One-Shot-Learning for Visual Lip-Based Biometric Authentication

Carrie Wright(✉) and Darryl Stewart

Queen's University Belfast, Belfast, N. Ireland
cwright32@qub.ac.uk
https://www.qub.ac.uk/schools/eeecs/

Abstract. Lip-based biometric authentication is the process of verifying an individual's identity based on visual information taken from lips whilst speaking. To date research in this area has involved more traditional approaches and inconsistent results that are difficult to compare. This work aims to push the field forward through the application of deep learning. A deep artificial neural network using spatiotemporal convolutional and bidirectional gated recurrent unit layers is trained end-to-end. For the first time one-shot-learning is applied to lip-based biometric authentication by implementing a siamese network architecture, meaning the model only needs a single prior example in order to authenticate new users. This approach sets a new state-of-the-art performance for lip-based biometric authentication on the XM2VTS dataset and Lausanne protocol with an equal error rate of 0.93% on the evaluation set and a false acceptance rate of 1.07% at a 1% false rejection rate.

Keywords: Lip-based · Biometric authentication · One-shot-learning · Siamese network · XM2VTS

1 Introduction

It is widely accepted that single passwords are not a secure means of authentication and this has led to increased attention for biometric authentication. Biometric authentication is the process for verifying the identity of a person based on a unique personal characteristic or trait. Replacing passwords with a biometric has many benefits; it cannot be forgotten, no one can steal your biometrics and it cannot be transferred to another person. Biometric authentication solutions, such as face recognition and fingerprint scanners have already been deployed on many state-of-the-art devices.

Physiological or behavioural data can be used as a biometric, examples of physiological biometrics include face, fingerprint and iris. Behavioural biometrics differ from physiological biometrics as they measure a behaviour or pattern such as voice, signature or gait. When incorporated into a biometric, liveness detection is used to confirm the user is live and present. Naturally, liveness detection is easier to incorporate into a behavioural biometric, making them more desirable.

G. Bebis et al. (Eds.): ISVC 2019, LNCS 11844, pp. 405–417, 2019.
https://doi.org/10.1007/978-3-030-33720-9_31

Behaviours are generally hard to mimic or replicate, however, they are also more difficult to collect, model and authenticate robustly.

Lip-Based Biometric Authentication (LBBA) involves authenticating an individual based on visual information captured from video data of a speaker's lips while they are talking. Lip-movements for authentication on mobile devices have a lot of potential, especially considering the popularity of mobile devices and laptops. It would require no dedicated hardware like with fingerprint and cannot be spoofed using a single image.

Authentication is a 2 stage process involving an enrollment stage, then authentication against the enrollment data. This process is illustrated in Fig. 1. During the enrollment stage it is important to consider usability, a desirable attribute of using facial recognition or fingerprint biometrics is that users do not need to provide large amounts of data to register.

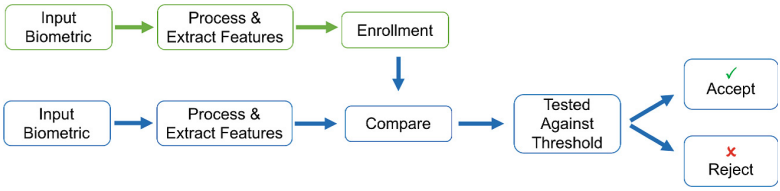

Fig. 1. The 2 stages of biometric authentication. Stage 1, green, is the enrollment phase and stage 2, blue, is the authentication stage. (Color figure online)

To date LBBA has not gained a significant amount of research, especially compared to other biometrics such as face, fingerprint or voice. Section 1.1 reviews the literature in this area, showing that much of the previous research into LBBA is sporadic and inconsistant, with results reported on small or private datasets, and different results metrics making comparison difficult. Section 1.1 also highlights many of the existing approaches require multiple enrollment videos, which could be a potential barrier to adoption.

The aim of this paper is to push the field forward by applying deep learning to improve the performance and usability of LBBA, setting a new state-of-the-art benchmark performance for a large, well known dataset and protocol. A deep Artificial Neural Network (ANN) is trained end-to-end and one-shot-learning applied.

1.1 Existing Publications and Motivation

Previous research into LBBA has stemmed from visual speech recognition and speaker verification. Multiple early publications [3,4,6,9,13,18] on speaker verification included research using visual lip-movements for authentication. These works reported results for the lip-movements as a solo biometric, in addition to their acoustic equivalent and multimodal approaches.

