2023 IEEE International Conference on Image Processing (ICIP) | 978-1-7281-9835-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICIP49359.2023.10221963

FEW-SHOT LIP-PASSWORD BASED SPEAKER VERIFICATION

Zhikai Hu, Yiu-ming Cheung, Mengke Li, Weichao Lan*

Department of Computer Science, Hong Kong Baptist University, Hong Kong SAR, China

ABSTRACT

Lip-password has provided a promising solution for speaker verification (Liu and Cheung 2014). Despite the potential of this technology, there are few related studies, largely attributed to the lack of corresponding public datasets. Furthermore, previous works in this field generally demand a substantial amount of training samples and negative samples, impeding their applications from a practical perspective. Therefore, this paper collects a lip-password dataset and proposes a novel few-shot lip-password based speaker verification model, which can be effectively deployed in real-world scenarios because only a small number of data are required for training. Specifically, with an analysis of lip-password features, a down-sampling strategy is presented to generate more training samples. To compensate for the information loss caused by this strategy, a few-shot model, consisting of global and local models, is designed to simultaneously verify the global and local information of the lip-password. Speaker identity is verified only if both stages are passed. The efficacy of the proposed method is demonstrated using the newly collected dataset.

Index Terms— Lip-password, speaker verification

1. INTRODUCTION

Biometric-based identification systems, including those based on fingerprints, faces, and irises, have been widely utilized in a variety of applications, such as financial transaction security [1], access control systems [2, 3], and human-computer interaction [4, 5]. Recently, the use of deep learning has led to significant progress in face-based identification systems [6, 7]. However, these static face recognition systems are still vulnerable to external attacks, such as photo attacks [8] and adversarial attacks [9]. One feasible solution to this problem is to leverage dynamic information of the face, such as lip motion [10, 11, 12], which can enhance the security of face recognition systems without requiring additional equipment.

Generally, both the linguistic information and lip features extracted from lip contour of a speaker [13] can be used for identity verification [14]. Along this line, Liu and Cheung [15, 16] have introduced the concept of lip-password, which consists of two parts: the password embedded in the lip motion, and the underlying characteristics of the lip motion. As a result, the lip-password can provide double security to a speaker verification system. That is, only a target person who says the correct password will be verified. Compared to traditional face and speech recognition systems, lip-password based systems have at least three merits: (1) They are insensitive to background noise as audio information is not used, (2) the introduction of dynamic lip information can effectively resist static photo attacks, and (3) even if the password is revealed, the system can still reject imposters who say the correct password based on the underlying characteristics of the lip motion. Considering these advantages, some works have proposed to use the 3D information of lip for text-dependent [17] and text-independent [18] speaker recognition, achieving satisfactory performance.

However, the application of lip-password based speaker verification systems to real-world scenarios is often hindered by several issues. One of the major problems is that these methods typically require a large number of data for training. For instance, Wang et al. [18] required each speaker to read 146 identical sentences. However, long-time recording may result in low-quality data due to fatigue of speakers. Additionally, obtaining negative samples for lip-password are also challenging. Some studies, e.g. see [15, 16], have required users to read some wrong lip-passwords to facilitate model training, which would increase the registration time further. Wang et al. [18] used the information of multiple users in the training stage, which raises concerns about user privacy. Furthermore, 3D information used in this method requires additional equipment for capturing, which increases the overall deployment costs of such systems. In addition, it is worth noting that the datasets used in these studies were collected in controlled indoor environments, with single and stable background conditions, which may limit the generalization ability of the above-mentioned methods to real-world scenarios where illumination conditions can be complex and varied. Therefore, it is crucial to develop more robust and adaptable methods that can cope with these challenges and improve the performance of lip-based speaker recognition systems in practical applications.

To this end, this paper proposes a few-shot lip-password based speaker verification model that minimizes the amount of lip-password data required for model training. To achieve

*YIU-MING CHEUNG IS THE CORRESPONDING AUTHOR.

