Multimedia Systems (2026) 32:561
https://doi.org/10.1007/s00530-026-02647-6

REGULAR PAPER

Biovid dataset and challenge: biometric user recognition on smartphone data in real-world conditions

Georgia Fargetta[1] · Massimo Orazio Spata[1] · Alessandro Ortis[1] · Sebastiano Battiato[1] · Riccardo Puddu[2] · Doris Dearca[2] · Alessandro Saba[2] · Gianpaolo Perelli[2] · Simone Maurizio La Cava[2] · Marco Micheletto[2] · Giulia Orrù[2] · Gian Luca Marcialis[2] · Venu Siddapura Govindaraju[3] · Stefano Marrone[3] · Carlo Sansone[3]

Received: 14 April 2026 / Accepted: 22 August 2026 / Published online: 6 September 2026
© The Author(s) 2026

Abstract

User authentication is more and more crucial in everyday life, with researchers developing sophisticated, yet easy-to-use, approaches to enforce secure authentication while providing a smooth user experience. This is particularly true in mobile authentication, in which the spread of AI-capable smartphones is opening up several biometric-based authentication methods, such as face and lips recognition. A more recently emerging setup consists of the matching of a spoken passphrase with lips movements. Despite being promising and well-suited for mobility authentication, the matching operation is not trivial, and the research around this topic is limited, as no publicly available dataset is available for experimental purposes. In this paper we thus introduce *BioVid*, a novel privacy-aware multimodal biometric dataset for user recognition from smartphone-recorded videos in real-world conditions. *BioVid* consists of 650 recordings from 43 participants, where each clip captures a user pronouncing predefined passphrases while simultaneously acquiring audio and lips movements. The dataset is designed to support dual-factor biometric authentication, combining speaker-unique identity (UID) and passphrase verification, and explicitly models realistic variability in device type, recording distance, background, and ambient noise. Together with the data, we provide a detailed statistical analysis of the linguistic, visual, and acoustic properties of the data, highlighting its diversity and suitability for robust biometric system evaluation. Moreover, to test the quality of the collected dataset, we organised an international challenge on multimodal user authentication leveraging *BioVid*. In this paper we thus also describe the challenge protocol, baseline systems, and evaluation metrics, while also summarising the best-performing submitted approaches as measured by Equal Error Rate (EER). *BioVid* and the accompanying challenge aim to establish a reproducible benchmark for multimodal biometric authentication on consumer-grade smartphones and to foster further research on privacy-aware, real-world biometric systems.

Keywords Speaker recognition · Multimodal signal analysis · Biometric challenge · Biometric dataset

1 Introduction

The integration of audio and visual information has emerged as a promising direction to improve robustness and security in biometric user recognition [1]. Multimodal systems that jointly analyse speaker-specific vocal traits and lip movements can cope better with adverse conditions and reduce the risk of spoofing or replay attacks. Such a dual-factor audio–visual scheme is relevant to applications such as secure access control, forensic speaker verification, high-value financial authorization, smart devices and IoT interfaces, and online learning platforms. However, the development and fair comparison of such systems are limited by the scarcity of publicly available datasets that (i) Are acquired with consumer-grade smartphones, (ii) Include both audio and lip video of passphrase pronunciation, (iii) Explicitly model real-world variability in recording conditions, and (iv) Are accompanied by a standardized

Georgia Fargetta, Massimo Orazio Spata These authors have contributed equally to this work.

Communicated by Bing-kun Bao.

Extended author information available on the last page of the article

benchmark protocol. To address these limitations, we introduce *BioVid*, a multimodal biometric dataset of videos recorded with smartphones in realistic environments (Fig. 1). Each recording contains a user speaking predefined passphrases, enabling the extraction of linguistic information (words), acoustic features (voice), and lip-based visual cues. The dataset targets scenarios in which identity verification depends on both who the user is (UID) and what the user says (password), under open-set conditions where impostors and wrong passphrases must be rejected.

In addition to describing the dataset and its statistical properties, this paper introduces the first *BioVid* Challenge on multimodal authentication [2]. The competition, based on the *BioVid* dataset, invited participants to design audio–visual biometric systems that verify users from short smartphone videos by combining speaker identity and passphrase consistency. The submitted systems were evaluated under two authentication scenarios, one based solely on user identity (UID only) and one combining identity and passphrase (UID + password), using standard biometric error rates. The study reports the challenge protocol, unimodal and multimodal baselines, evaluation criteria, and the best-performing solutions, with particular emphasis on the impact of fusion strategy and passphrase handling on performance. A preliminary investigation of the audio subset from the same acquisition campaign was presented in [3] for closed-set speaker recognition. However, it assumes that all target speakers are known at training time, and it is restricted to acoustic features, without considering open-set operation or content-dependent constraints. The present work formalizes the full multimodal resource as the *BioVid* dataset, extends the focus from audio-only recognition to joint audio-visual processing, and introduces an open-set dual-factor authentication benchmark in which both impostors and wrong passphrases must be rejected.

As a result, the main contributions of this work consist in the following two points:

- We introduce the *BioVid* dataset, a publicly available privacy-aware multimodal biometric dataset acquired with heterogeneous consumer smartphones, featuring synchronised audio and lip video of spoken passphrases and explicitly designed for dual-factor (UID + passphrase) authentication under realistic recording conditions with variable distance, noise, and device characteristics. We also provide a detailed statistical characterisation of the dataset, covering word-length distributions, video duration, visual quality metrics, and representative audio features, in order to support both unimodal and multimodal system design and benchmarking;
- We introduce and report the results and the take-aways from the first *BioVid* Challenge on dual-factor biometric user authentication, with an open-set evaluation protocol, baseline systems, and standard performance metrics. We also report the approaches of the participants, together with the obtained results and considerations, showing how different multimodal fusion strategies impact performance in both UID-only and UID + password scenarios.

The remainder of this paper is organised as follows. Section 2 reviews related work on multimodal biometric authentication and smartphone-based datasets. Section 3 introduces the *BioVid* acquisition protocol and dataset characteristics. Section 4 presents the statistical analysis of linguistic, visual, and acoustic properties. Section 5 reports baseline experiments on *BioVid* in unimodal settings and

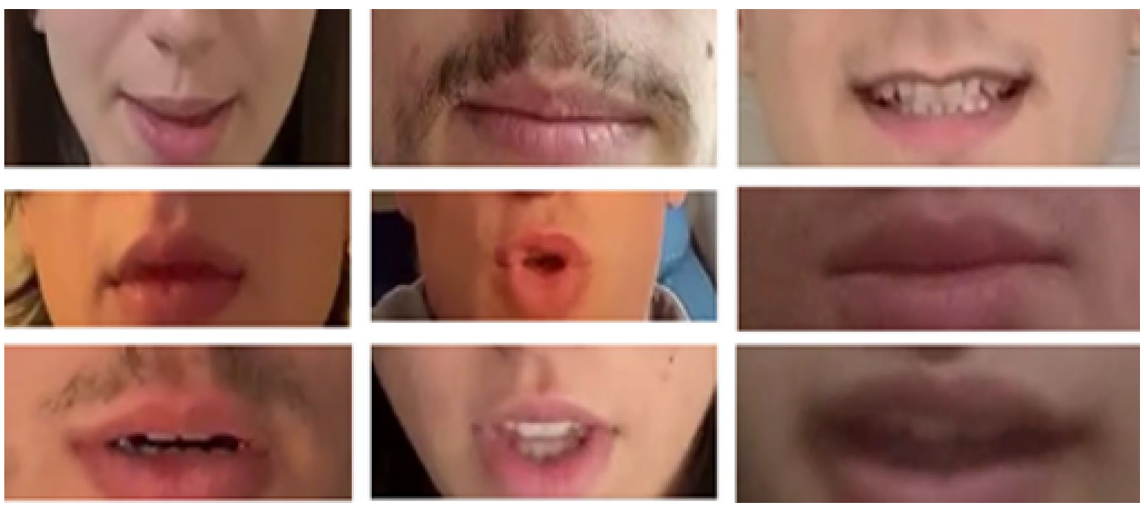


Fig. 1 Examples of the scenes contained in *BioVid* dataset. The acquisition has been made in real-world conditions, resulting in videos with non perfectly uniform illumination, potential lips misplacing, etc

includes reference evaluations with existing audio–visual architectures. Section 6 describes the *BioVid* Challenge setup, participating systems, and results. Section 7 concludes the paper and outlines future research directions.

2 Related work

Biometric user recognition has traditionally relied on unimodal systems based on either speech, voice, or facial appearance. Audio-based methods typically employ spectral representations such as MFCCs or spectrograms for speaker identity modelling [4]. Large-scale speaker verification pipelines such as VoxCeleb [5] have demonstrated the effectiveness of deep architectures under unconstrained conditions, while more recent works have explored multimodal extensions to improve robustness [6]. Visual biometrics, especially those based on facial regions or lip dynamics, have similarly benefited from deep convolutional models as feature extractors, including InceptionResNet-based architectures [7].

Beyond speaker verification, multimodal learning has been widely applied to a variety of security and perception tasks, including audiovisual speech recognition [8], event detection [9], speech enhancement [10], captioning [11, 12], user state classification [13], and behavioural biometrics for continuous authentication on smartphones [14]. Other works have explored gesture-based multimodal identification [15] and multimodal keystroke dynamics [16], further highlighting the growing interest in combining heterogeneous modalities for security applications. A recent related contribution is the work in [3], which proposed a Monte Carlo dropout–based CNN for closed-set speaker identification using smartphone-recorded audio under heterogeneous acquisition conditions. This study demonstrates the feasibility of robust speaker modelling from consumer-device audio alone, but it remains limited to a unimodal setting and a strict closed-set protocol. In contrast, the present work addresses a significantly more challenging problem involving multimodal fusion, lip-based features, passphrase-dependent verification, and open-set authentication within a standardized benchmark protocol. Existing multimodal datasets often rely on laboratory conditions, high-quality sensors, or scripted interactions, limiting their suitability for real-world biometric authentication. Similarly, smartphone-based datasets commonly used for behavioural or continuous authentication do not provide synchronised audio–visual speech acquisition nor controlled passphrase pronunciation.

To the best of our knowledge, no publicly documented dataset simultaneously provides (i) Synchronised audio and lip video of fixed passphrases, (ii) Acquisition exclusively on heterogeneous consumer smartphones across multiple devices, distances, and environments, (iii) Explicit wrong-passphrase recordings enabling dual-factor UID+passphrase verification, and (iv) An open-set benchmark protocol with multimodal challenge results, as summarised in Table 1. As described in the next sections, the *BioVid* dataset and the accompanying challenge introduced in this work address this gap by offering a privacy-aware, smartphone-recorded, multimodal resource tailored to both closed-set and open-set dual-factor authentication.

3 *BioVid* dataset

The *BioVid*[1] dataset consists of 650 short videos recorded from 43 users and covering 90 distinct Italian words, as summarised in Table 2. For 7 users, 26 videos are available, while the remaining 36 users contribute 13 videos each. Each clip is associated with a single spoken word and with the corresponding speaker identity. During acquisition, every participant repeated a predefined passphrase five times and uttered the remaining words twice each. This annotation scheme yields multiple instances per word and user and supports both user- and content-dependent evaluation protocols.

