Impact of Multimodal Face-Voice Integration and Feature Representation Type on the Performance of Authentication Recognition Systems

Lara Radovanovic

Introduction

Biometric recognition systems have become essential components in modern security infrastructure due to their reliability and usability. Among various biometric modalities, facial and voice recognition are particularly popular because of their non-intrusive nature and ease of data acquisition. However, systems relying solely on one modality, known as unimodal biometric systems, face significant challenges under real-world conditions. For example, facial recognition systems can be disrupted by lighting conditions or pose variations, while voice recognition systems are sensitive to background noise or changes in vocal tone [1]. To overcome these limitations, multimodal biometric systems combine two or more modalities to enhance overall recognition performance. By leveraging complementary biometric inputs, such as face and voice, a multimodal system can compensate for the weaknesses of each individual modality, leading to improved accuracy. For instance, when facial data is degraded, clear voice data can still contribute to a successful recognition outcome, and vice versa. Score-level fusion, a common strategy in multimodal systems, combines the confidence scores of each modality after independent processing, resulting in a more reliable decision-making process [1].

Beyond the fusion of modalities, the representation of biometric features, particularly in facial recognition, plays a critical role in system accuracy. Feature representations such as Euclidean distance, relative distance, and geometric ratios are used to quantify spatial relationships between facial landmarks. The choice of representation can significantly impact recognition outcomes by influencing both accuracy and computational efficiency. This study explores how different feature representations affect the performance of facial recognition and how they contribute to the overall performance of a multimodal system.

Research Questions

1. Does using a multimodal approach using face and voice improve recognition performance over using facial or voice recognition alone?
2. How do various types of feature representations (Euclidean, and relative distance, and ratio,) affect the performance of the facial recognition system, and in turn improve multimodal system?

Hypothesis

1. Multimodal systems offer greater robustness by reducing the limitations of single-modality systems.
2. Certain representations, particularly Euclidean distances and ratios, produce superior accuracy and resilience.

Related Work

Multimodal biometric systems have received significant attention due to their ability to overcome limitations seen in unimodal systems, such as vulnerability to environmental conditions, spoofing attacks, and low-quality data. Combining face and voice modalities is a particularly promising approach because of their complementary nature. When one modality fails, the other can often still perform successfully. In their foundational study, Ross and Jain [2] demonstrated that multimodal systems using score-level fusion achieved greater recognition accuracy compared to unimodal systems. Their work established score-level fusion as a practical and effective technique that balances performance gains with implementation complexity. Further research has validated the benefits of combining face and voice biometrics.

Faundez-Zanuy [3] explored different fusion strategies and concluded that score-level fusion not only improved performance over feature-level fusion but also offered better adaptability in real-time systems. More recently, Alharbi and Alshanbari [4] proposed a multimodal biometric system that combines voice and face recognition using Gaussian Mixture Models (GMM) and FaceNet, respectively, with scorelevel fusion. Their results showed that this approach significantly reduced the equal error rate (EER), affirming the performance gains achievable through multimodal fusion. In parallel, the performance of facial recognition systems has been closely tied to the choice of feature representation techniques. Traditional metrics such as Euclidean distance remain widely used due to their simplicity and effectiveness in measuring spatial relationships between facial landmarks.

However, alternative representations such as relative distances and geometric ratios have shown promise in capturing invariant features, particularly when subjects present variations in pose or expression. A study by Yilmaz and Ozer [5] compared Euclidean and ratio-based representations, finding that combining them reduced false accept rates in constrained datasets. Moreover, the configuration and quantity of facial landmarks directly affect recognition outcomes. While systems like Dlib and MediaPipe extract between 68 to 468 facial points, studies such as those by Xie et al. [6] show that not all landmarks contribute equally to recognition accuracy. They advocate for dimensionality reduction and the strategic selection of landmark-based features to reduce computational overhead while preserving discriminative power.

Despite these advances, few studies have systematically examined the interaction between feature representation techniques and multimodal fusion strategies, particularly under real-world conditions. This research addresses that gap by evaluating how Euclidean distances, relative distances, and ratios contribute to the accuracy of facial recognition. It also investigates how these techniques influence the overall performance of a face-voice multimodal biometric system.

Methods

Our approach consists of distinct stages: dataset selection, landmark detection, feature extraction, dimensionality reduction, data normalization, classification, and performance evaluation.

Dataset Selection and Preparation

To ensure the inclusion of a varied and challenging set of facial images, we selected the Caltech Faces dataset (1999) [7], which comprises 450 images from 27 individuals. The dataset includes natural variations in expression, pose, and background, making it appropriate for testing face recognition algorithms under semi-controlled conditions. For the voice modality, we used the Common Voice Delta Segment [8], a multilingual, publicly available dataset developed by Mozilla, which contains segmented voice recordings from speakers of diverse demographic backgrounds.

