Supplementary Material for "Enhancing the Security of Visual Speaker Authentication Based on Dynamic Lip-Print Analysis"

Supplementary Material

Viseme	Occur. (%)	Phonemes
/A	3.15	/f/, /v/
/B	15.49	/er/, /ow/, /r/, /q/, /w/, /uh/, /uw/, /axr/, /ux/
/C	5.88	/b/, /p/, /m/, /em/
/D	0.70	/aw/
/E	2.90	/dh/, /th/
/F	1.20	/ch/, /jh/, /sh/, /zh/
/G	1.81	/oy/, /ao/
/H	4.36	/s/, /z/
/I	31.46	/aa/, /ah/, /ay/, /eh/, /ey/, /ih/, /iy/, /y/, /ae/, /ax-h/, /ax/, /ix/
/J	21.10	/d/, /l/, /n/, /t/, /el/, /nx/, /en/, /dx/
/K	4.84	/g/, /k/, /ng/, /eng/
/S	-	/sil/, /pcl/, /tcl/, /kcl/, /bcl/, /dcl/, /gcl/, /h#/, /#h/, /pau/, /epi/

Table 1. The correspondence map from viseme to TIMIT phonemes proposed by Jeffers. The last viseme, /S is used for *silence*. The table shows the occurrence rate in spoken English.

α	AUC_{hm}	AUC_{fs}	AUC_{dfl}	AUC_{ss}	AUC_{ls}	AUC_{avg}
0.0	0.9889	0.9874	0.9809	0.9921	0.9877	0.9874
0.25	0.9963	0.9888	0.9922	0.9953	0.9925	0.9930
0.5	0.9992	0.9951	0.9983	0.9971	0.9954	**0.9970**
0.75	0.9996	0.9953	0.9954	0.9977	0.9913	0.9959
1.0	0.9978	0.9797	0.9758	0.9924	0.9672	0.9826

Table 2. Parameter selection experiment with different α values

1. Viseme Definition

Jeffers's viseme map [3] is selected as our viseme map, which had been proven to be the most effective viseme definition in speaker identification [1]. As shown in Table 1, it has 12 categories including "/A" to "/K" and *silence* viseme ("/S") . There is a one-to-many mapping between visemes and phonemes defined in TIMIT [2].

2. Parameter Selection

In this section, we explored the impact of the parameter α, which controls the trade-off between the global and local score components in our visual speaker authentication system. The experiments were conducted under the *unified_prompts* setting using VSA dataset.

Specifically, setting $\alpha = 1$ relies solely on the global score, while $\alpha = 0$ uses only the local score. As shown in Table 2, the global branch demonstrates stronger resistance against human imposters, whereas the local branch is more effective at detecting DeepFake attacks. This is because our multi-level dynamic-enhanced encoder captures fine-grained, speaker-specific talking styles that are difficult for DeepFake attackers to replicate.

Global features, extracted from all frames, include non-speech information that is filtered out during the local branch's viseme segmentation, making them more sensitive to static cues that can help distinguish genuine human impersonators. When both scores are combined, the overall AUC improves, highlighting the complementary nature of the two feature types. Ultimately, we selected $\alpha = 0.5$ as the optimal value, yielding the highest average AUC across all scenarios.

Viseme Segmentation	AUC_{hm}	AUC_{fs}	AUC_{dfl}	AUC_{ss}	AUC_{ls}
Original VFA	0.9992	0.9951	0.9983	0.9971	0.9954
Random shift 1 frame	0.9988	0.9922	0.9966	0.9967	0.9905
Random shift 2 frames	0.9988	0.9908	0.9960	0.9965	0.9915

Table 3. Authentication AUCs under variations of viseme segmentation (random temporal shifts).

3. Robustness of Authentication Performance to Viseme Segment Variations

To study how variations in viseme segmentation affect authentication, we compare three segment sources under the *unified-prompts* protocol on VSA dataset: (i) **Original VFA**, the viseme segments from our Visual Forced Aligner; (ii) **VFA + random shift 1 frame**, where each segment is randomly shifted by one frame; and (iii) **VFA + random shift 2 frames**, with two-frame random shifts. Table 3 reports the resulting authentication AUCs across five attack types. Although the average AUC slightly decreases as the random shift increases (from approximately 0.9970 with original segments to about 0.9947 with two-frame shifts), the absolute changes are minimal. This demonstrates that our authentication method is largely insensitive to the exact boundary precision of VFA viseme segmentation.

References

[1] Luca Cappelletta and Naomi Harte. Viseme definitions comparison for visual-only speech recognition. In *2011 19th European Signal Processing Conference*, pages 2109–2113. IEEE, 2011. 1

[2] Linguistic Data Consortium et al. The darpa timit acoustic-phonetic continuous speech corpus. *NIST Speech CD*, pages 1–1, 1990. 1

[3] J. Jeffers and M. Barley. *Speechreading (Lipreading)*. Charles C Thomas Pub Ltd, 1971. 1