The 90-word vocabulary employed in this study constitutes a *shared pool* of words distributed across the 43 participants, rather than a set of words uniquely assigned to a single individual; that is, individual words from the pool are reused across different users. Each participant was asked to record only their own predefined passphrase, repeated five times, together with a limited, participant-specific subset of additional words drawn from the shared pool, each repeated twice. This acquisition protocol yields the per-user video counts summarized in Table 2. Specifically, the 36 participants who completed a single acquisition session each recorded 5 repetitions of their passphrase plus 4 additional words repeated twice, for a total of $5 + (2 \times 4) = 13$ videos per participant. The remaining 7 participants, who took part in an additional acquisition session, contributed proportionally more material, for a total of 26 videos each. The overall dataset volume is therefore given by

$$(36 \times 13) + (7 \times 26) = 468 + 182 = 650, \tag{1}$$

which matches the total number of videos reported for the dataset.

Recordings were acquired with the front cameras of heterogeneous consumer smartphones at a fixed resolution of 1080×1920 pixels. Three acquisition sessions were carried out in a standard indoor environment, with ambient noise

[1] *BioVid* dataset is available at https://www.kaggle.com/datasets/maxspata/biovid/data.

Table 1 Comparison between *BioVid* and representative public datasets for audio–visual speech recognition and biometric authentication

Dataset	Type	Media	PP	RW	#Sbj	Main task
AVLetters [17]	VSR	Video	✗	✗	10	Visual letter recognition
AVLetters2 [18]	VSR	Video	✗	✗	5	Visual letter recognition
CUAVE [19]	AVSR	Both	✗	✗	36	AV digit recognition
AVICAR [20]	AVSR	Both	✗	✓	100	In-car AV speech recognition
GRID [21]	AVSR	Both	✗	✗	34	Sentence-level AVSR
OuluVS [22]	VSR	Both	✗	✗	20	Phrase-level VSR
OuluVS2 [23]	VSR	Both	✗	✗	53	Multi-view VSR
LRW [24]	VSR	Both	✗	✓	> 1000	Word-level lip reading
LRS2 [25]	AVSR	Both	✗	✓	> 1000	Continuous AV speech
LRS3 [26]	AVSR	Both	✗	✓	> 4000	Continuous AV speech
MIRACL-VC1 [27]	VSR	Both	✗	✗	15	Word/sentence VSR
XM2VTSDB [28]	AV-SV	Both	✗	✗	295	AV person authentication
RSR2015 [29]	TD-SV	Audio	✓	✗	300	Text-dependent verification
FAVLIPS [30]	Lip-Password	Video	✓	✗	42	Lip-based biometric auth
MMGatorAuth [31]	Multimodal	Both	✓	✗	63	Multimodal biometric auth
Smartphone Auth. [14]	Continuous Auth	None	✗	✓	~ 50	Continuous smartphone auth
BioVid (ours)		Both	✓	✓	43	**Open-set UID + passphrase**

Type denotes the main application domain of each dataset: VSR = Visual Speech Recognition, AVSR = Audio–Visual Speech Recognition, TD-SV = Text-Dependent Speaker Verification, Multimodal Auth. = Multimodal biometric authentication, Continuous Auth. = Continuous authentication. Audio and Video (lips) indicate the availability of acoustic and lip-motion data, respectively. Passphrase (PP in the table header) specifies whether an explicit secret utterance is used as a knowledge-based authentication factor. Real-world (RW in the table header) denotes recordings performed under unconstrained or semi-unconstrained conditions. Please note that none of the reported datasets, except BioVid, neither use a consumer mobile device for the acquisition, nor support open-set biometric evaluation. #Sbj reports the number of enrolled subjects

Table 2 *BioVid* dataset acquisition and description

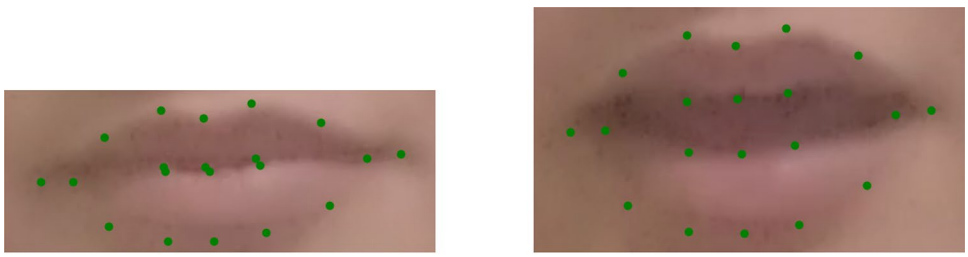
Description	Data
#users	43
#videos	650
#words	90
#videos with wrong passphrase	400
#videos with right passphrase	250
#different distances for recording	5
#locations for recording	21
#smartphones for recording	14

levels measured at approximately 18 dB, 19 dB, and 17.5 dB.[2] Participants held the phone at arm's length, at distances between 30 cm and 50 cm (30, 35, 40, 45, or 50 cm). For each video, a mouth-centred region of interest was automatically cropped to isolate the lip area for the visual analysis. This protocol introduces variability in device type, distance, background, and ambient noise that reflects typical handheld smartphone usage.

In addition to raw videos and audio tracks, the public release provides pre-computed lip-shape descriptors based on Shape Context [32]. Facial landmarks are first detected on each frame, and 20 key-points on the lip contour are retained for both closed-mouth and open-mouth configurations (see Fig. 2). For each frame, a Shape Context descriptor is computed by building a histogram of the relative positions of the contour points in a log-polar grid centred at each key-point. The resulting histograms form a compact representation of lip geometry that is approximately invariant to translation, scale, and in-plane rotation. The Kaggle repository includes a feature file containing Shape Context descriptors for the mouth region of every video, so that visual baselines can be reproduced without re-running the landmark detection pipeline.

Taken together, these characteristics define a multimodal resource specifically tailored to dual-factor authentication.

[2] For coherency reasons, these values have been masured by using a sound-mether app on the smartphone itself.



(a) Closed-mouth configuration (b) Open-mouth configuration

Fig. 2 Examples of facial landmarks detected on the mouth region under **a** closed-mouth and **b** open-mouth configurations

Synchronized speech and mouth-region video support both unimodal tasks (audio-only or visual-only user recognition) and joint audio–visual processing, while the acquisition protocol enforces realistic smartphone conditions and open-set operation. The dataset is therefore suited to the design and benchmarking of audio–visual biometric systems, and it provides the basis for the BioVid Challenge described in Sect. 6.

4 Statistical data analysis

This section summarizes the main statistical properties of the linguistic, visual, and acoustic data in the *BioVid* dataset.

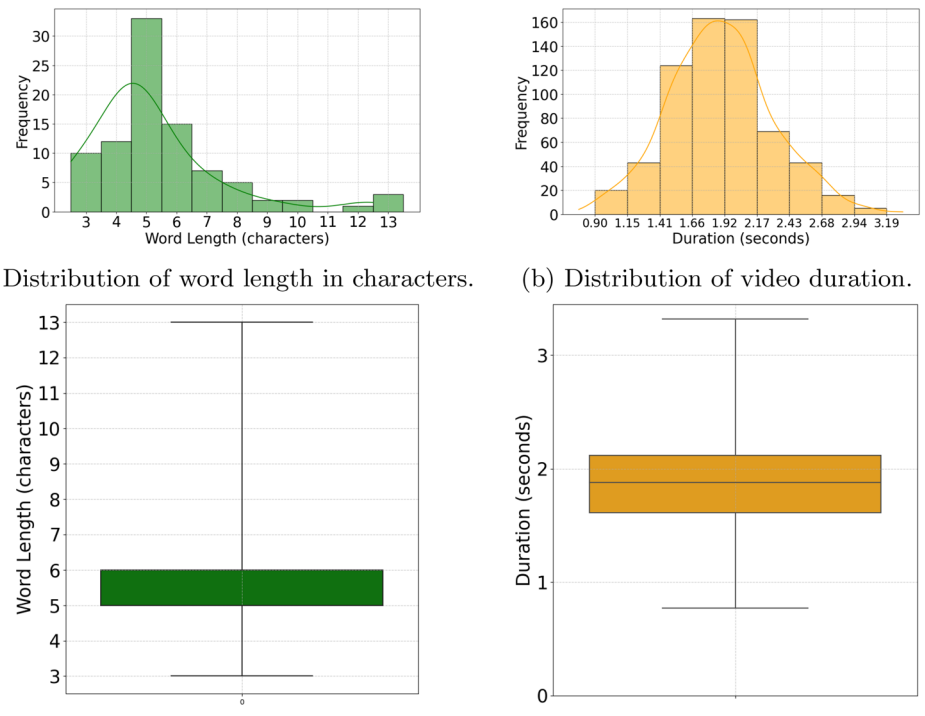
4.1 Word length and video duration

Table 2 presents the description of the acquisition counts, and Fig. 3 reports basic statistics for the 90 distinct Italian words and video duration included in *BioVid*. Word length ranges from 3 to 13 characters, with a mean of 5.68 and a standard deviation of 2.18. The first quartile, median, and third quartile are equal to 5, 5, and 6 characters, respectively, which indicates that most words are relatively short and concentrated in a narrow range around 5–6 characters. The distribution shown in Fig. 3a confirms that the large majority of words fall between 3 and 8 characters, while only a few reach 10 characters or more. The box-plot in Fig. 3c offers additional insights into the variability of word lengths and highlights a moderate asymmetry in the distribution, with a concentration of shorter words. Words exceeding 10 characters are identified as outliers within the distribution. Overall, the IQR and whiskers confirm that the dataset is centered on concise word lengths, while the outliers highlight rare but significant deviations.

Video duration exhibits a similar concentration around short values (see Fig. 3b,d). The shortest clips last approximately 0.9 s, while the longest reach about 3.5 s, with a median duration of 1.89 s. The inter-quartile range spans roughly from 1.66 s to 2.43 s, so half of the videos last between about 1.7 s and 2.4 s. The histogram in Fig. 3b shows that most recordings lie between 1.4 s and 2.2 s, whereas only a small fraction exceeds 2.9 s. In general, a clearer symmetry in the distribution of durations relative to the median is observed, compared to the analysis of word character counts. This indicates that the duration of a video showcasing the pronunciation of a word is not directly proportional to the number of characters in the word. Such a discrepancy can be attributed to both the opening and closing delays of the video and the differences between phonetic duration and the length of the corresponding word.

4.2 Audio-video quality metrics

Table 3 summarizes the visual quality metrics computed for each user in terms of mean and standard deviation of brightness, contrast, and sharpness. Sharpness is measured with three no-reference indicators: Laplacian variance,



(a) Distribution of word length in characters.

(b) Distribution of video duration.

(c) Box-plot of word length.

(d) Box-plot of video duration.