In their speaker verification research [18] explored single and multimodal solutions, using geometric features and dynamic time warping. Results were reported on the XM2VTS dataset. [18] reported an Equal Error Rate (EER) of 14% for lips alone and concluded that information from lips was not competitive with other solo biometrics. However, they also reported that including lip information with other modalities did improve the overall performance in all multimodel combinations investigated.

In [13] they used Hidden Markov Models (HMM) for lip-based speech and speaker recognition. Features were selected from the pixels using Principle Component Analysis (PCA) and Linear Discriminant Analysis (LDA), using the smaller M2VTS dataset. Using only the visual lip-movements, they recorded an EER of 19.7% with the LDA features, which was significantly worse than their acoustic equivalent.

While researching multimodal speaker authentication, [6] also investigated lip-based motion features, created using velocity vectors from 6 regions around the mouth and modelled using Gaussian Mixture Models (GMM). The XM2VTS dataset and Lausanne protocol were used. Results reported a 22% EER on their lip features, and 6% EER on acoustic features. They also reported an improved multimodal result of 2% EER when combining the audio and lip-based features.

Work in [17] used a form of geometric features from visual information for authentication, based on time-series information of points tracked around the lips and the geometric shape of the lips. Results were reported on a 43 person private dataset, achieving a 14.5% False Acceptance Rate (FAR) at a 3% False Rejection Rate (FRR).

All these approaches to LBBA have had limited success, showing that although including visual lip-based information improves performance, lips alone appear weak as a biometric. Results are reported on varying datasets of a wide range of sizes and lack of defined protocols makes comparison difficult.

In [23] they investigated visual lip information as a solo biometric. They used Discrete Cosine Transform (DCT) coefficients and their first and second order derivatives to create features. The features were modelled with GMMs and results reported on the XM2VTS dataset with the Lausanne Protocol. They reported an EER during evaluation of 2.2% and a FAR of 1.7% at a FRR of 3% on the test set. This work showed for the first time that visual information taken from moving lips can achieve comparable accuracy to other single biometric modalities, on a large dataset and known protocol.

LBBA research in [21] divided video data into time steps and categorised it based on the lips at rest, speaking or in transition. The feature representation consisted of 4 different lip properties: (i) shape descriptors calculated from the contour points, (ii) lip texture, (iii) motion vectors of the contour points and (iv) the Local Ordinal Contrast Pattern (LOCP). The features were modelled using HMMs and Support Vector Machines (SVM), and a Universal Background Model (UBM) was used during classification. A private 40 person dataset created for this work was used and a closed-set protocol defined. The small size of the dataset meant during testing there were only 9 registered clients to produce

returning client scores. Results were reported separately for the 3 stages (at rest, speaking, transition) and combined, the best reported result was on all stages combined and produced a Half Total Error Rate (HTER) of 1.26%.

Lip information was used when researching identification in [12], where identification differs from authentication as it aims to answer the question 'who out of this dataset is it?'. For this work 20 s RGB video segments containing only the mouth area were used to train an ANN with convolutional and LSTM layers. A 57 person dataset of individuals uttering the digits 0–9 in Chinese was divided into training and testing. The setup mimicked a closed-set protocol and the best reported result was 96.01% accuracy.

[20] proposed fusing face and dynamic visual information from speaker's lips for authentication. For this work they hand engineered features from spatial and temporal information over the frames. The face and lip features were fused using a single layer ANN. For this work the OuluVS corpus was used, which is made up of 20 individuals uttering 10 short phrases 10 times. Results on closed-set tests show, as a single modality, face (83.75%) outperforms lips (71%) but together an improved 93.25% accuracy can be achieved.

From the published literature it can be seen that LBBA using visual information is not a well researched area, which is surprising given the interest in its acoustic equivalent - speaker verification, or its potential for liveness unlike face or fingerprint. Results are still sporadic and frequently reported on evaluation sets, rather than evaluation and test sets. Datasets are often private [11,12,16,17,20,21], making it almost impossible to compare algorithms. Training and reporting results on such small datasets [12,17,20,21] leads to reservations about how scalable the system is, and raises concerns about the possibility of over fitting.

In addition to using a larger, more diverse dataset and defined protocol for comparison of results, there are a number of other considerations for a LBBA solution. Work in [23] currently have the state-of-the-art results on the XM2VTS dataset and Lausanne protocol for LBBA with a 2.2% EER during evaluation and FAR of 1.7% at a FRR of 3% on the test set. In [23] DCT coefficient features were modelled with GMMs. The setup required using 4 videos, each containing 20 digits to create individual GMMs for every individual during the enrollment stage. More recent publications in face recognition [19] and [22], implemented a one-shot-learning solution, where only a single image is required for enrollment during authentication. Both [19] and [22] train deep convolutional neural networks using a siamese network and achieved state-of-the-art results. Results in [19] outperformed human level performance for face recognition.