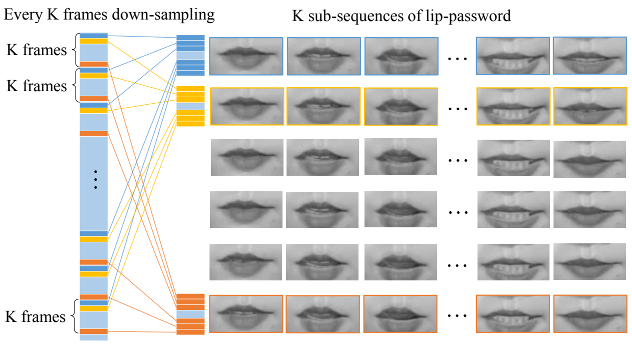

Fig. 1: The lip-password video typically contains a large amount of redundant information. To obtain more training data while avoiding the excessive use of data, we propose a down-sampling strategy in which we extract effective training data from the video every K frames.

this, we introduce a down-sampling strategy that captures the redundant information in the lip-password videos to obtain more training data. To obtain negative samples, we shuffle the order of frames and deform the lips. Furthermore, to compensate for errors resulting from the down-sampling strategy, we propose a few-shot verification model that divides the complete lip-password video into smaller sub-segments and down-samples them to obtain global and local features, respectively. The proposed model verifies the global features first, and then the local features to ensure the accuracy of the lip-password. Our main contributions are: we propose a lip-password verification model that requires only a small number of training samples, thus making it well-suited for deployment in real-world scenarios. To validate the effectiveness of our proposed model, we have collected a dataset that well resembles real-world scenarios.

2. PROPOSED METHOD

During the registration phase, users are required to record N lip-password videos $\mathbf{V} = \{V_1, V_2, ..., V_N\}$. The aim is to establish a discriminative model using these N pieces of data for speaker verification in the subsequent phase. In the verification phase, the proposed model will make a decision on whether a lip-password video V_{cand} should be accepted as a target person or rejected as an imposer.

2.1. Lip-password Video Preprocessing

In general, users expect to complete the registration process as quickly as possible. Thus, the number of training data is often limited, with N typically being less than 10.

The down-sampling technique is usually used to eliminate redundant information in videos, which can speed up video processing. However, in this work, we employ down-sampling to obtain more training samples for the proposed model. It is observed that the movements of lips in adjacent frames exhibit high levels of similarity, as shown in Fig. 1, and we leverage this observation to obtain multiple effective training samples from each video $V_i = [v_1, v_2, ...v_n]$, where n is the frame number of V_i and v_j is the j-th frame of V_i. Specifically, we generate a new piece of training data $x_1 = [v_1, v_{K+1}, ..., v_{K\times(L-1)+1}]$ by selecting one frame every K frames to form a new video of L frames. When we start down-sampling from different frames, we can get K different training data $x_i = [v_i, v_{K+i}, ..., v_{K\times(L-1)+i}] (i = 1, 2, ..., K)$. As illustrated in Fig. 1, the resulting K pieces of data have subtle differences, e.g. different degrees of mouth opening, while maintaining a high degree of similarity. This strategy removes redundant information in the lip-password videos, but yields more effective training data. We label these data as '$+1$' to form the positive training samples $\{x, y = +1\}$.

To obtain negative training samples, two different strategies are designed to simulate different attack scenarios, respectively. Firstly, we shuffle the order of the lip-password video frames and down-sample L frames from it to imitate the target person saying the wrong password, i.e., $x^r = [v_{\lfloor rand()\times N\rfloor}, ..., v_{\lfloor rand()\times N\rfloor}]$, where $rand()$ is a random number from 0 to 1. Secondly, we apply a deformation to the lip, e.g. vertical stretching, in the positive training samples to simulate the imposter saying the correct password, i.e., $x_i^{'} = [v_i^{'}, v_{K+i}^{'}, ..., v_{K\times(L-1)+i}^{'}]$, where $v_i^{'}$ denotes the deformation of v_i. We label these data '-1' to form the negative training samples $\{x, y = -1\}$. By combining these two sets of samples, we obtain a complete training set.