Fig. 3 Word length and video duration statistics for the *BioVid* dataset: distributions and box-plots

Table 3 Video-frame visual quality metrics across all users of the *BioVid* dataset

Visual quality metric	Mean ± Std
Brightness	111.318±18.327
Contrast	27.077±3.761
Laplacian	10.066±15.916
Tenengrad	535.804±717.539
CPBD	0.005 ± 0.002
S3	1.992 ± 0.261

Table 4 Audio quality metrics across all users of the *BioVid* dataset

Metric	Mean	Std	Min	Median	Max
Dynamic range	259.431	56.555	2.233	271.847	275.719
Spectral flatness	0.806	0.065	0.642	0.803	0.978

Table 5 Ordinary Least Squares (OLS) regression analysis of confidence score

Variable	Coef.	Std Err	t	p-value	95% CI
Intercept	1.0113	0.023	43.626	0.000	[0.965, 1.057]
Duration	−0.0101	0.011	−0.921	0.359	[−0.032, 0.012]
Word Length	0.0050	0.005	1.042	0.299	[−0.005, 0.015]
Syllables	−0.0144	0.012	−1.204	0.231	[−0.038, 0.009]
Model statistics:					
$R^2 = 0.018$ Adj. $R^2 = -0.005$ F = 0.776 p(F) = 0.510					
N = 130 AIC = −411.7 BIC = −400.3 Durbin–Watson = 1.98					

Cumulative Probability of Blur Detection (CPBD), and Spatial–Spectral Entropy-based Sharpness (S3) [33, 34]. Ten frames are uniformly sampled from each video, analyzed independently, and then averaged per user in order to obtain a representative estimate for each participant.

The reported values indicate a relatively consistent brightness and contrast across users, with moderate variability in sharpness. Such behaviour reflects the acquisition protocol, where participants record videos in similar indoor conditions with heterogeneous smartphones and small differences in focus and motion. The metrics provide a compact summary of visual quality that can be used to control for potential covariates in downstream experiments. Specifically, the analysis of video quality reveals a predominance of moderately blurred samples, as indicated by low median Laplacian values, with a small subset of sharp recordings acting as outliers. Brightness statistics show a balanced distribution consistent with indoor smartphone acquisition. The audio quality statistics in Table 4 indicate generally stable acoustic acquisition across the dataset. The dynamic range proxy shows high average values with limited dispersion, suggesting a consistent signal level across recordings, while a small number of low-range outliers reflect the presence of more challenging acoustic conditions. The spectral flatness proxy exhibits moderate values, indicating that background noise is present but relatively homogeneous across samples, which is consistent with real-world smartphone recordings.

The regression analysis in Table 5 was conducted to verify that the non-significant coefficients are consistent with the absence of shortcut effects within the scope of this diagnostic analysis, but they do not, by themselves, demonstrate that the model's confidence is fully independent of these covariates. A non-significant *p*-value indicates that no statistically detectable effect was found under the tested conditions and does not constitute proof of the null hypothesis.

We clarify that the "confidence score" analyzed in Table 5 corresponds to the posterior probability assigned by the classification model's softmax output layer to the correct/predicted identity for each trial. This score was used as the dependent variable in an Ordinary Least Squares (OLS) regression against three superficial linguistic and temporal covariates, namely utterance *Duration*, *Word Length*, and number of *Syllables*, in order to assess whether the model's confidence is systematically driven by these shortcut cues rather than by discriminative biometric information. We further clarify that the sample size $N = 130$ used in this diagnostic regression does not refer to the full dataset of 650 videos, but to the subset of non-passphrase word utterances drawn from the open-set test partition (Sect. 6), excluding passphrase repetitions in order to avoid confounding the analysis with the fixed passphrase content shared across trials.

As shown in Table 5, none of the three covariates reach statistical significance ($p = 0.359$ for Duration, $p = 0.299$ for Word Length, $p = 0.231$ for Syllables), and the overall model fit is weak ($R^2 = 0.018$, F-test $p = 0.510$).

The results show that video duration, word length, and number of syllables do not significantly predict the confidence output, with the overall model failing to reach statistical significance and explaining only a negligible fraction of the variance. None of the individual predictors is statistically significant, and all effect sizes are small. These findings indicate that the system does not rely on simple lexical complexity or clip duration as shortcut cues. Instead, the confidence estimates appear to be driven by identity-related audiovisual information rather than superficial linguistic attributes, mitigating concerns that performance might be biased by trivial word-level properties.

Specifically, the objective of this analysis was not to obtain a strong predictive regression model, but rather to verify that superficial linguistic and temporal factors (video duration, word length, and number of syllables) do not influence the model's confidence scores. From this perspective, the results are reassuring. The overall regression model is not statistically significant (Prob(F-statistic) $= 0.510$),

Table 6 Recording conditions by smartphone model, split into two distance ranges

Smartphone model	User IDs	Dist. range (cm)	Indoor	Outdoor
iPhone X	24	40	0	1
iPhone 11	43, 40, 39	23–29	3	0
	41, 21	38–50	0	2
iPhone 12	19, 30, 4	24–32	2	1
	22, 38	36–39	1	1
iPhone 13	14, 16	28–33	1	1
	2, 34	46–48	2	0
Samsung Galaxy S20	17, 15, 9	27–38	2	1
	5, 7	44–46	2	0
Samsung Galaxy S21	29	31	1	0
	27	38	1	0
Samsung Galaxy S22	37	26	1	0
	31	40	0	1
Google Pixel (5–8)	18, 32	28–33	1	1
	10, 6	35	2	0
Oppo	3, 42, 35	30–37	3	0
	13, 33	39–40	1	1
Redmi 13	11, 25	24–28	1	1
	20, 26	37–40	2	0
Redmi 14C	8, 36, 12	25–31	2	1
	28, 1, 23	34–50	3	0

indicating that the set of predictors does not explain the confidence output. Consistently, the coefficient of determination is extremely low ($R^2 = 0.018$; Adj. $R^2 = -0.005$), meaning that only 1.8% of the variance is accounted for and that the model does not improve over a constant baseline. Moreover, none of the individual predictors is statistically significant (duration: $p = 0.359$; word length: $p = 0.299$; syllables: $p = 0.231$), and their confidence intervals all include zero (e.g., duration: $[-0.032, 0.012]$), indicating the absence of measurable effects. Finally, the Durbin–Watson statistic (≈ 1.98) suggests no residual autocorrelation, confirming the reliability of the estimation. Overall, these results support the conclusion that the model's confidence is not driven by trivial lexical or temporal properties, but rather by identity-relevant audiovisual information.

4.3 Data bias analysis

In this paragraph, we conduct an in-depth analysis of the dataset to verify the absence of device-related or other potential biases. In principle, fixing the video resolution across heterogeneous devices (see Table 6) may disregard native sensor characteristics and introduce resampling artifacts. We acknowledge this potential limitation; however, device identifiers are not available in the public release of BioVid, preventing a direct device-level analysis. To address this concern, we conducted indirect robustness checks based on low-level visual acquisition statistics. In particular, we analyzed blur and brightness distributions (see Table 3) across the dataset, obtaining a Laplacian variance of 10.07 ± 15.92 (median $= 2.54$) and a Tenengrad sharpness of 535.80 ± 717.54, indicating substantial variability in visual sharpness with only a limited number of high-quality outliers. Brightness values were more stable, with a mean of 111.32 ± 18.33 over a wide range $[62.09,\ 153.44]$, confirming heterogeneous illumination conditions typical of uncontrolled smartphone recordings. Despite this variability, no strong correlation was observed between model confidence and these acquisition-related statistics, suggesting that performance is not driven by low-level device or resampling cues. Moreover, inference tests using different interpolation strategies (bilinear, bicubic, and area) yielded stable performance. Overall, these results indicate that the model does not rely on device-specific resizing artifacts, but rather on identity-related visual information.

4.4 Audio features

Figure 4 illustrates the temporal evolution of three Mel-Frequency Cepstral Coefficients (MFCCs) extracted from a representative subset of *BioVid* audio samples. MFCC1, MFCC2, and MFCC10 are considered as examples of coefficients that are often associated with overall energy, spectral shape, and finer spectral details, respectively. In the analyzed data, MFCC1 exhibits pronounced variations across time,

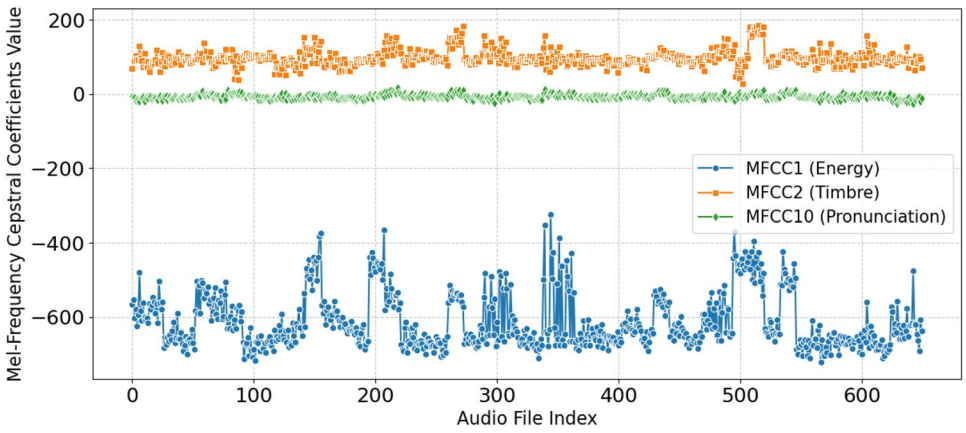


Fig. 4 Example of Mel-Frequency Cepstral Coefficients (MFCC) for three selected dimensions in *BioVid*: MFCC1, MFCC2, and MFCC10

with peaks and troughs that reflect changes in loudness and articulation. MFCC2 shows moderate variability, whereas MFCC10 appears comparatively more stable under the specific recording conditions and microphones used in *BioVid*. This observation is empirical and is not intended as a phonetic or linguistic interpretation of individual coefficients.

Mel-spectrograms form the standard time–frequency representation used in the audio baselines described in Sect. 5. An example is shown in Fig. 5, where short-time spectral energy is projected onto the mel scale. This representation preserves discriminative acoustic cues under moderate variations in pitch and speaking rate and provides a convenient input format for convolutional architectures in the unimodal audio setting.

5 Closed-set baseline experiments on *BioVid*

The experiments in this section evaluate a closed-set identification task. The 10-fold stratified split is performed at the video level and stratifies by user identity; consequently, the training and test partitions contain samples from the same enrolled identities and are not user-disjoint. Furthermore, the *BioVid* dataset naturally supports two complementary problem settings. At the unimodal level, each clip can be used for user recognition based either on voice cues in the audio stream or on lip-motion cues (video). At the multimodal level, the synchronized audio and mouth-region video enable biometric identification and authentication systems that combine voice and lip dynamics. In this section, we perform an exploratory baseline analysis, aimed at providing an initial quantitative characterization of the *BioVid* dataset before moving to the challenge results. In particular, we focus on an exploratory closed-set identification setting, where all test samples belong to the enrolled set and performance is reported as classification accuracy/Top-1. In this case, the reported *Accuracy* corresponds to the *Top-1 identification accuracy*, i.e., the percentage of probe samples whose predicted identity matches the ground-truth identity among the enrolled classes. In this setting, systems must discriminate between genuine and impostor access attempts, and performance is primarily assessed through threshold-based metrics such as FAR, FRR, and Equal Error Rate (EER), computed by varying the decision threshold on the similarity scores. In some implementations, the identity associated with the maximum score is first selected and subsequently validated against a decision threshold, which results in a final accept/reject decision. This clarification explains why identification-oriented metrics (e.g., Top-1 accuracy) are reported for the baseline experiments, while verification-oriented metrics (e.g., EER, FAR, and FRR) are used for the challenge evaluation.

