the specific context and objectives.

V. Experimental Results

This section unveils the findings from our research on multimodal biometrics, with a particular emphasis on score-level fusion through T-Norms. We used the T-Norm functions detailed in Table 2 for this purpose. After running our evaluation algorithms, we carried out recognition tests to gauge the system's effectiveness, offering valuable insights into how the system performs in various situations.

The table provided below displays the HTER (Half Total Error Rate) values achieved by combining two biometric characteristics, namely, facial and vocal features, at the score level. This combination was performed using five different T-Norm relationships. We will subsequently compare these

results with the findings of a prior study [15] that utilized mean relationships during the merging process.

A. Score level results for Lausanne Protocol I

In the first protocol divisions, there are 15 distinct binary combinations for face and voice. The Half Total Error Rate (HTER) values for these combinations ranged from a minimum of 0.533, associated with (DCTb, GMM) (LFCC, GMM), to a maximum of 4.225 for (PAC, GMM) (SSC, GMM) [14]. To improve HTER values, various methodologies were applied, with the best result achieved at HTER=0.505 using the T_2 algorithm in the combination (DCTb, GMM) (LFCC, GMM). These values showed variability across experiments. Additionally, comparisons were made with alternative techniques and previous research [14] as showing in Table 2. Referring to the table, the optimal match is ((DCTb, GMM) (LFCC, GMM)), which achieves the lowest HTER value. A lower HTER indicates a more secure system with a reduced error rate.

TABLE II
LAUSANNE PROTOCOL I AND BASELINE SYSTEM DESCRIPTION

Fusion candidates	***HTFR***					
	Mean [14]	$T_1(x,y)$	$T_2(x,y)$	$T_3(x,y)$	$T_4(x,y)$	$T_5(x,y)$
((FH, MLP) (LFCC, GMM))	0.785	1.559	0.781	1.718	1.881	0.711
((FH, MLP) (PAC, GMM))	1.133	1.663	1.179	1.733	1.883	1.178
((FH, MLP) (SSC, GMM))	0.868	1.696	0.763	1.748	1.883	0.825
((DCTs, GMM) (LFCC, GMM))	0.526	0.606	0.523	0.599	1.055	2.028
((DCTs, GMM) (PAC, GMM))	1.436	1.419	1.409	1.429	3.531	2.670
((DCTs, GMM) (SSC, GMM))	1.144	1.280	1.150	1.613	3.393	2.630
((DCT$_b$, GMM) (LFCC, GMM))	**0.553**	**0.525**	**0.505**	**0.590**	**0.941**	**0.703**
((DCT$_b$, GMM) (PAC, GMM))	1.127	1.115	1.409	1.014	1.316	3.400
((DCT$_b$, GMM) (SSC, GMM))	0.747	0.751	0.827	0.767	1.586	3.200
((DCTs, MLP) (LFCC, GMM))	0.841	1.658	1.147	2.143	3.241	0.803
((DCTs, MLP) (PAC, GMM))	1.119	2.276	1.478	2.885	3.347	1.935
((DCTs, MLP) (SSC, GMM))	1.326	2.556	1.388	3.217	3.361	1.871
((DCT$_b$, MLP) (LFCC, GMM))	1.621	4.580	2.740	5.085	5.883	1.762
((DCT$_b$, MLP) (PAC, GMM))	3.653	4.968	2.905	5.536	6.221	3.889
((DCT$_b$, MLP) (SSC, GMM))	3.017	5.588	3.193	6.052	6.225	3.105
((FH, MLP) (DCTs, GMM))	1.280	1.728	1.361	1.756	1.882	2.074
((FH, MLP) (DCT$_b$, GMM))	1.122	1.737	1.179	1.745	1.882	1.528
((FH, MLP) (DCTs, MLP))	1.513	1.480	1.510	1.595	1.748	1.873
((FH, MLP) (DCT$_b$, MLP))	1.960	1.657	2.261	2.268	2.496	2.586
((LFCC, GMM) (SSC, GMM))	1.595	1.213	1.721	1.104	1.140	4.437
((PAC, GMM) (SSC, GMM))	4.225	4.794	4.457	4.820	6.096	4.129
((DCTs, GMM) (DCTs, MLP))	2.388	3.092	2.437	3.354	3.363	3.332
((DCT$_b$, GMM) (DCT$_b$, MLP))	3.063	5.365	3.738	5.662	6.196	2.559