2.2. Proposed Model

Since we have employed down-sampling to obtain both positive and negative training samples, this could result in imposters being mistakenly verified as the target person due to the loss of information. To ensure the security of the lip-password system, we have designed a few-shot model comprising of global and local models to provide a double verification of the lip-password. In the first stage, we focus on extracting the global features from the training data. Therefore, we directly down-sample the lip-password videos $\mathbf{V}$ to obtain the training data $\{x, y\}$. As the contour and orientation features of lips are crucial for lip-password verification [19], we extract the hog (histogram of oriented gradients) feature, denoted as $f(x)$, from the down-sampled training data x. Next, we utilize the extracted $f(x)$ and the corresponding label y to train a global classifier

$$y_{pred1} = \mathbf{G}_{global}(f(x); \Theta_{global}), \tag{1}$$

where $\mathbf{G}_{global}$ and Θ_{global} denote the global classifier and its parameters, respectively. Given a sample that is predicted to be positive by the global classifier $\mathbf{G}_{global}$, we need a second stage of validation to further verify whether it is a true positive or not, due to the loss of details during the down-sampling process.

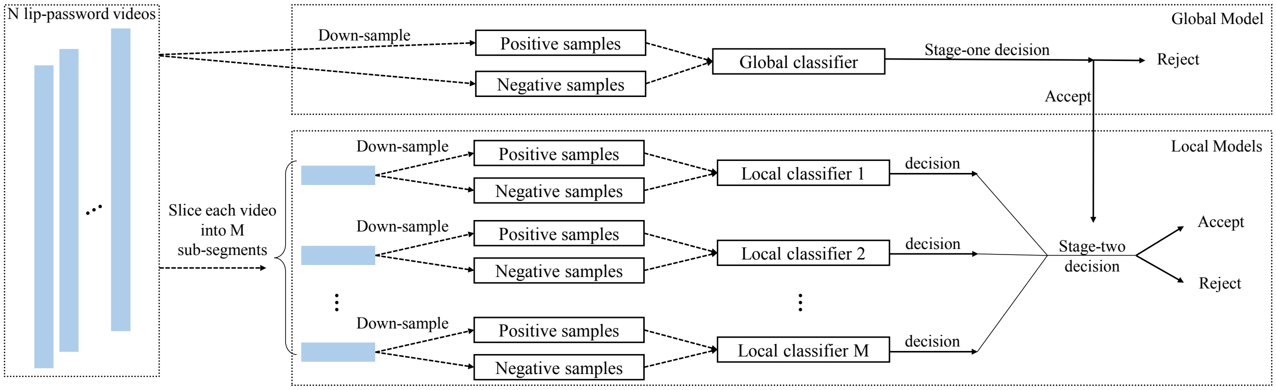

Fig. 2: In the first stage, we down-sample multiple training samples by extracting a frame every K frames from the original lip-password videos and train a global classifier. Given a sample to be validated, once the global classifier predicts it as a positive sample, the local verification module will be triggered. In this stage, we divide each lip-password video into M sub-segments and train M local classifiers for each segment. The predictions of all local classifiers are combined to make the final decision.

For the second stage, we divide each lip-password video into M sub-segments $V_i = [V_i^{(1)}, V_i^{(2)}, ..., V_i^{(M)}]$, where $V_i^{(j)} = [v_{j\times\lfloor N/M\rfloor+1}, v_{j\times\lfloor N/M\rfloor+2}, ..., v_{(j+1)\times\lfloor N/M\rfloor}]$. For each sub-segment $V_i^{(j)}$, we also employ a down-sampling strategy to obtain training data $\{x^{(j)}, y^{(j)}\}$. As $V_i^{(j)}$ represents a local segment of the lip-password, this stage captures more detailed information when down-sampling $V_i^{(j)}$ than in the first stage. For each sub-segment, we train a local classifier

$$y_{pred}^{(j)} = \mathbf{G}_{local}^{(j)}(f(x^{(j)}); \Theta_{local}^{(j)}), \quad j = 1, 2, ..., M, \qquad (2)$$

where $\mathbf{G}_{local}^{(j)}$ and $\Theta_{local}^{(j)}$ denote the j-th local classifier of sub-segment $V_i^{(j)}$ and its corresponding parameters, respectively. The final decision of the second stage depends on the prediction results of all local classifiers. That is,

$$y_{pred2} = \sigma(\frac{1}{M}\sum_{j}^{M} y_{pred}^{(j)}), \sigma(x) = \left\{ \begin{array}{ll} -1 & ,\tau > x \\ 1 & ,\tau \leq x \end{array} \right. , \qquad (3)$$

where τ controls the strictness of the models.