From an audio perspective, mel-spectrograms are adopted as a standard input representation for subsequent experiments (see Fig. 5). Mel-spectrograms encode the short-time power spectrum on a mel-scaled frequency axis and capture spectral and temporal patterns that are relevant for spoken passphrases and speaker identity. This representation is widely used in speech and speaker recognition, since it preserves discriminative cues under moderate variations in pitch, speaking rate, and background noise. For the visual unimodal baselines, Shape Context features [32]

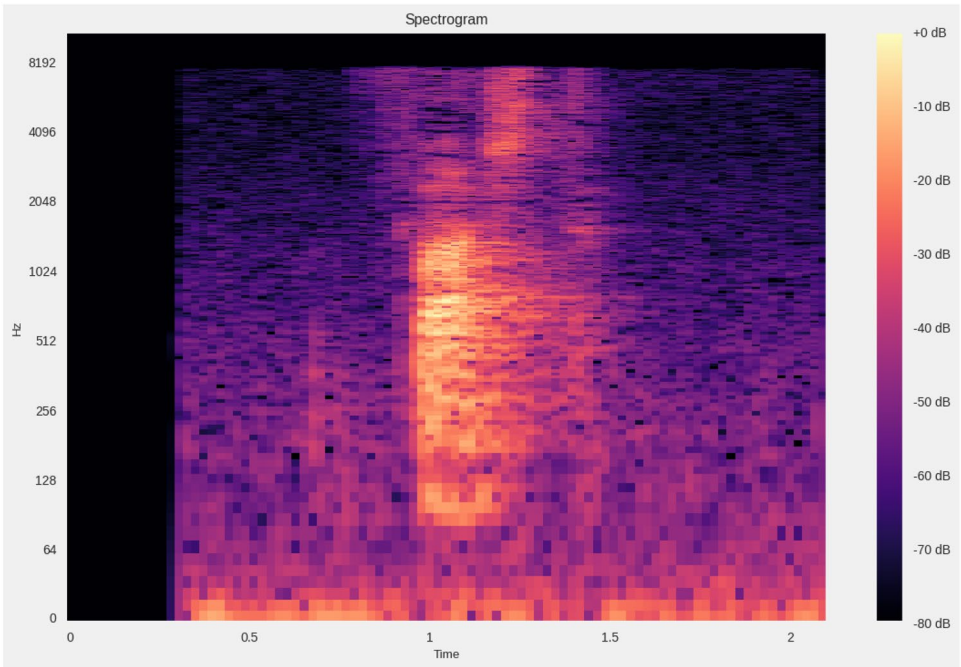


Fig. 5 Example of a mel-spectrogram computed from a *BioVid* audio sample and used as input representation in the unimodal speaker-identification baselines

Table 7 Baseline performance of different unimodal classification models on the *BioVid* dataset using shape context features extracted from lip key-points. The models are evaluated with a 10-fold stratified cross-validation

Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (s)
ETC	91.98%	0.9976	91.98%	92.13%	91.88%	0.9175	0.9176	16.2240
LR	91.73%	–	91.73%	91.80%	91.70%	0.9149	0.9150	177.8820
MLP	91.42%	0.9980	91.42%	91.67%	91.39%	0.9118	0.9119	123.5960
RFC	91.14%	0.9971	91.14%	91.30%	91.01%	0.9089	0.9090	11.2250
SVM	87.39%	–	87.39%	89.27%	87.53%	0.8702	0.8708	56.3430
KNC	84.54%	0.9711	84.54%	84.88%	84.28%	0.8410	0.8412	7.9600
LDA	83.24%	–	83.24%	83.81%	83.20%	0.8277	0.8278	6.5140
RC	81.40%	–	81.40%	82.48%	80.75%	0.8085	0.8090	1.8590
DT	75.69%	0.8751	75.69%	75.88%	75.63%	0.7500	0.7501	3.3450
NB	45.97%	0.8695	45.97%	56.38%	42.88%	0.4478	0.4520	2.6970
QDA	18.10%	–	18.10%	44.20%	17.84%	0.1634	0.1999	13.2080
ADA	18.09%	–	18.09%	12.38%	10.84%	0.1461	0.1629	9.6270

are extracted from 20 lip key-points to obtain a compact description of mouth geometry over time. Facial landmarks are detected on each frame and 20 points on the lip contour are retained, as described in Sect. 3. The corresponding Shape Context descriptors provide a fixed-length representation for each clip and are used as input to a set of standard classification models.

Table 7 reports the performance of a set of standard classification models applied to Shape Context features: Multi-Layer Perceptron (MLP), ADA Boost Classifier, Decision Tree Classifier (DT), Random Forest Classifier (RFC), Extra Trees Classifier (ETC), Ridge Classifier (RC), Linear Discriminant Analysis (LDA), Logistic Regression (LR), Support Vector Machine (SVM), K-Neighbors Classifier (KNC), Naïve Bayes (NB), and Quadratic Discriminant Analysis (QDA). Models are evaluated with 10-fold stratified cross-validation on the 70% training portion of the dataset, while a separate 30% stratified hold-out set is kept aside. The preprocessing pipeline applies automatic type inference and mean imputation for numerical features, without additional scaling, dimensionality reduction, or feature engineering, in order to preserve a simple and reproducible experimental setting.

It is important to clarify the scope of the data split adopted for the baseline experiments presented in this section. The task addressed here is a *closed-set identification* problem, in which, by design, every enrolled identity (all 43 users) must be represented in both the training and test partitions. Consequently, the "stratified" split reported in Table 7 stratifies by class label (i.e., user identity) at the video level, ensuring balanced class representation across folds; it is not intended to be user-disjoint. This property is instead a distinguishing feature of the open-set verification protocol adopted for the challenge, detailed in Sect. 6. We highlight this distinction explicitly to avoid any ambiguity when comparing the closed-set baseline accuracy reported below with the open-set challenge results, as the two protocols target fundamentally different recognition tasks and are therefore not directly comparable.

Table 8 Baseline performance of existing audio and audio–visual speaker identification architectures evaluated on the *BioVid* dataset, reported in terms of Top-1 accuracy

Method	Top-1%
I-vectors + PLDA + SVM [35]	58.70%
CNN [35]	77.30%
Network A [35]	81.90%
Network C [6]	85.00%
Network B [35]	85.80%
VGGVox Baseline [35]	88.00%

Among all models, ETC achieves the highest accuracy (91.98%), together with strong AUC, F1-score, Cohen's kappa, and Matthews Correlation Coefficient values. These results indicate that shape-context features derived from lip key-points are highly informative for user identification on *BioVid* and that ETC can be regarded as a reference visual baseline for future work.

To complement the frame-wise analysis, a time-series baseline based on a Temporal Convolutional Network (TCN) [16] is also considered. In this configuration, shape-context features extracted from consecutive frames form a sequence that is first processed by a 3D convolutional layer (Conv3D) to capture joint spatial and short-term temporal patterns. The resulting representation is then fed to a stack of causal and dilated convolutions, as in standard TCN architectures, to model longer-range temporal dependencies in lip motion. Experimental results yield a weighted average accuracy of 76% and F1-score and recall equal to 74%, which confirms that a temporal model can exploit dynamic information from lip movements, although static shape-context features combined with ETC remain the strongest visual baseline in this study.

From a multimodal perspective, Table 8 summarises the Top-1 accuracy obtained on *BioVid* by a set of audio–visual speaker identification architectures originally proposed in

[6, 35]. In the present evaluation, all configurations operate in an audio–visual setting and jointly process the speech signal and the synchronised mouth-region video extracted from *BioVid* clips. I-vectors with PLDA combined with a shallow SVM classifier reach 58.70% Top-1 accuracy, while a convolutional network trained on spectral audio representations attains 77.30%. Deep architectures tailored to speaker recognition further improve performance: Network A and Network B from [35] achieve 81.90% and 85.80% , respectively, and the VGGVox baseline reaches 88.00%. Network C [6], which explicitly fuses video-based features with a Mixture-of-Experts audio front-end, attains 85.00% Top-1 accuracy. Overall, results highlight the advantage of deep audio–visual architectures over traditional i-vector-based systems on *BioVid* and provide multimodal reference points that complement the unimodal baselines described in this section. While Section 5 reports closed-set identification baselines, Section 6 moves to an open-set biometric authentication protocol, where the training and test identities are disjoint and unseen users must be rejected.

6 Biometric user identification open-set challenge

Unlike the closed-set baseline experiments of Section 5, the challenge protocol in this section adopts an open-set setting with disjoint train/test splits and previously unseen users in the test set. To provide a first open-set benchmark on *BioVid*, a dedicated challenge was organized and hosted at a major international conference in the field.[3] The objective was to assess and compare different strategies for audio–visual biometric user authentication on the *BioVid* dataset. Participating teams were asked to develop multimodal systems that verify a user's identity and passphrase from synchronized video and audio recordings acquired with smartphones, with emphasis on robustness under realistic conditions.

6.1 Challenge description

The *BioVid* Challenge focuses on dual-factor biometric authentication from videos in an open-set scenario. To ensure full transparency and reproducibility of the reported results, we detail here the complete specification of the open-set evaluation protocol adopted in the BioVid Challenge (ICIAP 2025).[4] The training and test releases are constructed from completely disjoint recordings, i.e., no video is shared between the two splits. Furthermore, the protocol evaluates an open-set verification scenario, in which entirely unseen identities are introduced in the test set alongside the identities enrolled during training (Table 9).

The training set comprises 457 videos in total, covering 38 enrolled identities: 7 users contribute 8 genuine authentication attempts (correct passphrase) and 13 wrong attempts (incorrect passphrase) each, while the remaining 31 users contribute 4 genuine and 6 wrong attempts each. The test set comprises 193 videos in total, and includes the same 38 enrolled identities used to evaluate verification performance, plus 5 entirely new, unseen (non-enrolled) identities, introduced to simulate and evaluate open-set impostor rejection. Within the test set, the 38 enrolled identities contribute 5 videos each (7 users with 2 genuine and 3 wrong attempts; 31 users with 1 genuine and 2 wrong attempts), while the 5 unseen identities contribute additional videos used exclusively as impostor trials, without any enrolled template. Submitted systems must jointly verify user identity and spoken passphrase based on audio and lip-motion information extracted from each clip. Evaluation is based on the EER computed on a held-out test set. We clarify that the reported EER was obtained via a post-hoc sweep of the decision threshold $\tau \in [0, 1]$, with a step size of 0.02, performed directly on the test-set similarity scores. No separate validation partition was used for threshold tuning. Two evaluation scenarios are considered. In the UID-only scenario, genuine trials correspond to enrolled users and impostor trials correspond to non-enrolled users, regardless of the spoken content. In the UID + password scenario, genuine trials correspond to enrolled users pronouncing the correct passphrase, while impostor trials include either non-enrolled users or enrolled users uttering an incorrect passphrase. Participants were encouraged to explore deep learning, traditional machine learning, and multimodal fusion approaches in order to improve authentication performance, generalization, and efficiency. Teams of two to five members were allowed, and each participant could belong to at most one team.

Table 9 Challenge results based on Equal Error Rate (EER) in two scenarios: using (i) UID only and (ii) UID + password

Team	EER (UID + password)	EER (UID only)
Team A	24.84%	22.51%
Team B	**3.18%**	**3.18%**

6.2 Challenge best teams

Two teams, referred to here as Team A and Team B, achieved the most competitive results and adopted notably different technical approaches.

[3] Challenge: Dual Factor Lip-Based Biometric User Authentication at the 23rd International Conference on Image Analysis and Processing (ICIAP 2025), 15–19 September 2025, Sapienza University of Rome, Italy, https://www.dmi.unict.it/spata/BiovidChallenge/

[4] BioVid Challenge, ICIAP 2025.