Feature Extraction

Facial Landmark detections were performed using the Dlib 68-point shape predictor [9], which maps 68 anatomically relevant facial points covering key regions such as the jawline, nose, eyes, eyebrows, and lips. This Configuration was chosen for its balance between detail and computation cost. Each landmark is returned as a coordinate pair (x, y) and provides the foundation for the geometric feature execution process. For each image, the following features were extracted using a custom extract_face_features function:

Euclidean distances between landmark pairs:

$$d_E = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Relative difference between landmark pairs:

$$d_M = |x_2 - x_1| + |y_2 - y_1|$$

Ratios of x and y coordinates landmark pairs:

$$r = \frac{x_2 - x_1}{y_2 - y_1}$$

These features reflect both fine-grained and broader aspects of facial geometry, delivering a detailed representation of a face's structure.

Voice feature extraction was performed by calculating Mel-Frequency Cepstral Coefficients (MFCCs) for each clip and the coefficient's time-series by its mean and standard deviation. This yields a compact (N × 2·13) feature matrix that captures both the spectral envelope and its variability.

Feature Preprocessing

To enhance model accuracy and manage the challenges of high-dimensional facial data, preprocessing of the data needed to be handled. The two useful techniques were: feature scaling and dimensionality reduction. Feature scaling was carried out using scikit-learn's StandardScaler, which standardizes data by removing the mean and scaling it to unit variance. The transformation follows the formula:

$$z = \frac{x - \mu}{\sigma}$$

here z represents the standardized value, x is the original data point, μ is the mean, and σ is the standard deviation. This process ensures that no single feature dominates due to its scale, allowing all features to contribute equally to the model.

Principal Component Analysis (PCA) was employed to reduce the dimensionality of the extracted facial and voice features while retaining the most informative components. PCA works by identifying the directions (principal components) in which the data varies the most and projecting the original features onto these orthogonal axes. This transformation reduces redundancy among features and helps improve the efficiency and performance of classifiers by minimizing noise and computational overhead.

In this study, PCA was applied after initial feature extraction and prior to model training. The number of components retained was selected based on cumulative variance, ensuring that 95% of the original data variance was preserved. This step was particularly useful for the facial recognition pipeline, where geometric features often contain correlations that PCA effectively consolidates into a compact, discriminative representation [11].

$$x^{1} = \frac{x - \mu}{\sigma}$$

Experimental Procedure

The experimental procedure was designed to evaluate recognition performance across unimodal (face-only and voice-only) and multimodal (face + voice) systems, while also assessing how different combinations of facial features affect ultimodal system performance.

To ensure the validity and generalizability of the results, identity-level stratified sampling was used to divide the dataset. The filtered facial dataset, containing one-hot encoded identity labels (y_face_filtered), was split using scikit-learn's train_test_split() function. A 67% training and 33% testing partition was implemented with stratification to preserve class distribution, and a fixed random_state=42 ensured reproducibility. This approach ensured that each identity was proportionally represented in both sets, minimizing bias due to unbalanced sampling.

Baseline performance for each modality was first established. A One-vs-Rest K Nearest Neighbors (KNN) classifier was trained on face-only data using Euclidean features, and another KNN was trained separately on the voice data. The models were evaluated using Equal Error Rate (EER) and d-prime (d′) to measure verification accuracy. These unimodal results served as the control group for comparison against fusion-based systems. Next, we created a basic multimodal system that combined both facial and voice data. For this fusion model, separate KNN classifiers were trained for each modality, and their predicted class probabilities were averaged using score-level fusion. This combined output was then evaluated using a broader set of metrics including accuracy, EER, and d-prime. This phase tested the first hypothesis whether combining modalities improves performance over using face or voice alone.

To address the second research question investigating how the type and combination of facial features influenced recognition performance, we calculated several types of geometric characteristics derived from the 68 facial landmarks. In this phase, the voice feature set remained constant while we varied the facial feature inputs across seven configurations: Euclidean only, relative only, ratio only, Euclidean + relative, Euclidean + ratio, relative + ratio, and a combined set of all three types. Each configuration was evaluated using the same pipeline: the selected facial features were used to train the KNN classifier, fused with the voice classifier at the score level, and assessed using EER and d-prime. This allowed us to test whether incorporating more types of facial features leads to improved recognition accuracy.