To validate these results, we created plots illustrating the distributions of agent and imposter degrees, allowing us to evaluate the degree of overlap between them. Indeed, assessing the overlap between the client and imposter score distributions is crucial in evaluating the effectiveness of a biometric verification system. When the client and imposter score distributions do not overlap or overlap very little, it is a positive indicator of the system's performance in separating genuine clients from impostors, as shown in the figure below. The absence of or minimal overlap implies that genuine clients' scores are clearly distinct from those of impostors. As a result, the system is less likely to mistakenly accept impostors as genuine clients. This leads to a low FAR, which means that the system has a high level of security in terms of verifying genuine users. However, it's important to note that achieving absolutely no overlap in practice can be challenging, and there is often a trade-off between FAR and False Rejection Rate (FRR). Sometimes, setting the system's threshold to eliminate all overlap may result in a higher FRR, which could inconvenience genuine users. Therefore, it's essential to strike the right balance between security (FAR) and user convenience (FRR) based on the specific requirements and use cases of the system. Different applications may have different tolerance levels for security and user experience, and the system's parameters should be adjusted accordingly.

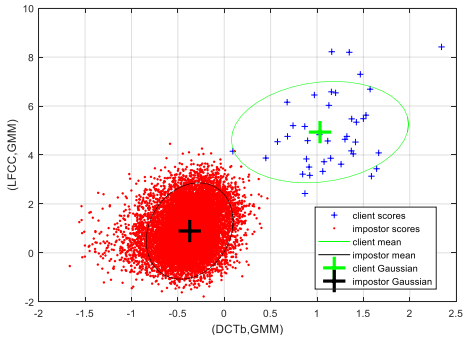

Fig.1. Distribution of client and imposter scores in the best matching

By examining this distribution, we gain valuable insights into how scores are spread across these different groups. This

visual representation aids in understanding the variations in scores and their implications for face and speech matching accuracy and security.

B. Score level results for Lausanne Protocol II

In a technical evaluation of a biometric system's performance, nine distinct combinations of face and voice features were tested in protocol II, with Half Total Error Rate (HTER) values ranging from 0.133 to 2.891 [14]. Various methods were applied to enhance HTER, with the best result achieved using the "T1 algorithm" in the combination (DCTb, GMM) (LFCC, GMM). Comparisons with alternative techniques were made, and the optimal match ((DCTb, GMM) (LFCC, GMM)) achieved the lowest HTER, indicating a more secure system with fewer errors as showing in Table. When examining the results, it is evident that the protocol II outperforms the protocol I. The second protocol consistently yields lower HTER values, which implies a more secure system with a reduced error rate. This outcome is highly desirable in biometric applications, as it means that the system is better at accurately verifying the identity of genuine users and minimizing the risk of granting access to imposters.

The reasons for this performance difference between the two protocols could be attributed to various factors, such as the choice of features, the data collection process, or the algorithms applied for matching. Careful analysis and understanding of these differences are crucial for improving biometric systems and enhancing their overall security and reliability. It's worth noting that the choice of the optimal protocol depends on the specific requirements and constraints of the application.

VI.TABLE III
LAUSANNE PROTOCOL II AND BASELINE SYSTEM DESCRIPTION

Fusion candidates	***HTFR***					
	Mean [14]	$T_1(x,y)$	$T_2(x,y)$	$T_3(x,y)$	$T_4(x,y)$	$T_5(x,y)$
((FH, MLP) (LFCC, GMM))	0.688	1.184	0.488	1.472	1.769	0.681
((FH, MLP) (PAC, GMM))	1.144	1.405	0.758	1.614	1.760	1.508
((FH, MLP) (SSC, GMM))	0.981	1.367	0.636	1.623	1.767	0.694
((DCTb, GMM) (LFCC, GMM))	**0.133**	**0.133**	**0.251**	**0.139**	**0.441**	**0.682**
((DCTb, GMM) (PAC, GMM))	1.175	0.395	2.229	0.439	0.681	5.260
((DCTb, GMM) (SSC, GMM))	0.177	0.273	1.070	0.294	0.695	3.393
((FH, MLP) (DCT_b, GMM))	0.962	1.475	0.990	1.627	1.771	0.900
((LFCC, GMM) (SSC, GMM))	0.828	0.819	1.986	0.824	1.365	3.393
((PAC, GMM) (SSC, GMM))	2.891	3.068	2.941	3.050	3.553	4.289