Team A: Gated Audio–Visual Fusion for Open-Set Authentication Team A developed a multimodal biometric authentication framework specifically tailored to the *BioVid* Challenge. The system processes short smartphone-recorded videos and leverages the synchrony between lip motion and spoken passphrases to verify users in an open-set scenario, where it must both authenticate enrolled users and correctly reject impostors or incorrect phrases.

Figure 6 provides an overview of the proposed dual-stream architecture. The visual stream encodes spatio-temporal lip movements, while the audio stream produces speaker-specific embeddings. These two modalities are merged through a Gated Multimodal Unit (GMU), which performs adaptive audio–visual fusion by learning dimension-wise gates that determine the relative contribution of each modality. The resulting fused embedding is then optimized using a hybrid objective that combines triplet loss with binary cross-entropy, enabling both discriminative embedding learning and reliable genuine/impostor classification. During inference, verification is carried out via cosine similarity against stored user templates, followed by threshold-based decision-making to support open-set rejection.

Focusing on the visual-data branch (top in Fig. 6), each video is uniformly sampled into 30 RGB frames and resized to 96×96 pixels. A 3D ResNet-18 [36] backbone extracts local spatiotemporal patterns associated with mouth movements. To capture longer-range temporal dependencies, the encoded features are passed through a Bidirectional GRU (Bi-GRU), which aggregates information in both forward and backward directions, resulting in a 256-dimensional visual embedding $\mathrm{v} \in \mathbb{R}^{256}$, where v denotes a compact representation of the lip-motion dynamics, with each dimension encoding discriminative spatiotemporal cues that are informative for user authentication in the open-set verification scenario.

Similarly, in the audio branch (bottom in Fig. 6) the audio component is extracted from the input video, resampled to 16 kHz, and normalized. A fine-tuned ECAPA-TDNN

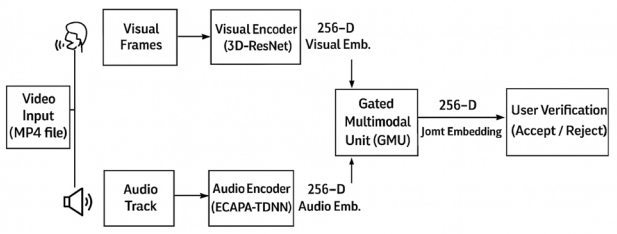


Fig. 6 Team A multimodal authentication architecture combining spatiotemporal lip dynamics (3D ResNet + BiGRU) with speaker embeddings (ECAPA-TDNN). The Gated Multimodal Unit (GMU) performs adaptive audio–visual fusion, producing a robust representation for open-set biometric verification

model [37] is used to generate a 192-dimensional speaker embedding $\mathrm{a} \in \mathbb{R}^{192}$, where the vector a captures speaker-specific acoustic characteristics such as timbre, prosodic rhythm, and vocal tract resonances. Before multimodal fusion, the 192-dimensional audio embedding needs to be projected into the same 256-dimensional space as the visual embedding. To this aim, a simple linear projection is applied, obtainig a new 256-dimensional speaker embedding $\tilde{\mathbf{a}} \in \mathbb{R}^{256}$, where the vector $\tilde{\mathbf{a}}$ still preserves the linear correlations insisting on the speaker-specific acoustic characteristics.

The two modalities are then fused together through an element-wise gating and fusion procedure performed by the Gated Multimodal Unit (GMU) [38]. In particular, the GMU learns a data-dependent gating vector to control the contribution of the visual and audio embeddings defined as

$$\mathrm{z} = \sigma(\mathrm{W}_v \mathrm{v} + \mathrm{W}_{\tilde{a}} \tilde{\mathrm{a}} + \mathrm{b})$$

where $\sigma(\cdot)$ denotes the sigmoid activation function, while W_v, $\mathrm{W}_{\tilde{a}}$ and b are the weights and the bias terms learnt by the GMU module. Each element of z lies in the range [0, 1] and determines, on a per-dimension basis, whether the fused representation should rely more heavily on the visual or the audio modality. The final multimodal embedding is obtained through an adaptive interpolation

$$\mathrm{f} = \mathrm{z} \odot \tanh(\mathrm{v}) + (1 - \mathrm{z}) \odot \tanh(\tilde{\mathrm{a}}), \qquad \mathrm{f} \in \mathbb{R}^{256}$$

where $\odot$ indicates element-wise multiplication. This formulation ensures that the model selectively emphasizes the modality that provides the most discriminative information for each input sample. GMU component is the only one needing to be trained, as the 3D ResNet-18 and the fine-tuned ECAPA-TDNN are used as feature extraction modules. The GMU training is thus performed using a hybrid loss function that combines metric learning with binary classification. The former is achieved by a triplet loss function that encourages the fused embedding to be discriminative by enforcing intra-class compactness and inter-class separation:

$$\mathcal{L}_{\text{triplet}} = \max\{0,\, d(f_a, f_p) - d(f_a, f_n) + m\},$$

where f_a, f_p, and f_n correspond to the anchor, positive, and negative embeddings, respectively, while $d(\cdot)$ denotes the distance metric and m is a margin ensuring that the negative example is sufficiently farther from the anchor than the positive one. To stabilize the genuine–impostor discrimination, the triplet formulation is complemented with a binary cross-entropy term:

$$\mathcal{L}_{CE} = -[\, y \log(\hat{y}) + (1-y) \log(1-\hat{y}) \,],$$

where $y \in \{0, 1\}$ is the ground-truth label and $\hat{y}$ is the predicted probability of a genuine attempt. The final training loss is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{triplet}} + \lambda \mathcal{L}_{CE}, \qquad \lambda = 1,$$

allowing the model to jointly optimize for robust embedding learning and reliable verification performance.

At inference time, verification relies on cosine similarity between the fused embedding f and each enrolled user template t_u:

$$S(u) = \frac{\mathrm{f} \cdot \mathrm{t}_u}{\|\mathrm{f}\| \, \|\mathrm{t}_u\|}.$$

The predicted identity is:

$$\hat{u} = \arg\max_u S(u),$$

and an attempt is accepted only if the maximum similarity score exceeds a predefined threshold:

$$\max_u S(u) \geq \tau, \qquad \tau = 0.60.$$

This thresholding mechanism enables effective open-set rejection of unseen users and incorrect passphrases.

Across the three validation folds, the proposed system achieves an average accuracy of 71.36%, with an EER of 28.61%. The error profile remains balanced, with APCER and BPCER values of 28.68% and 28.55%, respectively, indicating that the model does not exhibit modality or class bias in rejecting impostors versus genuine attempts. On the hidden challenge test set, the model attains an accuracy of 71.00%, demonstrating that the GMU-based audio–visual fusion and hybrid metric-learning objective generalize effectively to unseen users and passphrase conditions. The results in Table 10 further confirm the robustness of the proposed multimodal authentication framework in real-world open-set biometric verification scenarios.

Team B: LOKI-Authentication The team proposed *LOKI-Authentication* (Localized lip patch and Oral Key Integration), a modular multimodal pipeline designed to overcome the inherent weaknesses of unimodal biometric systems. The architecture (Fig. 7) comprises three key modules: lip-based visual analysis combined with an audio-based speaker recognition, followed by passphrase verification for final authentication.

Table 10 Performance of Team A across the three validation folds and the hidden test set

Dataset split	Accuracy (%)	EER (%)	APCER (%)	BPCER (%)
Fold 0	72.48	27.53	27.47	27.58
Fold 1	68.46	31.57	31.39	31.75
Fold 2	73.15	26.74	27.17	26.31
Average (3-fold)	71.36	28.61	28.68	28.55

For the visual stream, the lip region of interest (ROI) is localized through the *Segment Anything Model* (SAM) [39]. A validity check ensures segmentation quality; if it fails, the frame is automatically re-processed using an adaptive zoom strategy. The segmented lip patch is then processed by an *InceptionResNetV1* network [7], pre-trained on VGGFace2 [40], to obtain frame-level embeddings. These are aggregated via median pooling and compared to the enrolled templates using cosine similarity, yielding a visual similarity score $S_V(u)$ for each enrolled identity u.

For the audio stream, to improve generalizability, the system extracts speaker embeddings using two complementary toolkits: *SpeechBrain* [41, 42] and *WeSpeaker* [43, 44]. SpeechBrain provides robust spectrogram-based representations, while WeSpeaker offers lightweight yet discriminative embeddings optimized for speaker verification. The cosine similarities between the extracted and enrolled embeddings produce two audio scores, later combined via a weighted average to obtain the final audio score $S_A(u)$.

The multimodal fusion then integrates both modalities at the score level through a weighted combination:

$$S_F(u) = \omega_A S_A(u) + \omega_V S_V(u),$$

where ω_A, ω_V are tuned on a validation set. Specifically, these have been set based on the EER assessed on the single modalities, according to the following formula:

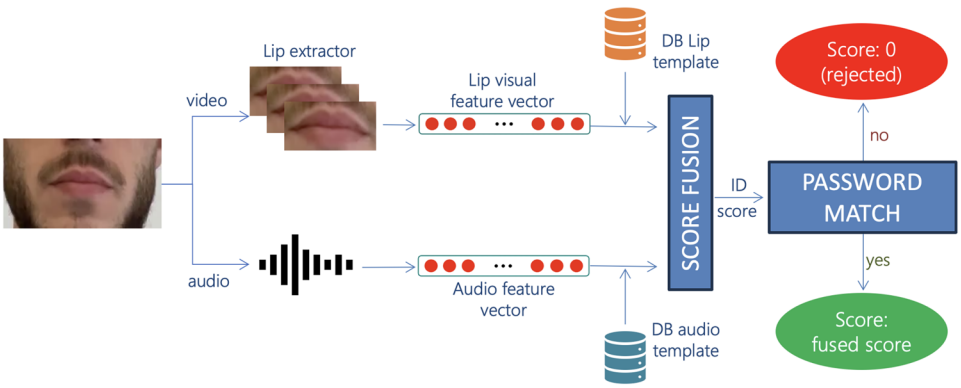
$$\omega_i = \frac{\frac{1}{EER_i}}{\sum_j \frac{1}{EER_j}}$$

The identity with the highest fused score is selected as the candidate user:

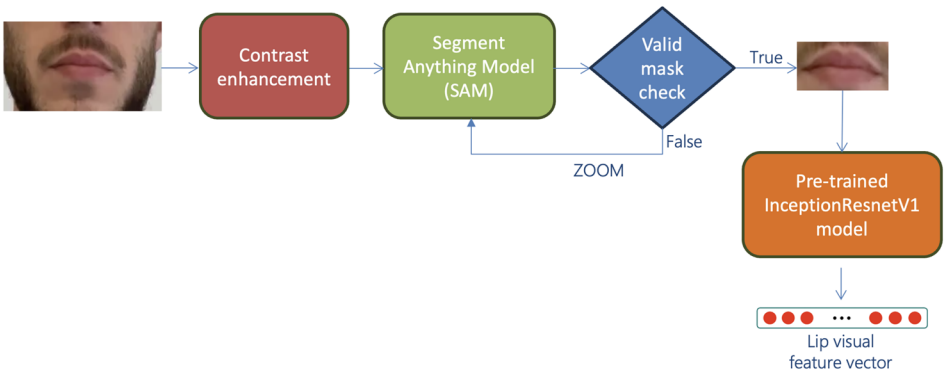
$$\hat{u} = \arg\max_u S_F(u).$$

Once the most likely identity $\hat{u}$ has been determined, the system verifies whether the spoken content matches the passphrase associated with that user.