Drawing from the success of these works, this paper explores the possibility of employing one-shot-learning for LBBA using a siamese network. A one-shot-learning solution would mean authenticating a claimed identity using only one enrollment video. The model trained for this work is referred to as LipAuth. To the best of the authors knowledge, this is the first time one-shot-learning and end-to-end training has been applied to LBBA. In addition this is the first time recurrent convolutional neural networks have been implemented within a

siamese network architecture to handle variable length RGB video data for lip-based biometric authentication.

2 Methods

2.1 Siamese Network Overview

A siamese network is used to learn the similarity function between inputs, the architecture consists of two branches, each containing an identical model. See Fig. 2. The similarity function between two inputs is learned by passing each input to one of the branches to obtain a feature embedding for each input. The distance between the feature embeddings is calculated and model weights updated in both arms of the network identically, to minimize the distance between similar inputs and maximise distance between different inputs. The duplicated model weights ensure that if identical inputs are passed to the network they will be mapped to the same feature embedding.

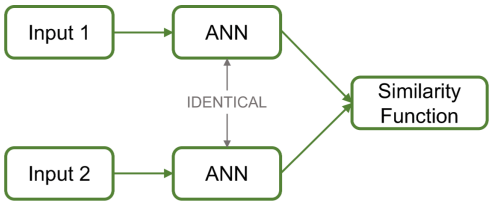

Fig. 2. Siamese Network overview

2.2 Artificial Neural Network Layers

STCNN Layers. Spatio-Temporal convolution is a modification of 2D convolution typically used in CNNs. To process video data STCNNs include an additional summation over time. For an input video $\mathbf{x} \in \mathbb{R}^{C \times T \times W \times H}$ and a STCNN layer with C' kernels of size $k_t \times k_w \times k_h$, the output volume is computed as:

$$[stconv\,(\mathbf{x}, \mathbf{w})]_{c'tij} = \sum_{c=1}^{C} \sum_{t'=1}^{T} \sum_{i'=1}^{W} \sum_{j'=1}^{H} w_{c'ct'i'j'} x_{c,t+t',i+i',j+j'} \tag{1}$$

where x_{ctij} is the pixel at location i, j in the c^{th} channel of the video frame at timestep t and $w_{c'ct'i'j'}$ indexes the STCNN layer weights. The equation above ignores the bias and assumes both a stride of 1 and zero padding of frames when $i + i'$ or $j + j'$ are greater than W or H respectively.

Bi-directional Gated Recurrent Units Layer. The RNN portion of the LipNet architecture uses GRU layers [5] formulated as:

$$\Gamma_r \quad = \quad \sigma\left(\mathbf{W}_r \left[\mathbf{c}^{\langle t-1 \rangle}, \quad \mathbf{x}^{\langle t \rangle}\right] + \mathbf{b}_r\right) \tag{2}$$

$$\tilde{\mathbf{c}}^{\langle t \rangle} = \tanh\left(\mathbf{W}_c \left[\Gamma_r \bullet \mathbf{c}^{\langle t-1 \rangle}, \quad \mathbf{x}^t\right] + \mathbf{b}_c\right) \tag{3}$$

$$\Gamma_u = \sigma\left(\mathbf{W}_u \left[\mathbf{c}^{\langle t-1 \rangle}, \quad \mathbf{x}^{\langle t \rangle}\right] + \mathbf{b}_u\right) \tag{4}$$

$$\mathbf{c}^{\langle t \rangle} = \Gamma_u \bullet \tilde{\mathbf{c}}^{\langle t \rangle} + (1 - \Gamma_u) \bullet \mathbf{c}^{\langle t-1 \rangle} \tag{5}$$

where $\mathbf{x}^{\langle t \rangle}$ is the output of the STCNN, $\mathbf{c}^{\langle t-1 \rangle}$ is the previous timestep's activations and $\sigma(\mathbf{z}) = \frac{1}{(1+e^{(-z)})}$. $[\mathbf{W}_r, \mathbf{b}_r]$ and $[\mathbf{W}_u, \mathbf{b}_u]$ denote the parameters of the reset and update gates. A bi-GRU [7] is used to take advantage of information contained in all video frames, not just previous frames.

2.3 Loss Function

There are 2 popular loss functions that are suitable for this task, Binary Cross Entropy Loss (BCE) or Triplet Loss.

