WhisperNet: Deep Siamese Network For Emotion and Speech Tempo Invariant Visual-Only Lip-Based Biometric

Abdollah Zakeri
a.zakeri@shahroodut.ac.ir
Faculty of Computer Engineering, Shahrood University of Technology

Hamid Hassanpour
h.hassanpour@shahroodut.ac.ir
Faculty of Computer Engineering, Shahrood University of Technology

2021 7th International Conference on Signal Processing and Intelligent Systems (ICSPIS) | 978-1-6654-0938-4/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICSPIS54653.2021.9729394

Abstract— **In the recent decade, the field of biometrics was revolutionized thanks to the rise of deep learning. Many improvements were done on old biometric methods which reduced the security concerns. Before biometric people verification methods like facial recognition, an imposter could access people's vital information simply by finding out their password via installing a key-logger on their system. Thanks to deep learning, safer biometric approaches to person verification and person re-identification like visual authentication and audio-visual authentication were made possible and applicable on many devices like smartphones and laptops. Unfortunately, facial recognition is considered to be a threat to personal privacy by some people. Additionally, biometric methods that use the audio modality are not always applicable due to reasons like audio noise present in the environment. Lip-based biometric authentication (LBBA) is the process of authenticating a person using a video of their lips' movement while talking. In order to solve the mentioned concerns about other biometric authentication methods, we can use a visual-only LBBA method. Since people might have different emotional states that could potentially affect their utterance and speech tempo, the audio-only LBBA method must be able to produce an emotional and speech tempo invariant embedding of the input utterance video. In this article, we proposed a network inspired by the Siamese architecture that learned to produce emotion and speech tempo invariant representations of the input utterance videos. In order to train and test our proposed network, we used the CREMA-D dataset and achieved 95.41% accuracy on the validation set. Keywords—Biometrics; Deep Siamese Network; Lip-Based Biometrics; Video Processing**

I. INTRODUCTION

In recent years, biometric approaches to person verification and person re-identification have attracted interest due to their enhanced security compared to normal passwords. Biometric identifiers are categorized into two main groups, namely physiological and behavioral traits. The former includes several characteristics such as the face, fingerprint, iris, palm print, etc., and the latter includes handwriting, signature, gait, etc.

In some cases, a single characteristic would not suffice for the biometric person authentication system due to its lack of discriminative power. For example, signature authentication is less fraud-resistant compared to fingerprint authentication because a fake signature can be easily forged in order to bypass a signature verification system, and hence, it is not secure enough. Additionally, some traits like handwriting simply do not have the required discriminative power to distinguish a large number of people because when sampling the handwriting of a large enough group of people, we would certainly observe cases that are similar enough to deceive a handwriting recognition system.

One of the traditional methods for person authentication is the use of their acoustic signals as a unique trait [1]. Although this method is fast and computationally affordable, it does not consider the problem of liveness and hence, doesn't have any robustness against recorded voices [2].

Another old and widely-used method of person authentication is authentication based on visual data e.g., face images. Many approaches provide methods for authenticating people using their faces. One-Shot Learning Perspective [3] and transfer learning [4] are two of the famous deep learning-based approaches to visual authentication. Facial recognition modality is widely used as an authentication method thanks to its high discriminative power. Nevertheless, there are some downsides to facial recognition like some people considering it a threat to personal privacy and hence, do to consent to upload an image/video of their faces into a database. Additionally, as a result of recent advances in Generative Adversarial Networks (GANs), forging fake videos of people talking has become quite easy and this would threaten the security of a facial recognition system [5]. Furthermore, processing a video of a face requires an abundance of processing power which reduces its applicability.

In order to enhance the security of a biometric system, we can take advantage of multiple traits simultaneously. For instance, we can use both visual and audio modalities in order to authenticate a person by processing a video of the person's face while uttering a secret phrase [6]. Although the aforementioned method will enhance security, it will result in several issues due to dependence on the audio modality like lacking the required quality and discriminative power because of the audio noises present in the environment. Additionally, using the audio modality is impossible when a person is in a library or is speech impaired. Furthermore, uttering the secret phrase out loud would jeopardize the security since other people who are present will hear it. Consequently, finding a way to achieve the goal of high accuracy person re-identification without the audio modality is a necessity.