Finally, we analyzed and compared results from all phases. We first compared the performance of the unimodal systems (face-only and voice-only) to the multimodal fusion results to evaluate whether combining modalities improved recognition performance. We then compared the results across all seven facial feature configurations within the multimodal system to identify which feature combination yielded the best results.

Results

This section presents the performance analysis of three biometric configurations:, face-only, voiceonly, and fused face + voice systems, using common evaluation metrics including Equal Error Rate (EER), Area Under the Curve (AUC), d-prime (d'), and operating-point metrics such as True Accept Rate (TAR), False Accept Rate (FAR), and False Reject Rate (FRR).

Impact of Modality Combination

As seen in figures 1 and 2, the face only recognition system performed well. At a low FAR of 1%, the TAR remained reasonably high (64.5%), indicating strong usability in low false-alarm applications. The clear separation between genuine and impostor scores in the score distribution and high AUC validate the face system's reliability. In contrast, the voice-only system underperformed significantly, as depicted in table 1. These results indicate near-random classification performance, with extremely poor separation between genuine and impostor scores. The ROC curve barely rises above the diagonal, and the DET curve remains far from the optimal region. This result draws attention to serious limitations in the voice data or its processing. The scores show how weak or misconfigured modalities can significantly weaken a system's performance, especially in multimodal score-level fusion scenarios.

The fused multimodal face + voice system offered better performance than voice alone and comparable results to face-only. As seen in the ROC curve, the purple line representing the fused system tracks closely with the blue face-only system, showing similarly strong TAR at low FAR. However, the DET curve in Image reveals that the fused system introduces a slightly higher FRR than the face-only configuration at the same FAR. The fused face + voice system was expected to leverage the strengths of each modality to improve performance. However, results showed that the fused system did not outperform the face-only system, and in some cases (e.g., TAR@0.01, FRR), slightly underperformed. What is surprising is finding that even with the extremely poor voice recognition performance, the fused scores performed close to just face alone. These findings illustrate a critical flaw in the score-level fusion: without quality gating or modality weighting, the underperforming modality (voice) weakens the overall system.

The voice-only system's high EER (0.5087) and low d-prime (0.01) indicate very poor separation capability, which can reduce the benefit of fusion unless quality-dependent weighting or modality reliability is used. It is important to note that these findings do not indicate that using multimodal fusion will lead to poor performance, but rather that there was something wrong with the code logic or dataset.

Despite expecting that the multimodal system would outperform all unimodal systems, it did not significantly surpass the face-only system, and in some cases (e.g., AUC and TAR@0.01), it slightly underperformed. This outcome likely results from the extremely poor performance of the voice modality, which introduced noise during score-level fusion.