2.3. User Verification

In the verification phase, given a lip-password video V_{cand} to be verified, the proposed model will down-samples $Vcand$ to obtain K pieces of data $\{x_1, x_2, ..., x_K\}$. These data are then fed into $\mathbf{G}_{global}$ to make the first-stage decision:

$$y_{pred1} = (\sum_{i}^{K} \mathbf{G}_{global}(f(x_i); \Theta_{global})) > 0. \qquad (4)$$

If more than half of the K pieces of data are classified as positive samples by $\mathbf{G}_{global}$, V_{cand} will pass the verification in the first stage; otherwise, it will be rejected.

Once V_{cand} has passed the initial verification in the first stage, it is then segmented into M sub-segments $V_{cand} = [V_{cand}^{(1)}, V_{cand}^{(2)}, ..., V_{cand}^{(M)}]$. For each sub-segment $V_{cand}^{(j)}$, K pieces of data $\{x_1^{(j)}, x_2^{(j)}, ..., x_K^{(j)}\}$ are down-sampled and fed into the corresponding local classifier $\mathbf{G}_{local}^{(j)}$. The final decision in the second stage is made by aggregating the prediction results from all K pieces of down-sampled data $x_i^{(j)}$. That is,

$$y_{pred2} = \sigma(\frac{1}{M}\sum_{j}^{M}(\sum_{i}^{K} \mathbf{G}_{local}^{(j)}(f(x_i^{(j)}); \Theta_{local}^{(j)})) > 0). \quad (5)$$

Finally, the speaker will be verified as target person only if V_{cand} passes the verification of both stages, i.e., $y_{pred1} = y_{pred2} = 1$.

3. EXPERIMENT

3.1. Dataset and Evaluation Metrics

We collected a lip-password dataset containing 30 speaker[1]. The i-th speaker was required to read four kinds of lip-passwords: 1) $\mathcal{P}_i^1$ only includes the correct lip-password 4536708. The speaker was asked to repeat it in English for 16 times, with 6 recordings used for training and the remaining 10 used for testing. 2) $\mathcal{P}_i^2$ consists of one-number-wrong passwords generated by randomly modifying any one number in the correct lip-passwords, e.g., $4\underline{3}36708$ and $453\underline{9}708$. Speaker was asked to read 10 different one-number-wrong passwords in English. 3) $\mathcal{P}_i^3$ consists of two-number-wrong password generated by randomly modifying any two numbers in the correct lip-password, e.g., $4\underline{3}3670\underline{9}$ and $453\underline{9}7\underline{1}8$. Speaker was asked to read 10 different two-number-wrong passwords in English. All of them were used as test set. 4) $\mathcal{P}_i^4$ consists of random-wrong passwords, e.g., 1234567. Speaker

[1] Volunteers recorded data using their personal mobile phone's front-facing cameras, thus the background and lighting were more random.

Table 1: The FAR of three scenarios including imposers saying wrong password, imposers saying correct password, and target person saying wrong password and FRR of target person saying the correct password.

Person		Password		FAR/FRR(%)(↓)	
Target	Imposer	Correct	Wrong	[16]	Ours
	✓		✓	0.00	0.00
	✓	✓		2.41	0.00
✓			✓	12.67	7.11
✓		✓		67.67	35.67

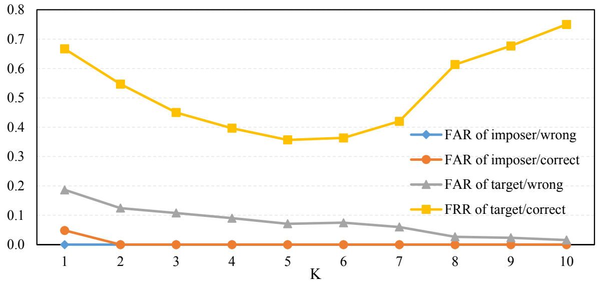

Fig. 3: The FAR and FRR of four scenarios with different K.