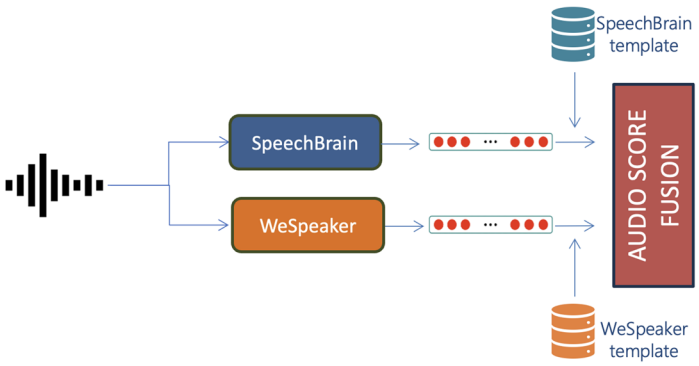
The recorded audio is transcribed using *OpenAI Whisper* [45], and the resulting text is compared with the enrolled passphrase via the normalized *Levenshtein distance*, which measures the minimal number of character-level edits (insertions, deletions, substitutions) required for alignment. If the distance is below a predefined threshold, the passphrase verification is considered successful and the fused score $S_F(\hat{u})$



(a) Overview of the LOKI-Authentication architecture.



(b) Lip visual feature extraction pipeline.



(c) Audio feature extraction pipeline.

is retained. Otherwise, the authentication attempt is rejected by setting the final score to zero. This two-stage scheme, first biometric identification and then semantic validation, reduces mismatches in speech content while maintaining tolerance for minor transcription or pronunciation errors.

The system exhibits highly consistent performance, achieving a ROC AUC of 0.97 and showing well-separated score distributions between genuine and impostor trials (Fig. 8-9). A modality-wise analysis, reported in Table 11, further highlights the relative contribution of each component. In terms of EER, the audio-based systems consistently outperform the video-only configuration (39.86%), with WeSpeaker achieving the best unimodal performance (22.97%), followed by SpeechBrain (27.03%). The fusion of the two audio systems results in an EER of 25.00%, indicating limited complementarity between the audio backbones. The multimodal fusion (Audio+Video) achieves an EER of 24.32%, substantially improving over the video-only setup and slightly outperforming the audio-only fusion, although it does not surpass the strongest unimodal audio model.

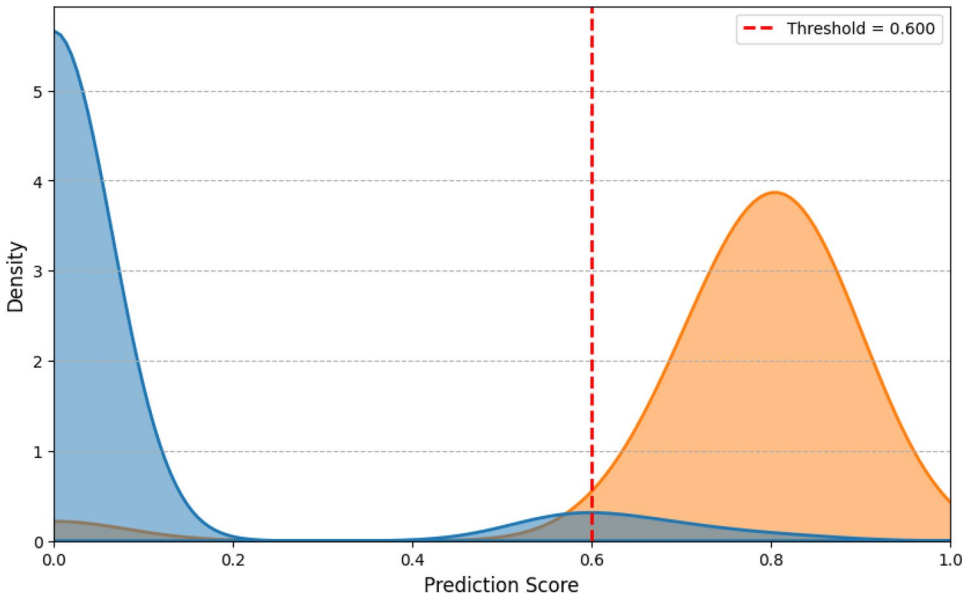


Fig. 8 Distribution of fused prediction scores for genuine and impostor trials for the LOKI-Authentication system (Team B)

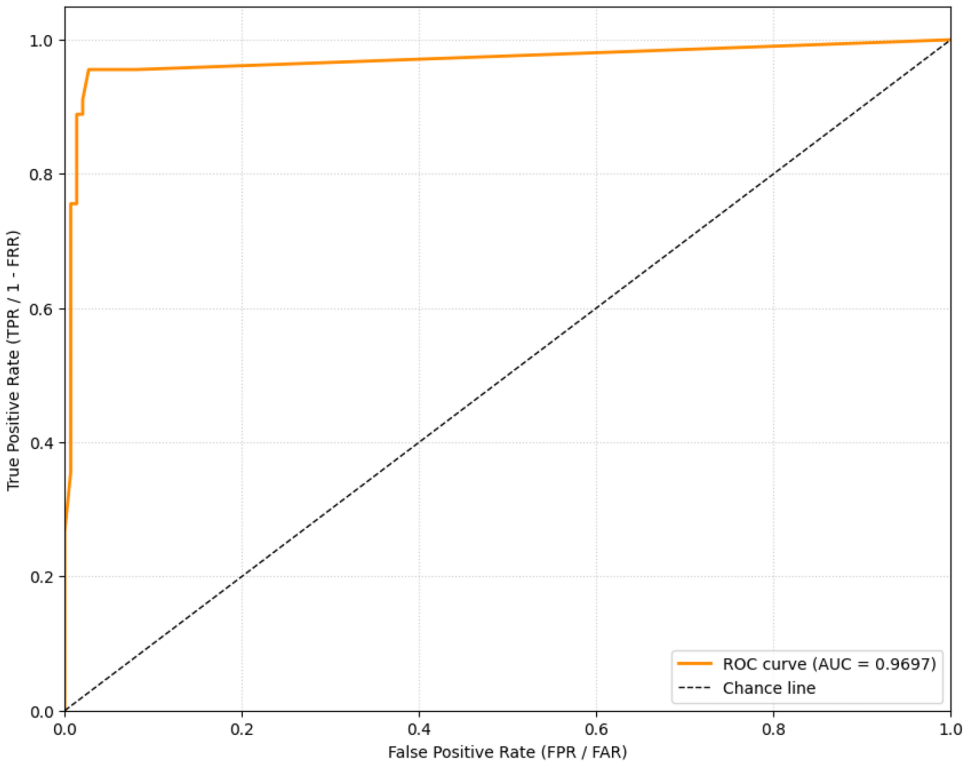


Fig. 9 ROC curve of the Team B LOKI-Authentication system (AUC = 0.97)

Table 11 Performance of Team B obtained on validation data across unimodal and multimodal settings before the passphrase-verification stage

	Modality				
	Video	Audio (SpeechBrain)	Audio (WeSpeaker)	Audio (fused)	Multi-modal
EER	39.86%	27.03%	22.97%	25.00%	24.32%

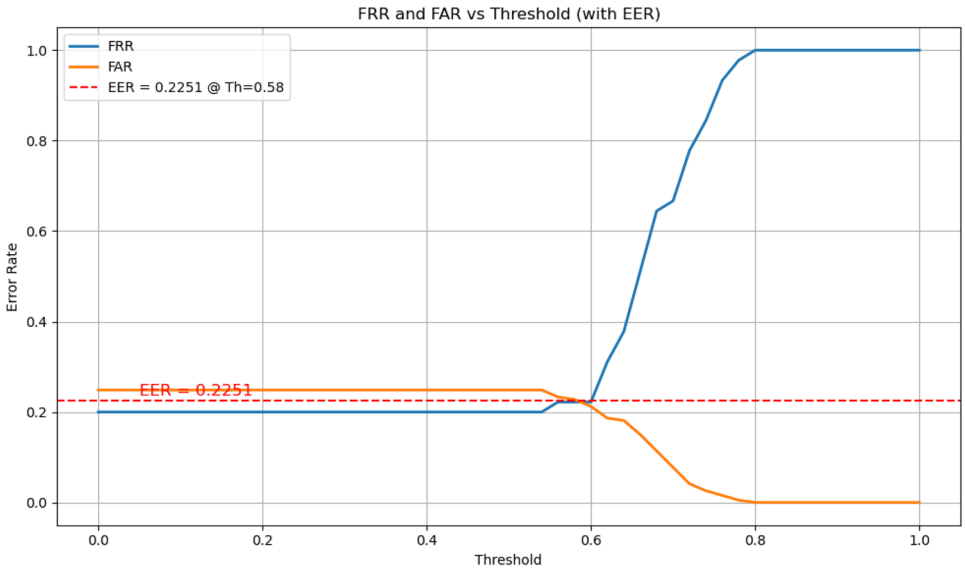
Overall, these results suggest that the audio modality primarily drives performance, while the visual stream provides complementary information within the current fusion framework.

6.3 Challenge best results

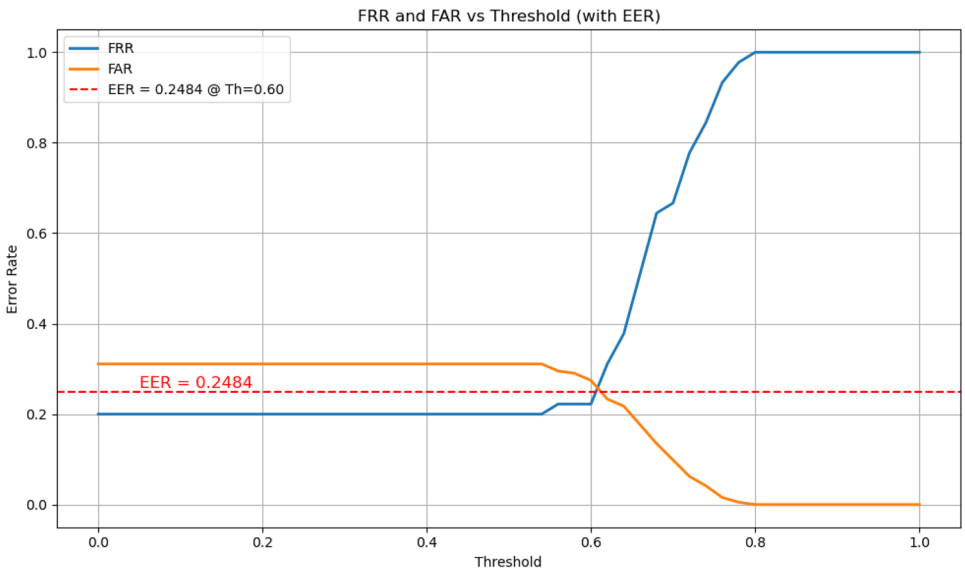
To quantitatively compare the effectiveness of the submitted systems, we adopted the EER under two authentication scenarios: *UID only*, where only the speaker identity

is verified, and *UID + password*, where both identity and spoken passphrase must be correct. The similarity scores provided by each participant were evaluated over a set of thresholds $\tau \in [0, 1]$ with step 0.02. For each threshold, we computed the corresponding False Acceptance Rate (FAR) and False Rejection Rate (FRR), then derived the EER from the operating point where FAR and FRR curves intersect, in line with standard biometric practice. Table 9 and Figs. 10, 11 summarize the final results.