978-1-6654-0938-4/21/$31.00 ©2021 IEEE

In order to solve the aforementioned issues, we can use an authentication method leveraging a video of the person's lips in which they utter a secret passphrase silently instead of the multimodal facial recognition [1][7][8]. This method will combine physiological traits like the shape of a person's lips and mouth with behavioral traits like the way the person moves their lips while whispering the secret phrase. Each person might utter the same word in a unique way. This method of authentication does not need to know the meaning of the secret phrases that people use, and as far as the authentication method concerns, the secret phrases could be meaningless phrases, but with a minimum length criterion. This method will not only have sufficient discriminative power thanks to the uniqueness of the combination of physiological and behavioral traits, but also will not suffer from the aforementioned issues because considering that the phrase is whispered silently, the audio modality will not be present or used and therefore, all of the issues caused by the audio modality will be resolved. Additionally, the movements of the lips are not related to the audio and as a result, removing the audio will not affect the movement of the lips and makes it applicable to speech-impaired people. Moreover, this method only requires a video of the person's lips and does not threaten their privacy as much as a video of their full face would.

When lip-based biometric authentication is employed in a real-world application like on mobile devices, the authentic person, i.e., the owner of the mobile phone who should be successfully authenticated, might have different emotional states that might affect their utterance of the secret phrase or change their speech tempo. Numerous valuable researches were conducted on the subject of lip-based biometric authentication [1][7][8], but to the best of our knowledge, none of them considered the challenges of different emotional states that the person might have while uttering the secret phrase or their speech tempo.

II. Dataset

In order to train our model, we used the CREMA-D dataset [10]. This dataset consists of 7,442 videos of 91 different people, uttering 12 different phrases with 6 emotional states including neutral, happy, sad, anger, disgust, and fear. Nevertheless, there was less available data for three of these people, hence the data from the remaining 88 people were used to form our initial dataset. Out of these 88 people, 66 were used for training, 11 were used for validation and the remaining 11 were used for test. Since there are no common people among train and test data, the trained model would be evaluated with unseen data.

III. Proposed Method

A. Pre-Processing

Our dataset consisted of full-face videos of people uttering phrases; thus, we had to extract the lip Region Of Interest (ROI) from the videos and crop each one to that ROI so that our embedding network would not use any data other than lips. To extract the lip ROI, we first needed to localize the lips in the input image and extract the lip landmarks; thus, we trained a landmark extraction network that took a single frame of the video that contained the lip region and outputted 24 different coordinates (x, y pairs) which represented the lips.

Our landmark extraction network used multiple convolutional and pooling layers to extract the desired landmarks from the input lip image and was trained using 300W dataset [9]. Furthermore, this network did not need the full face in order to output the desired lip landmarks and any input image that contained the lip region sufficed for the landmark extraction task. A sample output of the landmark extraction network is shown in Fig. 1.

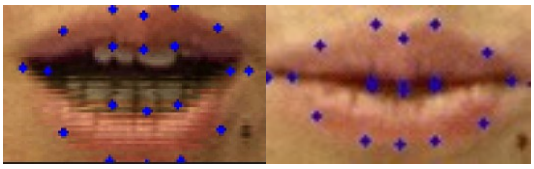

Fig. 1. Output of the landmark extraction network. The network only outputs 24 (x,y) pairs, not the image above. The output landmarks are plotted on the input image and exhibited here for clarity.