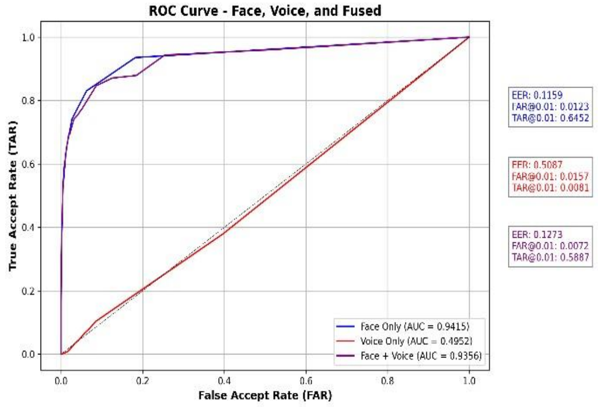

Figure1. ROC Curve Plot using Score Fusion but not Feature fusion

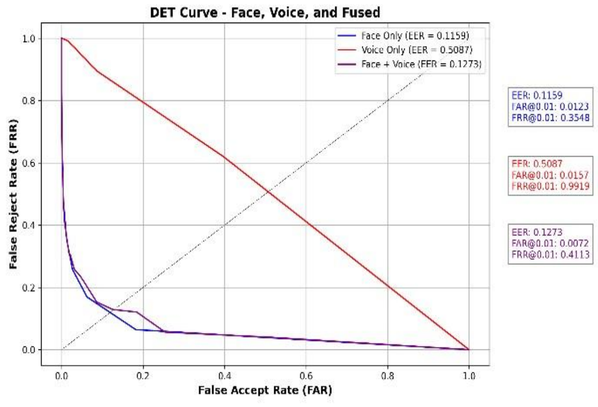

Figure 2.DET Curve plot using Score fusion but not Feature fusion

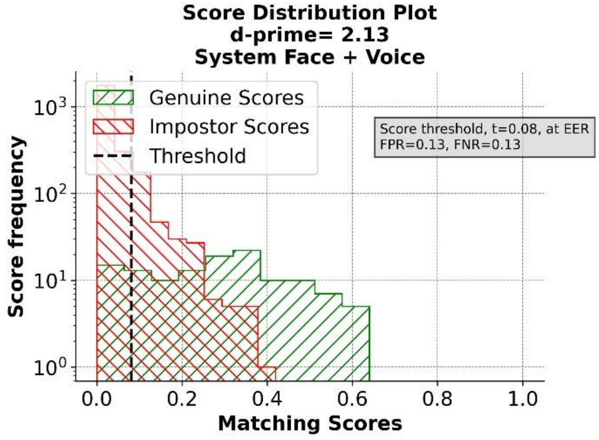

Figure 3. Score Distribution Plot using Score fusion, but not feature fusion

Impact of Feature Representations

This research investigated multiple types of feature representations and evaluated multiple combinations to identify the most effective combination. The three distinct types of features were: Euclidean distances, relative distances, and ratios between pairs of facial landmarks. Initially, each feature type was tested independently, without being combined with the others.

TABLE I. Decision parameters obtained from different feature representations rounded to two decimal places.

Feature representation	***d-prime value***	***AUC (in ROC curve)***
Euclidean distance	*2.29*	*0.94*
Relative distance	*2.33*	*0.95*
Ratio	*1.90*	*0.93*

Subsequently, feature representations were evaluated in pairs such as Euclidean distance with relative distance, Euclidean distance with ratio, and ratio with relative distance.

TABLE II. Decision parameters obtained from combinations of different feature representations rounded to two decimal places.

Feature representation	***d-prime value***	***AUC (in ROC curve)***
Euclidean + Relative	*2.49*	*0.96*
Euclidean + Ratio	*2.23*	*0.94*
Ratio + Relative	*2.14*	*0.94*
All Three Features	*2.29*	*0.95*

Comparing the individual feature representations with the combined ones, there is little difference in performance except for two representations. Ratio when combined with any of the other feature representations saw an increase in d-prime value of up to 0.33, and the combination of Euclidean + Relative reached the highest d-prime value of 2.49. Since “All Three Features” performed worse than Euclidean + Relative, the results suggest that Ratio reduces performance when included and that Euclidean + Relative is the optimal combination for this biometric system.

Impact of Score and Feature Level Fusion

The team's fusion strategy demonstrated significant impacts on the performance of the face-voice biometric authentication system, as evidenced by the comparison of results across different combinations of scorelevel and feature-level fusion:

- Score-level fusion with optimized facial feature inputs (Euclidean + Relative) achieved the best multimodal performance with a d-prime of 2.49 and an AUC of 0.96.
- Feature-level fusion (combining Euclidean, Relative, and Ratio features) without optimizing features slightly lowered performance, achieving a d-prime of 2.29.
- Using only feature-level fusion without score fusion (simply combining facial features without modality fusion) failed to significantly outperform the face-only system.
- The weakest performance occurred when indiscriminately combining all facial features (Euclidean + Relative + Ratio) and fusing with a poor-quality voice score, highlighting that both feature selection and score combination are critical.

These results indicate that score-level fusion played a crucial role in enhancing multimodal system performance, especially when supported by careful feature-level fusion for the face modality. The mean dprime across all tested fusion scenarios was approximately 2.20, with a range of 0.48 between the maximum and minimum scores.

Key Findings

Feature representations have a considerable effect on the performance of the facial recognition system, sometimes enhancing or weakening performance, as indicated by the d-prime value, in measures of roughly 0.30 in either direction. The results suggest that the Euclidean + Relative combination provides the optimal performance among the feature representations and thus proves the second hypothesis.

- Feature Fusion Impact: Carefully selected feature-level fusion (specifically Euclidean + Relative features) enhanced facial recognition and, in turn, boosted fusion performance. However, indiscriminate feature fusion (adding Ratio features) reduced system effectiveness.
- Fusion Efficacy: The strongest results were achieved when both optimized feature-level fusion and score-level fusion were applied together, underscoring the importance of feature quality prior to modality fusion. The results at face value indicate that using a multimodal approach weakens recognition performance over using facial recognition alone. However, the voice recognition system was never fully fleshed-out, therefore it cannot be concluded that this would be the case if this experiment were to be replicated.