Under the adopted protocol, performance falls in two clearly separated regimes: Team B (LOKI-Authentication) attains an EER of 3.18% in both the *UID + password* and *UID only* scenarios, whereas Team A yields EER values of 24.84% and 22.51%, respectively. The observed difference points to a relevant influence of fusion strategy and passphrase handling, although the available evidence is restricted to two submitted systems. LOKI-Authentication keeps audio and visual subsystems independent, fuses their outputs at the score level, and constrains the final decision through an explicit passphrase check. The GMU-based model instead integrates the two modalities at the feature level into a single joint representation optimized with a hybrid triplet plus binary cross-entropy objective. In the *BioVid* setting, characterised by limited training data and heterogeneous smartphone devices, the results suggest that score-level fusion combined with explicit content verification exploits

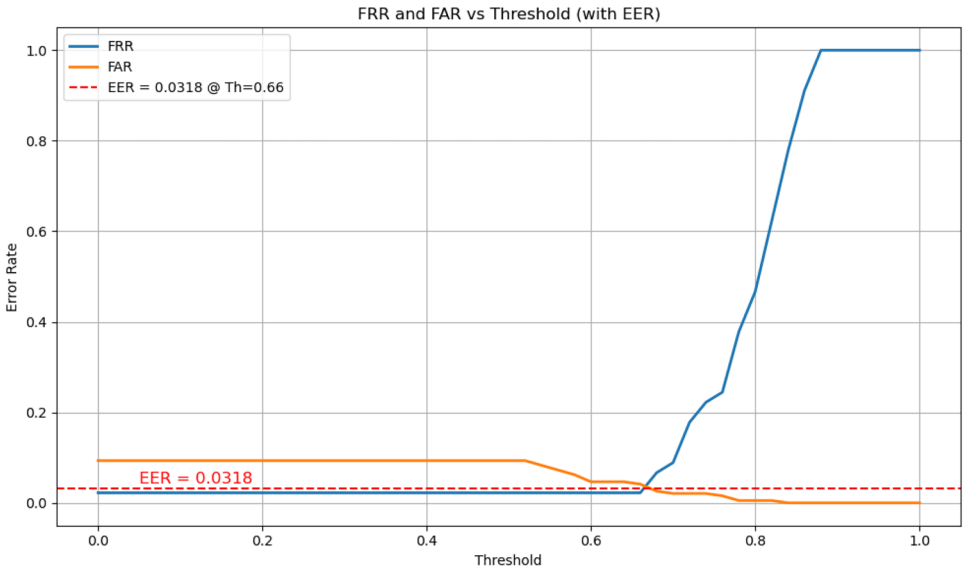


(a) Team A: UID only

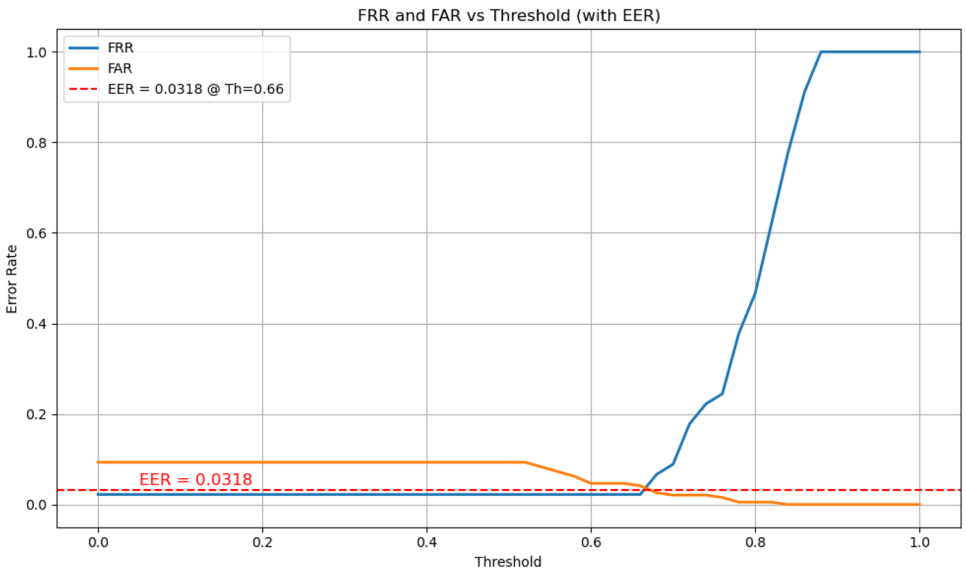


(b) Team A: UID + password

Fig. 10 False Acceptance Rate (FAR) and False Rejection Rate (FRR) curves for Team A under the UID-only and UID+password scenarios. For each configuration, the Equal Error Rate (EER) corresponds to the intersection point between the FAR and FRR curves



(a) Team B: UID only



(b) Team B: UID + password

Fig. 11 False Acceptance Rate (FAR) and False Rejection Rate (FRR) curves for Team B under the UID-only and UID+password scenarios. For each configuration, the Equal Error Rate (EER) corresponds to the intersection point between the FAR and FRR curves

the complementary information in voice and lip motion more effectively than monolithic feature-level fusion. Such indications should be regarded as preliminary, since the two submissions differ along multiple architectural dimensions and additional systems would be required to disentangle the contribution of each design choice.

7 Conclusions and future works

This paper introduced BioVid, a privacy-aware audio-visual biometric dataset designed for user recognition and passphrase verification from videos acquired with consumer-grade smartphones. The dataset comprises 650 recordings from 43 participants and covers 90 Italian words, including both correct and incorrect passphrase utterances. Its acquisition protocol incorporates variability in recording device, distance, location, background, illumination, and acoustic conditions, providing a challenging resource for the evaluation of biometric systems beyond tightly controlled laboratory settings.

BioVid supports two complementary experimental settings. The closed-set experiments provide an initial characterization of the biometric information contained in the dataset, showing that lip-shape descriptors can be highly discriminative for enrolled-user identification. In particular,

the Extra Trees Classifier achieved a Top-1 accuracy of 91.98% using Shape Context features, while the evaluated deep audio-visual architectures reached accuracies of up to 88.00%. These results should not be directly compared with the open-set challenge results, as the two settings address different tasks and adopt different data-partitioning and evaluation protocols.

The BioVid Challenge extended the evaluation to a more demanding open-set authentication scenario, requiring systems to reject both unseen identities and enrolled users pronouncing an incorrect passphrase. The two submitted solutions produced markedly different results. Team B achieved an EER of 3.18% in both the UID-only and UID-plus-passphrase scenarios, whereas Team A obtained EER values of 22.51 and 24.84%, respectively. Within the adopted protocol, these results suggest that maintaining separate biometric subsystems, combining their outputs at the score level, and performing an explicit passphrase-consistency check can be effective when training data are limited and acquisition conditions are heterogeneous. Nevertheless, this observation should be considered preliminary as the two systems differ in several architectural and training choices, and a larger number of controlled submissions would be required to isolate the contribution of the fusion strategy itself.

The modality-level analysis also indicates that the audio stream currently provides most of the discriminative power, while the visual stream supplies complementary information. This finding highlights that an important research direction for BioVid is in making future models exploit not only static lip appearance but also fine-grained temporal articulation and audio-visual synchronization. More effective lip-motion representations, including landmark trajectories, keypoint-based descriptors, spatiotemporal encoders, and cross-modal attention mechanisms, may increase the contribution of the visual modality, particularly under noisy acoustic conditions.

The present study also has several limitations. BioVid remains relatively small in terms of participants, repetitions, and unseen identities, and its vocabulary is restricted to short Italian utterances. The number of recordings is not uniform across participants, and device identifiers are not included in the public release, preventing a direct evaluation of device-specific effects. Furthermore, the challenge EER was estimated through a post-hoc threshold sweep on the test scores rather than by fixing the operating threshold on an independent development set. Consequently, the reported results establish an initial benchmark but should not yet be interpreted as evidence of deployment-level performance.

Future releases will therefore focus on increasing the number and diversity of participants, devices, sessions, languages, and passphrases, while including longitudinal acquisitions to evaluate cross-session biometric stability. A revised benchmark should provide separate training, development, and hidden test partitions, together with fixed threshold-selection procedures, confidence intervals, and cross-device and cross-condition evaluation protocols. Further work should also investigate robustness to missing or degraded modalities, calibration of authentication scores, computational efficiency for on-device inference, and resistance to replay, presentation, voice-conversion, and audio-visual deepfake attacks. Finally, extending the challenge to a larger set of independently developed systems would enable more reliable comparisons among feature-level, score-level, and decision-level fusion strategies.

Overall, BioVid provides a reproducible starting point for studying dual-factor audio-visual authentication on consumer smartphones. By releasing both the dataset and an open-set evaluation protocol, this work aims to support systematic research on biometric systems that are not only accurate, but also robust, secure, and suitable for realistic mobile acquisition conditions.

Acknowledgements G.F. is supported by PNRR MUR project PE0000013-FAIR. M.O.S. is supported by HICONNECTS, Grant 101097296. S.B. is supported by Spoke 1 Future HPC & Big Data of the Italian Research Center on High-Performance Computing.

Author contributions Georgia Fargetta: Writing, Original draft preparation, Editing, Data curation, Visualization, Investigation. Massimo Orazio Spata: Conceptualization, Methodology, Software, Writing, Editing, Original draft preparation. Alessandro Ortis: Writing, Original draft preparation, Editing, Data curation, Visualization, Investigation. Sebastiano Battiato: Supervision, Reviewing. Riccardo Puddu: Writing, Original draft preparation, Editing, Data curation. Doris Dearca: Writing, Original draft preparation, Editing, Data curation. Alessandro Saba: Writing, Original draft preparation, Editing, Data curation. Gianpaolo Perelli: Writing, Original draft preparation, Editing, Data curation. Simone Maurizio La Cava: Writing, Original draft preparation, Editing, Data curation. Marco Micheletto: Writing, Original draft preparation, Editing, Data curation. Giulia Orrù: Writing, Original draft preparation, Editing, Data curation. Gian Luca Marcialis: Writing, Original draft preparation, Editing, Data curation. Venu Siddapura Govindaraju: Writing, Original draft preparation, Editing, Data curation. Stefano Marrone: Writing, Original draft preparation, Editing, Data curation. Carlo Sansone: Writing, Original draft preparation, Editing, Data curation.

Funding Open access funding provided by Università degli Studi di Catania within the CRUI-CARE Agreement. M.O.S. is supported by HICONNECTS, Grant 101097296; G.F. and V.S.G. have been supported by PNRR MUR under Project PE0000013-FAIR.