In order to calculate the proper bounding box (ROI) for cropping the input image, we take advantage of these 24 points:

$$Landmarks = \{\langle x, y\rangle_i\}_{i \in \{1,\dots,24\}}$$

$$\begin{cases} x_{min} = \min\{x \mid \langle x, y\rangle_i \in Landmarks\} \\ x_{max} = \max\{x \mid \langle x, y\rangle_i \in Landmarks\} \end{cases}$$

$$\begin{cases} y_{min} = \min\{y \mid \langle x, y\rangle_i \in Landmarks\} \\ y_{max} = \max\{y \mid \langle x, y\rangle_i \in Landmarks\} \end{cases}$$

$$\begin{cases} BoundingBoxTopLeft = \langle x_{min}, y_{min}\rangle \\ BoundingBoxBottomRight = \langle x_{max}, y_{max}\rangle \\ BoundingBoxWidth = x_{max} - x_{min} \\ BoundingBoxHeight = y_{max} - y_{min} \end{cases}$$

$$BoundingBoxAspectRatio = \frac{BoundingBoxWidth}{BoundingBoxHeight}$$

While observing the distribution of different calculated bounding-box aspect ratios, we observed that the value $\frac{5}{3}$ has the highest absolute frequency, hence we defined the aforementioned value as the main aspect ratio and tried to surround the lip region in a bounding box with this aspect ratio for all frames. For the aspect ratio of all the frames to match this value, we slightly expanded the width or the height of the bounding box:

$$\begin{cases} W_{new} = H_{current} * \frac{5}{3} & if\ AspectRatio < \frac{5}{3} \\ H_{new} = W_{current} * \frac{3}{5} & if\ AspectRatio > \frac{5}{3} \end{cases}$$

Finally, we resized all the extracted bounding boxes to a size of (30×18). This size was chosen because it was the minimum size of a bounding box containing a lip, and resizing all other bounding boxes to this minimum size would minimize the amount of interpolation during the resizing process. Additionally, we saved the sequence of extracted ROIs with the desired size and aspect ratio along with the sequence of

extracted landmark points for each video to form our new dataset. Each sample of our new dataset consists of a video of the extracted ROIs and a sequence of landmarks. Although different samples of the dataset did not have the similar number of frames and landmark sequence lengths, the number of frames in each video was equal to the length of the sequence of extracted landmarks corresponding to the same sample.

B. *Embedding Network Architecture*

During the past decade, deep learning has revolutionized numerous fields including computer vision, time-series processing, biometrics, etc. Due to the high potential of deep learning methods, especially in tasks related to image or video processing, we decided to take advantage of deep learning in order to improve the current solutions to the problem of person re-identification using visual lip passwords.

Our proposed network is inspired by the Siamese network architecture with the triplet loss function [11][12]. The Siamese network maps the input data to the latent space using an embedding network. Leveraging this embedding network, the Siamese network learns to discriminate different input data. For example, in our use-case, the network must learn to differentiate between people and the phrases they utter. The embedding network performs the task of extracting spatio-temporal features from the input data and since the data is sequential, the embedding network must be able to perform the task of feature extraction for sequences with arbitrary lengths.

Our embedding network was inspired by the LipNet architecture which was used for visual speech recognition [13]. This network consists of three Spatio-Temporal CNNs (STCNN) followed by two bi-directional Gated Recurrent Unit (GRU) layers that extract spatio-temporal features from the input video sequence. STCNNs are a variation of CNNs that have the ability to process videos by adding a summation over time [13]. Given an input video $x \in R^{C\times T\times W\times H}$ to an STCNN layer with f number of filters with size $k_t \times k_w \times k_h$, the output volume is computed as follows:

$$[stcnn(x,w)]_{ftij} = \sum_{c=1}^{C}\sum_{t'=1}^{T}\sum_{i'=1}^{W}\sum_{j'=1}^{H} w_{fct'i'j'}x_{c,t+t',i+i',j+j'},$$

where x_{ctij} represents the pixel at location i, j in the cth channel of the video frame at time-step t, and $w_{fct'i'j'}$ represents the STCNN layer weights.

There are two branches in our proposed embedding network. The first one extracts spatio-temporal features using the STCNN layers from the visual lip data and the second one extracts spatio-temporal features from lip landmark data using 3x Time distributed dense layers followed by a Long Short-Term Memory (LSTM) layer. The results of these two branches are then concatenated and processed further to produce the final representation vector. The architecture of our proposed embedding network is presented in Fig. 2.