Ethics Statement

The study used publicly available datasets (the Caltech Faces 1999 dataset and Mozilla's Common Voice Delta Segment) who were intended for research purposes with appropriate consent. Although informed consent was given, ethical concerns are not nullified as personal data such as race or sexual activity, which may be protected by the law, could be collected if the subject has voluntarily consented [12, p.444]. Because voice and face data qualify as permissible traits, collecting them may require specialized equipment and the physical presence of the person to harvest data [13]. The voice datasets don't explicitly mention if any one of these conditions were met, so it's possible the subject recorded the lines remotely. The possibility of a malicious actor probe for further personal information on subjects is possible but would prove difficult as no personally identifiable information is stated. However, it's possible for law enforcement investigations of minor offences to probe numerous databases of organizations who may collect biometric data for research purposes [14]. For example, Smith and Miller mention the company Clearview AI, that developed a biometric facial recognition algorithm to search images on the internet to identify suspects, indicating that law enforcement and sometimes even private companies may infringe user privacy by going through numerous databases that might include the data used in the study. This is an ethical issue because, despite working towards fighting crime, this action would involve innocent people and violate their right to privacy. Cases like these highlight the importance of upholding ethics when performing experiments that involve massive amounts of sensitive information.

Conclusion

This study compared unimodal and multimodal biometric systems using facial and voice data, along with various facial feature representations. The face-only model successfully achieved a high accuracy, but the voice-only model was inadequate due to our execution. Combining facial feature types enhanced recognition performance, and PCA effectively reduced complexity while preserving key information. The ROC, DET, and score distribution analyses revealed a possible mistake in the handling of voice recognition. These findings highlight the practical value of analyzing performance data and learning from the challenges that are faced when creating biometric authentication systems.

References

[1] A. K. Jain, A. Ross, and S. Pankanti, "Biometrics: A tool for information security," IEEE Transactions on Information Forensics and Security, vol. 1, no. 2, pp. 125–143, Jun. 2006, doi: 10.1109/TIFS.2006.873653.

[2] A. Ross and A. K. Jain, "Information fusion in biometrics," Pattern Recognition Letters, vol. 24, no. 13, pp. 2115–2125, Sep. 2003, doi: 10.1016/S0167-8655(03)00079-5.

[3] M. Faundez-Zanuy, "Biometric security technology: A review," IEEE Aerospace and Electronic Systems Magazine, vol. 21, no. 6, pp. 15–26, Jun. 2006, doi: 10.1109/MAES.2006.1667039.

[4] B. Alharbi and H. S. Alshanbari, "Face-voice based multimodal biometric authentication system via FaceNet and GMM," PeerJ Computer Science, vol. 9, p. e1468, Jul. 2023, doi: 10.7717/peerj-cs.1468.

[5] M. Yilmaz and B. Ozer, "Comparison of facial feature representations for robust face recognition," IEEE Access, vol. 8, pp. 14672–14681, Jan. 2020, doi: 10.1109/ACCESS.2020.2966931.

[6] Y. Xie, T. Li, and K. Wang, "Lightweight and discriminative facial landmarks for face recognition," in Proc. 2020 IEEE Intl Conf on Image Processing (ICIP), Abu Dhabi, UAE, Oct. 2020, pp. 2825–2829, doi: 10.1109/ICIP40778.2020.9191311.

[7] M. Weber, "Caltech Face Dataset 1999," CaltechDATA, 2022. https://data.caltech.edu/records/6rjahhdv18

[8] Mozilla Foundation, "Common Voice: Delta Segment," Mozilla Common Voice, 2024. Retrieved March 17, 2025 from https://commonvoice.mozilla.org/en/datasets

[9] A. Rosebrock, "Facial landmarks with dlib, OpenCV, and Python," PyImageSearch, 2017. https://pyimagesearch.com/2017/04/03/facial-landmarks-dlib-opencv-python/

[10] M. Afifi, "Optimizing Face Recognition with PCA and KNN: A Machine Learning Approach," ResearchGate, 2023. Retrieved April 20, 2025, from https://www.researchgate.net/publication/387406529_Optimizing_Face_Recognition_with_PCA_and _KNN_A_Machine_Learning_Approach

[11] I. T. Jolliffe and J. Cadima, “Principal component analysis: A review and recent developments,” Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 379, no. 2191, 2021, doi: 10.1098/rsta.2020.0161.

[12] A. North-Samardzic, “Biometric technology and ethics: Beyond security applications,” Journal of Business Ethics, vol. 167, no. 3, pp. 433–450, 2020.

[13] L. H. X. Ng, A. C. M. Lim, A. X. W. Lim, and A. Taeihagh, “Digital ethics for biometric applications in a smart city,” *Digital Government: Research and Practice*, vol. 4, no. 4, Art. no. 26, pp. 1–6, Dec. 2023.

[14] M. Smith and S. Miller, “The ethical application of biometric facial recognition technology,” *AI & Society*, vol. 37, no. 1, pp. 167–175, Mar. 2022.