Data availability BioVid dataset: the URL for downloading the dataset is available at: https://www.kaggle.com/datasets/maxspata/biovid/data

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate

if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Srihari, M., Bhaskaran, V.: A multimodal fusion approach: emotion identification from audio and video using convolutional neural network. In: 2024 IEEE International Conference on Computer Vision and Machine Intelligence (CVMI), pp. 1–5. (2024)
2. Appasamy, K., Shanmugasundaram, R.S.: Deepfusenet: a novel multimodal biometric authentication model with deep neural networks. In: 2025 International Conference on Frontier Technologies and Solutions (ICFTS), pp. 1–8. (2025)
3. Spata, M.O., Ortis, A., Fargetta, G., Battiato, S.: Cnnmc: a convolutional neural network with monte carlo dropout for speaker recognition. EURASIP J. Inf. Secur. **2025**(1), 34 (2025)
4. Garofolo, J.S., Lamel, L.F., Fisher, W.M., Pallett, D.S., Dahlgren, N.L., Zue, V., Fiscus, J.G.: TIMIT Acoustic-Phonetic Continuous Speech Corpus. Linguistic Data Consortium (1993). https://doi.org/10.35111/17gk-bn40. https://cir.nii.ac.jp/crid/1881146593179904768
5. Nagrani, A., Chung, J.S., Zisserman, A.: Voxceleb: a large-scale speaker identification dataset. In: Interspeech (2017)
6. Goel, G., Sridevi, K., Maitra, A., Piplani, P., SS, P., etal.: Security identity: an audio-video biometric approach. In: 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), pp. 1–5 (2023). IEEE
7. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, 31, (2017)
8. Chen, H., Du, J., Dai, Y., Lee, C.-H., Siniscalchi, S.M., Watanabe, S., Scharenborg, O., Chen, J., Yin, B., Pan, J.: Audio-visual speech recognition in misp2021 challenge: dataset release and deep analysis. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, vol. 2022, pp. 1766–1770 (2022)
9. Thapa, S., Rauniyar, K., Jafri, F., Veeramani, H., Jain, R., Jain, S., Vargas, F., Hürriyetoğlu, A., Naseem, U.: Extended multimodal hate speech event detection during russia-ukraine crisis-shared task at case 2024. In: Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024), pp. 221–228 (2024)
10. Blanco, A.L.A., Valentini-Botinhao, C., Klejch, O., Gogate, M., Dashtipour, K., Hussain, A., Bell, P.: Avse challenge: audio-visual speech enhancement challenge. In: 2022 IEEE Spoken Language Technology Workshop (SLT), pp. 465–471. IEEE (2023)
11. Chen, S., Li, L., Ren, S., Gao, R., Liu, Y., Bi, X., Sun, X., Hou, L.: Towards multimodal video paragraph captioning models robust to missing modality. arXiv preprint arXiv:2403.19221 (2024)
12. Ragusa, F., Tomaselli, V., Furnari, A., Battiato, S., Farinella, G.M.: Food vs non-food classification. In: Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management, pp. 77–81. (2016)
13. Chaptoukaev, H., Strizhkova, V., Panariello, M., Dalpaos, B., Reka, A., Manera, V., Thümmler, S., Ismailova, E., Todisco, M., Zuluaga, M.A., et al.: Stressid: a multimodal dataset for stress identification. Adv. Neural. Inf. Process. Syst. **36**, 29798–29811 (2023)
14. Yang, Q., Peng, G., Nguyen, D.T., Qi, X., Zhou, G., Sitová, Z., Gasti, P., Balagani, K.S.: A multimodal data set for evaluating continuous authentication performance in smartphones. In: Proceedings of the 12th ACM Conference on Embedded Network Sensor Systems, pp. 358–359. (2014)
15. Choi, H., Park, H.: A multimodal user authentication system using faces and gestures. Biomed. Res. Int. **2015**(1), 343475 (2015)
16. Spata, M.O., Russo, V.M., Ortis, A., Battiato, S.: A new pipeline for snooping keystroke based on deep learning algorithm. IEEE Access **13**, 24498–24514 (2025)
17. Matthews, I., Cootes, T.F., Bangham, J.A., Cox, S., Harvey, R.: Extraction of visual features for lipreading. IEEE Trans. Pattern Anal. Mach. Intell. **24**(2), 198–213 (2002)
18. Cox, S.J., Harvey, R.W., Lan, Y., Newman, J.L., Theobald, B.J.: The challenge of multispeaker lip-reading, pp. 179–184. AVSP (2008)
19. Patterson, E.K., Gurbuz, S., Tufekci, Z., Gowdy, J.N.: Cuave: a new audio-visual database for multimodal human-computer interface research. In: 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 2, p. 2017 (2002). IEEE
20. Lee, B., Hasegawa-Johnson, M., Goudeseune, C., Kamdar, S., Borys, S., Liu, M., Huang, T.S.: Avicar: audio-visual speech corpus in a car environment, pp. 2489–2492. Interspeech (2004)
21. Cooke, M., Barker, J., Cunningham, S., Shao, X.: An audio-visual corpus for speech perception and automatic speech recognition. J. Acoust. Soc. Am. **120**(5), 2421–2424 (2006)
22. Zhao, G., Barnard, M., Pietikainen, M.: Lipreading with local spatiotemporal descriptors. IEEE Trans. Multimed. **11**(7), 1254–1265 (2009)
23. Anina, I., Zhou, Z., Zhao, G., Pietikäinen, M.: Ouluvs2: a multi-view audiovisual database for non-rigid mouth motion analysis. In: 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 1, pp. 1–5 (2015). IEEE
24. Chung, J.S., Zisserman, A.: Lip reading in the wild. In: Asian Conference on Computer Vision, pp. 87–103. Springer (2016)
25. Chung, S., Senior, J., Vinyals, A., Zisserman, O., A.: Lip reading sentences in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6447–6456. (2017)
26. Afouras, T., Chung, J.S., Zisserman, A.: Lrs3-ted: a large-scale dataset for visual speech recognition, (2018). arXiv preprint arXiv:1809.00496
27. Rekik, A., Ben-Hamadou, A., Mahdi, W.: An adaptive approach for lip-reading using image and depth data. Multimed. Tools Appl. **75**(14), 8609–8636 (2016)
28. Messer, K., Matas, J., Kittler, J., Luettin, J., Maitre, G.: Xm2vtsdb: the extended m2vts database. In: Second International Conference on Audio and Video-based Biometric Person Authentication, 964, pp. 965–966, Washington, DC (1999)
29. Larcher, A., Lee, K.A., Ma, B., Li, H.: The rsr2015: database for text-dependent speaker verification using multiple pass-phrases. In: Annual Conference of the International Speech Communication Association (Interspeech), (2012)
30. Wright, C., Stewart, D.W.: Understanding visual lip-based biometric authentication for mobile devices. EURASIP J. Inf. Secur. **2020**(1), 3 (2020)
31. Morrison-Smith, S., Aloba, A., Lu, H., Benda, B., Esmaeili, S., Flores, G., Smith, J., Soni, N., Wang, I., Joy, R.: Mmgatorauth: a novel multimodal dataset for authentication interactions in gesture and voice. In: Proceedings of the 2020 International Conference on Multimodal Interaction, pp. 370–377. (2020)

32. Belongie, S., Malik, J., Puzicha, J.: Shape context: a new descriptor for shape matching and object recognition. Adv. Neural. Inf. Process. Syst. 13, (2000)
33. Narvekar, N.D., Karam, L.J.: A no-reference image blur metric based on the cumulative probability of blur detection (cpbd). IEEE Trans. Image Process. **20**(9), 2678–2683 (2011)
34. Vu, C.T., Phan, T.D., Chandler, D.M.: S_3: a spectral and spatial measure of local perceived sharpness in natural images. IEEE Trans. Image Process. **21**(3), 934–945 (2011)
35. Nagrani, A., Chung, J.S., Xie, W., Zisserman, A.: Voxceleb: large-scale speaker verification in the wild. Comput. Speech Lang. **60**, 101027 (2020)
36. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778. (2016)
37. Dawalatabad, N., Ravanelli, M., Grondin, F., Thienpondt, J., Desplanques, B., Na, H.: Ecapa-tdnn embeddings for speaker diarization, (2021). arXiv:2104.01466 arXiv preprint
38. Arevalo, J., Solorio, T., Montes-y-Gómez, M., González, F.A.: Gated multimodal units for information fusion. arXiv preprint arXiv:1702.01992 (2017)
39. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y.: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4015–4026. (2023). (**Segment anything**)
40. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: a dataset for recognising faces across pose and age. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pp. 67–74 (2018). IEEE
41. Ravanelli, M., Parcollet, T., Moumen, A., Langen, S., Subakan, C., Plantinga, P., Wang, Y., Mousavi, P., Libera, L.D., Ploujnikov, A., Paissan, F., Borra, D., Zaiem, S., Zhao, Z., Zhang, S., Karakasidis, G., Yeh, S.-L., Champion, P., Rouhe, A., Braun, R., Mai, F., Zuluaga-Gomez, J., Mousavi, S.M., Nautsch, A., Nguyen, H., Liu, X., Sagar, S., Duret, J., Mdhaffar, S., Laperrière, G., Rouvier, M., Mori, R.D., Estève, Y.: Open-source conversational ai with speechbrain 1.0. J. Mach. Learn. Res. **25**(333) (2024)
42. Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., Subakan, C., Dawalatabad, N., Heba, A., Zhong, J., Chou, J.-C., Yeh, S.L., Fu, S.W., Liao, C.F., Rastorgueva, E., Grondin, F., Aris, W., Na, H., Gao, Y., Mori, R.D., Bengio, Y.: SpeechBrain: a general-purpose speech toolkit, (2021). arXiv:2106.04624
43. Wang, S., Chen, Z., Han, B., Wang, H., Liang, C., Zhang, B., Xiang, X., Ding, W., Rohdin, J., Silnova, A., et al.: Advancing speaker embedding learning: wespeaker toolkit for research and production. Speech Commun. **162**, 103104 (2024)
44. Wang, H., Liang, C., Wang, S., Chen, Z., Zhang, B., Xiang, X., Deng, Y., Qian, Y.: Wespeaker: a research and production oriented speaker embedding learning toolkit. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5 (2023). IEEE
45. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: International Conference on Machine Learning, pp. 28492–28518 (2023). PMLR

Publisher"s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Georgia Fargetta[1] · Massimo Orazio Spata[1] · Alessandro Ortis[1] · Sebastiano Battiato[1] · Riccardo Puddu[2] · Doris Dearca[2] · Alessandro Saba[2] · Gianpaolo Perelli[2] · Simone Maurizio La Cava[2] · Marco Micheletto[2] · Giulia Orrù[2] · Gian Luca Marcialis[2] · Venu Siddapura Govindaraju[3] · Stefano Marrone[3] · Carlo Sansone[3]

✉ Massimo Orazio Spata
massimo.spata@unict.it

Georgia Fargetta
georgia.fargetta@unict.it

Alessandro Ortis
alessandro.ortis@unict.it

Sebastiano Battiato
sebastiano.battiato@unict.it

Riccardo Puddu
r.puddu24@studenti.unica.it

Doris Dearca
d.dearca@studenti.unica.it

Alessandro Saba
a.saba30@studenti.unica.it

Gianpaolo Perelli
gianpaolo.perelli@unica.it

Simone Maurizio La Cava
simonem.lac@unica.it

Marco Micheletto
marco.micheletto@unica.it

Giulia Orrù
giulia.orru@unica.it

Gian Luca Marcialis
marcialis@unica.it

Venu Siddapura Govindaraju
v.siddapuragovindaraju@studenti.unina.it

Stefano Marrone
stefano.marrone@unina.it

Carlo Sansone
carlo.sansone@unina.it

[1] Department of Mathematics and Computer Science, University of Catania, Catania, Italy

[2] Department of Electrical and Electronic Engineering, University of Cagliari, Cagliari, Italy

[3] Department of Electrical Engineering and of Information Technologies, University of Naples Federico II, Naples, Italy