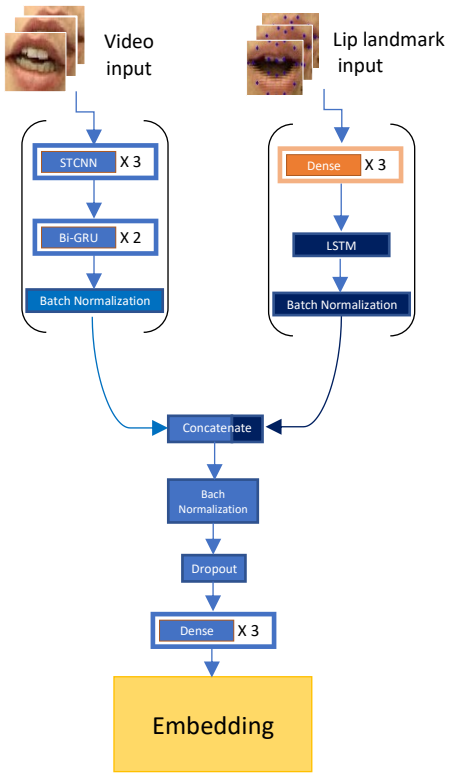

Fig. 2. The embedding network architecture for generating the representation (embedding) vector from the input sequences

C. *Siamese Network Architecture*

The Siamese network with triplet loss function takes a triplet of anchor, positive and negative embeddings and tries to simultaneously minimize the distance between the anchor and positive embeddings while maximizing the distance between the anchor and negative embeddings. The triplet loss is calculated as follows:

$$L(A,P,N) = \max(0, D(A,P) - D(A,N) + margin),$$

where $D(x,y)$ is the distance metric used to calculate the distance between x and y. L2 distance or (1 – cosine similarity) can be used as the distance metric. The margin term represents the minimum required distance between $D(A,P)$ and $D(A,N)$. Our experiments suggested that setting this term to a high value may cause overfitting of the Siamese network. The architecture of the used Siamese network is represented in Fig. 3.

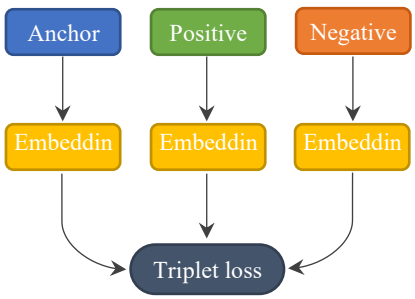

Fig. 3. The Siamese network architecture

D. Triplet selection

Considering that we choose a video of person *P,* uttering a phrase *S* to be the anchor video, all the other utterances of *S* done by *P*, are considered to be positive. There are three methods of selecting the negative videos to generate a triplet:

(i). **None-Selective approach:** In this approach, all other videos, including videos of phrases other than *S*, uttered by *P* are considered to be negative. In this case, the total number of train triplets would be:

$$\underbrace{(66 \times 12 \times 6)}_{Anchor} \times \underbrace{5}_{Positive} \times \underbrace{\left(\underbrace{(65 \times 12 \times 6)}_{\substack{Other\ People,\\ all\ Phrases}} + \underbrace{(1 \times 11 \times 6)}_{\substack{Same\ Person,\\ Other\ Phrases}} \right)}_{Negative}$$

$$= 112{,}764{,}960$$

(ii). **Selective Approach:** In this approach, instead of using all the other phrases of all other people, we only select videos of phrase *S*, uttered by people other than *P*. In this case, the total number of train triplets would be:

$$\underbrace{(66 \times 12 \times 6)}_{Anchor} \times \underbrace{5}_{Positive} \times \underbrace{\left(\underbrace{(65 \times 1 \times 6)}_{\substack{Other\ People,\\ Same\ Phrase}} + \underbrace{(1 \times 11 \times 6)}_{\substack{Same\ Person,\\ Other\ Phrases}} \right)}_{Negative}$$

$$= 10{,}834{,}560$$

(iii). **Combination of Both:** This approach is a combination of the other two meaning that we select some of the phrases including *S,* uttered by people other than *P* as negative samples. In this case, the total number of train triplets would be:

$$\underbrace{(66 \times 12 \times 6)}_{Anchor} \times \underbrace{5}_{Positive} \times \underbrace{\left(\underbrace{(65 \times 6 \times 6)}_{\substack{Other\ People,\\ 6\ Phrases\ including\ S}} + \underbrace{(1 \times 11 \times 6)}_{\substack{Same\ Person,\\ Other\ Phrases}} \right)}_{Negative}$$

$$= 57{,}166{,}560$$

Since the same person should be authenticated only when they utter the same phrase, utterance videos of phrases other than *S*, done by *P*, are considered to be negative in all approaches. But these triplets are considered to be the hardest triplets for the network because the network cannot use any physiological traits to distinguish between the negative and positive samples. On the other hand, when we use the video of the same phrase uttered by people other than *P* as the negative sample, the network cannot take much advantage of behavioral traits like the lip dynamics since the lip movements while uttering the sample phrase are similar among different people. So, these triplets are considered to be hard as well. According to our experiments, the network performs the best when using a combination of hard and easy triplets.

E. Accuracy metric

In order to have a better understanding of the network's performance, we did various tests on different functions to obtain an accuracy metric. We employed the distance metrics used in the loss function to calculate the distance between anchor-positive and anchor-negative pairs to get a sense of how good the representations were.

$$D_{PN} = D(A,P) - D(A,N).$$

Since the network must learn to minimize $D(A,P)$ while maximizing $D(A,N)$, the distance D_{PN} must be negative so that the produced embeddings are acceptable. But in this case, being acceptable is not binary, meaning that D_{PN}'s higher negative value results in more acceptable embeddings. If D_{PN} has a high negative value, this means that not only $D(A,P)$ is less than $D(A,N)$ and the anchor is far more similar to the positive comparing to the negative, but there is a noticeable difference between them so the network must achieve high accuracy. On the other hand, if $D(A,P)$ has a high positive value, the network failed to produce acceptable embeddings of the input data, so the accuracy must be fairly low.

There is also a third case that occurs when D_{PN} is near zero and the network cannot distinguish between the positive and the negative embeddings.

Having the aforementioned points in mind, it is logical to use the negative of sigmoid function to calculate the desired accuracy since it has the expected behavior:

$$ACC = \frac{1}{1 + e^{2 \times (D(A,P) - D(A,N))}}$$

Our accuracy metric's plot is demonstrated in Fig. 4

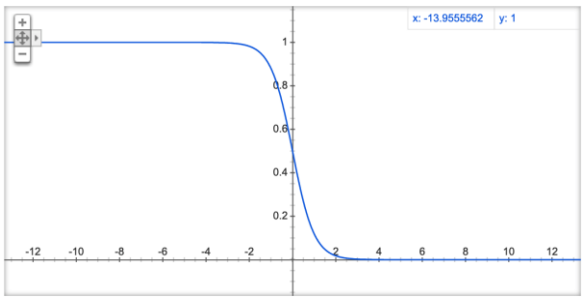

Fig. 4. Accuracy metric plot. In this plot, the X and Y axes represent D_{PN} and the corresponding accuracy value respectively.

IV. Results

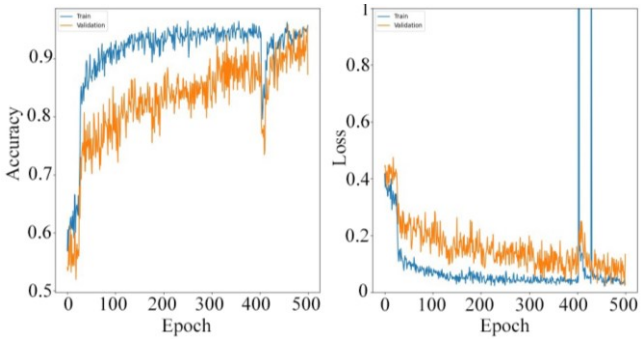

Fig. 5. Model training accuracy and loss plots

Our proposed Siamese network achieved 95.41% accuracy on validation set and 96.57% accuracy on train set while trained on CREMA-D dataset (Fig. 5). These results demonstrate that our proposed method has the potential to be used in a real-word application like on mobile devices since our model is able to

distinguish between the authentic person and an imposter due to the fact that the validation set and the train set were built on disjoint sets of people. Additionally, there are several utterances of the same phrase by the same person, with different speech tempos and various emotional states present in the dataset and our network learned to overcome these challenges to a certain point, and had the ability to produce an emotional and speech tempo invariant embedding of the input video.