Table 2: The mean FAR and FRR of four scenarios of each speaker with different number of local classifier.

	FAR(%)(↓)			FRR(%)(↓)
M	$\mathcal{P}^4$	$\mathcal{P}^3$	$\mathcal{P}^2$	$\mathcal{P}^1$
0	4.00	22.33	29.67	15.00
3	0.00	6.00	15.33	35.67
5	0.00	0.00	0.00	75.67

was asked to read 10 different random-wrong passwords in English. All of $\mathcal{P}_i^{2,3,4}$ were used as test set.

We utilized the False Acceptance Rate (FAR=$\frac{\#FalseAccept}{\#TotalFalse}$) to evaluate three scenarios: imposers saying the wrong password, imposers saying the correct password, and the target person saying the wrong password. For the i-th speaker, the test sets are $\bigcup_k^{2,3,4} \bigcup_{j \neq i} \mathcal{P}j^k$, $\bigcup_{j \neq i} \mathcal{P}j^1$, and $\bigcup k^{2,3,4} \mathcal{P}_i^k$ for these three scenarios. We utilized the False Rejection Rate (FRR=$\frac{\#FalseReject}{\#TotalTrue}$) to evaluate the scenario of the target person saying the correct password, where only $\mathcal{P}_i^1$ is used as the test set for the i-th speaker. When training global classifier, we set $L = 50$ and $K = 5$. When training local classifiers, we empirically set $M = 3$, $L = 20$ and $K = 5$.

3.2. Analysis of Experimental Results

We first compare the proposed model with the only 2-D lip-password based method [16]. The results are reported in Table 1. For the threshold settings in both methods, we have followed the principle of selecting the minimum threshold τ in the case where the maximum FAR is achieved for scenarios where imposters say the correct or wrong lip-password. This is because, in practical scenarios, identifying imposters as the target person is generally more harmful than identifying the target person as the imposter. It is worth noting that even at the maximum FAR, [16] still recognizes a very small number of imposters saying the correct password as the target person. By contrast, the proposed method can reject all imposters saying the correct or wrong password in all cases we have tried so far. Moreover, the proposed method can more effectively reject the target person when they say the wrong password. In the case where the target person says the correct password, the pass rate of our proposed method is also much higher than that of [16].

Further, we evaluated the effect of the down-sampling parameter K on the performance of the system, whose results are shown in Figure 3. It is observed that selecting a suitable value of K in the range of 4 to 7 can effectively improve the system's recognition performance. Otherwise, a large value of K (i.e. $K > 7$) will have a negative impact. This is because increasing K will weaken similarity between the down-sampled positive samples, thus leading to decreased performance. Besides, we tested the effect of the local classifiers on the performance by reporting the mean FAR and FRR of four scenarios for each speaker in Table 2. In this experiment, we focused only on the scenarios where the target person is saying the correct or wrong password. The results show that the incorporation of local classifiers can significantly reduce the FAR on the test sets $\mathcal{P}_{2,3,4}$, but also lead to an increase in the threshold for accepting the target person.

4. CONCLUSION AND FUTURE WORK

In this paper, we have proposed a few-shot model for efficient lip-password based speaker verification. Our down-sampling strategy and global-local model architecture compensate for the information loss caused by limited training data. The proposed method has achieved promising results and can serve as a practical solution for secure speaker verification with low training overhead.

Although this paper solves the problem of small-sample training in lip-password speaker verification, there still exist some challenges from a practical perspective. For example, speakers may not always speak at a consistent speed. In this paper, we required volunteers to maintain a constant speed as much as possible during data recording to mitigate this issue. Furthermore, the passwords used in this paper are limited to English numbers. In practice, users may prefer more flexible password settings, e.g. text-independent and language-independent. To address these issues, we plan to collect more diverse data in future work and develop a lip-password speaker verification system that is less restricted and more adaptable to user preferences.