VII. CONLUSIONS

In conclusion, our paper has unveiled the untapped potential of the XM2VTS biometric database and introduced an exciting multimodal biometric system that employs innovative score-level fusion. The research findings not only hold promise but also pave the way for future algorithmic experiments geared toward bolstering the accuracy and resilience of biometric identification methods. Our overarching goal is to make the HTER value as low as possible compared to previous works and so elevate the precision and trustworthiness of biometric identification techniques to new heights.

REFERENCES

[1] A.K. Jain, A. Ross, S.Prabhakar,"An introduction to biometric recognition, " IEEE Trans. Circuit Syst. Video Technol, vol. 14, pp. 4–20, 2004.

[2] A.K. Jain, P. Flynn, A.A. Ross, "Handbook of Multibiometrics," Springer, NJ, USA,2008, pp. 1–22.

[3] K. Vishi, V.Mavroeidis, "An evaluation of score level fusion approaches for fingerprint and finger-vein biometrics," Proceedings of the 10th Norwegian Information Secuirity Conference, Oslo, Norway, pp. 27–29 November 2017.

[4] N. D. AL-Shakarchy, H.K. Obayes, Z. Najm Abdullah, " Person identification based on voice biometric using deep neural network," International Journal of Information Technology, vol. 15, pp. 789–795, 2023.

[5] J. H. Ortega, J. Fierrez, A. Morales. J. Galbally,"Introduction to Presentation Attack Detection in Face Biometrics and Recent Advances," Advances in Computer Vision and Pattern Recognition book series (ACVPR), Handbook of Biometric Anti-Spoofing,2023, pp. 203–23024.

[6] A. A. Joseph, A. N. H. Lian, K. Kipli, "Person Verification Based on Multimodal Biometric Recognition," Science & Technology, Pertanika J. Sci. & Technol, vol. 30 (1), pp. 161–183, 2022.

[7] Y. S. Ismael, M. Y. Shakor, A. A. Peshraw, "Deep Learning Based Real-Time Face Recognition System," NeuroQuantology, vol. 20, pp. 7355–7366, 2004.

[8] J. M. D. Zuo, Kevin, J. M. D. Saun, J. M. D. Tomas, R. M.D. Christopher, "Facial Recognition Technology: A Primer for Plastic Surgeons," Journal of the American society of plastic, vol. 143(6), pp. 1298–1306, June 2019.

[9] A. Kassis, U. Hengartner, "Breaking Security-Critical Voice Authentication," IEEE Symposium on Security (2023).

[10]] I. Trancoso, F. Teixeira, C. Botelho & A. Abad, "Treating Speech as Personally Identifiable Information and Its Impact in Machine Translation," Machine Translation: Technologies and Applications book series (MATRA), vol. 4, pp. 215–233, 2022.

[11] A. Fathima, K. Vaidehi, "Review on Facial Expression Recognition System Using Machine Learning Techniques," Advances in Decision Sciences. Image Processing, Security and Computer Vision, 2019, pp. 608–618.

[12] Smith, J. et al, "Advancements in Facial Biometrics: The M2VTS Database and Its Significance.' Journal of Biometric Research," 2023, pp. 123–135.

[13] D. Sadhya, S. K. Singh, "A comprehensive survey of unimodal facial databases in 2D and 3D domains," Elsevier, Neurocomputing, vol. 358, pp188–210, 2019.

[14] M.Cheniti, N. E. Boukezzoula1, Z. Akhtar, "Symmetric sum-based biometric score fusion," IET Biometrics, 2017, pp 2047-4938.

[15] N. Poh, S, Bengio, "Database, protocols and tools for evaluating score-level fusion algorithms in biometric authentication," Pattern Recognition, vol. 39, pp. 223–233, 2006.