V. Conclusion

In this article, we introduced WhisperNet, a deep Siamese network which can be used for the purpose of visual-only lip-based biometrics. We used the CREMA-D dataset to train and test our network and we achieved 95.41% accuracy on the test set. Considering that the CREMA-D dataset consists of videos of different people uttering phrases with various emotional states and speech tempos, since our network's performance was acceptable on this dataset, we can conclude that our network has learned to distinguish people, being inattentive of their emotions and their speech tempo which may affect their utterance.

VI. References

[1] Lu, L., Yu, J., Chen, Y., Liu, H., Zhu, Y., Liu, Y., & Li, M. (2018). "LipPass: Lip Reading-based User Authentication on Smartphones Leveraging Acoustic Signals. IEEE INFOCOM 2018 - IEEE Conference on Computer Communications." 1466–1474. doi: 10.1109/INFOCOM.2018.8486283

[2] J. Gałka, M. Masior and M. Salasa, "Voice authentication embedded solution for secured access control." in IEEE Transactions on Consumer Electronics, vol. 60, no. 4, pp. 653-661, Nov. 2014, doi: 10.1109/TCE.2014.7027339.

[3] Chanda, Sukalpa, GV AsishChakrapani, Anders Brun, Anders Hast, Umapada Pal and David S. Doermann. "Face Recognition - A One-Shot Learning Perspective." 15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS) (2019): 113-119.

[4] Mehdipour Ghazi, Mostafa & Ekenel, Hazım. (2016). "A Comprehensive Analysis of Deep Learning Based Representation for Face Recognition." doi: 10.1109/CVPRW.2016.20.

[5] K. Vougioukas, S. Petridis, and M. Pantic, "Realistic Speech-Driven Facial Animation with GANs," International Journal of Computer Vision, vol. 128, no. 5, pp. 1398–1413, May 2020, doi: 10.1007/s11263-019-01251-8.

[6] Q. Memon, Z. AlKassim, E. AlHassan, M. Omer, and M. Alsiddig, "Audio-Visual Biometric Authentication for Secured Access into Personal Devices," in Proceedings of the 6th International Conference on Bioinformatics and Biomedical Science, 2017, pp. 85–89. doi: 10.1145/3121138.3121165.

[7] C. Wright and D. Stewart, "Understanding visual lip-based biometric authentication for mobile devices," EURASIP Journal on Information Security, vol. 2020, Mar. 2020, doi: 10.1186/s13635-020-0102-6.

[8] X. Liu and ming Yiu-Cheung, "Learning Multi-Boosted HMMs for Lip-Password Based Speaker Verification," IEEE Transactions on Information Forensics and Security, vol. 9, pp. 233–246, 2014.

[9] Sagonas, C., Antonakos, E., Tzimiropoulos, G., Zafeiriou, S., & Pantic, M. (2016). "300 Faces In-The-Wild Challenge: database and results. Image and Vision Computing", 47, 3–18. doi: 10.1016/j.imavis.2016.01.002

[10] Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., & Verma, R. (2014). "CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset." IEEE transactions on affective computing, 5(4), 377–390. doi:10.1109/TAFFC.2014.2336244

[11] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature Verification Using a 'Siamese' Time Delay Neural Network." in Proceedings of the 6th International Conference on Neural Information Processing Systems, 1993, pp. 737–744.

[12] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 2005, vol. 1, pp. 539–546 vol. 1. doi: 10.1109/CVPR.2005.202.

[13] Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas, "LipNet: End-to-End Sentence-level Lipreading," arXiv: Learning, 2016.

[14] Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., & Chen, T. (2018). "Recent advances in convolutional neural networks." Pattern Recognition, 77, 354–377. doi: 10.1016/j.patcog.2017.10.013