5. REFERENCES

[1] Abhinav Muley and Vivek Kute, "Prospective solution to bank card system using fingerprint," in *2018 2nd International Conference on Inventive Systems and Control (ICISC)*. IEEE, 2018, pp. 898–902.

[2] Yongjian Zhao and Rui Guo, "Hybrid techniques for identity authentication," in *2019 11th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA)*. IEEE, 2019, pp. 62–65.

[3] Meng Pang, Yiu-ming Cheung, Binghui Wang, and Risheng Liu, "Robust heterogenous discriminative analysis for face recognition with single sample per person," *Pattern Recognition*, vol. 89, pp. 91–107, 2019.

[4] Mohamed Hamidi, Hassan Satori, Ouissam Zealouk, Khalid Satori, and Naouar Laaidi, "Interactive voice response server voice network administration using hidden markov model speech recognition system," in *2018 Second World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4)*. IEEE, 2018, pp. 16–21.

[5] Shih-Chung Hsu, Yu-Wen Wang, and Chung-Lin Huang, "Human object identification for human-robot interaction by using fast r-cnn," in *2018 Second IEEE International Conference on Robotic Computing (IRC)*. IEEE, 2018, pp. 201–204.

[6] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690–4699.

[7] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu, "Cosface: Large margin cosine loss for deep face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5265–5274.

[8] Roberto Tronci, Daniele Muntoni, Gianluca Fadda, Maurizio Pili, Nicola Sirena, Gabriele Murgia, Marco Ristori, Sardegna Ricerche, and Fabio Roli, "Fusion of multiple clues for photo-attack detection in face recognition systems," in *2011 International joint conference on biometrics (IJCB)*. IEEE, 2011, pp. 1–6.

[9] Stepan Komkov and Aleksandr Petiushko, "Advhat: Real-world adversarial attack on arcface face id system," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 819–826.

[10] Chi Ho Chan, Budhaditya Goswami, Josef Kittler, and William Christmas, "Local ordinal contrast pattern histograms for spatiotemporal, lip-based speaker authentication," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 2, pp. 602–612, 2011.

[11] Maycel-Isaac Faraj and Josef Bigun, "Synergy of lip-motion and acoustic features in biometric speech and speaker recognition," *IEEE transactions on computers*, vol. 56, no. 9, pp. 1169–1175, 2007.

[12] Mustafa Nazmi Kaynak, Qi Zhi, Adrian David Cheok, Kuntal Sengupta, Zhang Jian, and Ko Chi Chung, "Analysis of lip geometric features for audio-visual speech recognition," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 34, no. 4, pp. 564–570, 2004.

[13] Xin Liu, Yiu-ming Cheung, Meng Li, and Hailin Liu, "A lip contour extraction method using localized active contour model with automatic parameter selection," in *2010 20th International Conference on Pattern Recognition (ICPR)*. IEEE, 2010, pp. 4332–4335.

[14] Maycel Isaac Faraj and Josef Bigun, "Person verification by lip-motion," in *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)*. IEEE, 2006, pp. 37–37.

[15] Xin Liu and Yiu-ming Cheung, "A multi-boosted hmm approach to lip password based speaker verification," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 2197–2200.

[16] Xin Liu and Yiu-ming Cheung, "Learning multi-boosted hmms for lip-password based speaker verification," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 2, pp. 233–246, 2014.

[17] Jianguo Liao, Shilin Wang, Xingxuan Zhang, and Gongshen Liu, "3d convolutional neural networks based speaker identification and authentication," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 2042–2046.

[18] Jianrong Wang, Tong Wu, Shanyu Wang, Mei Yu, Qiang Fang, Ju Zhang, and Li Liu, "Three-dimensional lip motion network for text-independent speaker recognition," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 3380–3387.

[19] Krzysztof Wrobel, Rafal Doroz, Piotr Porwik, Jacek Naruniec, and Marek Kowalski, "Using a probabilistic neural network for lip-based biometric verification," *Engineering Applications of Artificial Intelligence*, vol. 64, pp. 112–127, 2017.