Binary Cross Entropy Loss Function. The BCE loss function is used when comparing 2 inputs at a time. When inputs are from the same class the true label will be $y = 1$, else $y = 0$. The loss function $J(y, \hat{y})$ is calculated using:

$$J\left(y, \hat{y}\right) = -\frac{1}{m} \sum_{i=1}^{m} \left(y_i \log\left(\hat{y}_i\right) + (1 - y_i)\, log\left(1 - \hat{y}_i\right)\right) \tag{6}$$

Where the loss is an average of the errors between the predicted labels, $\hat{y}$, and the actual labels y. The predicted labels are represented by the sigmoid of the distance: $\sigma\left(d\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right)\right)$.

Triplet Loss Function. The triplet loss function differs from the BCE loss function in that the network is passed 3 examples at a time. During training the network described in Fig. 2 has an additional third arm, with another identical copy of the model. Instead of passing the network examples in pairs, the network is trained by passing it triplets where a triplet contains a positive pair and an additional negative example. One of the triplets is the **anchor**, A, and every triplet contains a **positive**, P, and **negative**, N, example. The triplet loss function is defined as:

$$J\left(y, \hat{y}\right) = -\frac{1}{m} \sum_{i=1}^{m} \max\left(d(\mathbf{A}, \mathbf{P}) - d(\mathbf{A}, \mathbf{N}) + \alpha, \quad 0\right) \tag{7}$$

where a margin α, is the minimum distance between the positive and negative pair that the network tries to satisfy. If the triplet used as a training example is 'too easy', it will have no contribution to the loss and not have any impact on the weight updates, hence the model cannot learn from it. There are 3 groupings of triplets to select from when training:

1. *Easy Triplets*: refers to those which do not contribute to the loss because the positive is very similar to the anchor and the negative is extremely unlike the anchor, therefore: $d(\mathbf{A}, \mathbf{P}) + \alpha < d(\mathbf{A}, \mathbf{N})$ so the loss will be 0.
2. *Semi-Hard Triplets*: are those for which the positive is still closer to the anchor than the negative, but produce a positive loss because the negative falls within the margin: $d(\mathbf{A}, \mathbf{P}) < d(\mathbf{A}, \mathbf{N}) < d(\mathbf{A}, \mathbf{P}) + \alpha$.
3. *Hard Triplets*: refers to triplets where the negative example is more similar to the anchor than the positive example: $d(\mathbf{A}, \mathbf{P}) > d(\mathbf{A}, \mathbf{N})$

2.4 Selecting Training Data

The selection of pairs or triplets used during training has been investigated for both triplet loss and BCE loss. [10] trained a state-of-the-art model for character recognition using randomly sampled triplets, whereas for face recognition [19] trained using only semi-hard triplets. In [19] it was found that when triplets were randomly sampled only a few contributed to the loss and therefore the model took longer to converge, whereas using only the hard triplets caused the model to fail to converge. Work in [1] on person re-identification trained their model using BCE and found a 2:1 ratio of the negative to positive samples, followed by fine tuning the final layers using only hard pairs was optimal.

It is also worth noting when training a siamese network it can be computationally expensive and slow compared to other architectures, because the pairs/triplets need selected each time after updating the network weights as the embeddings of the inputs will have changed with each update.

3 Dataset and Protocol

The XM2VTS dataset [15] was used for the work in this paper. The XM2VTS is a commonly used because of its size, availability and diversity. The dataset contains video data of 295 individuals recorded over 4 months in 4 separate sessions. The time frame between sessions makes the dataset particularly desirable for authentication tasks as it captures general changes in appearance over time. During each session full face audio-visual video data was collected of each individual speaking a 20 digit sequence; '0123456789 5069281374'. Each speaker uttered the sequence 2 times during each session, producing 8 videos per person and 2,360 available videos. Additional video data was captured as part of the XM2VTS but only the digit sequence video data was used in this work. All XM2VTS videos were captured in a well lit up room, with minimal background noise, a blue background and recorded and 25 frames per second.

The XM2VTS is accompanied by the Lausanne Protocol [14]. Configuration II of the protocol was used for this work. The Lausanne protocol is a closed-set protocol because all users are enrolled during training and the protocol does not take new users into consideration during evaluation and testing.

For this work the XM2VTS was cropped to only contain the lips and surrounding mouth area using open-source library, DLib [8].

Training data contains 4 videos per individual, this provided 6 combinations of anchor-positive per person, with $200 \times 6 = 1,200$ postive examples in training. As the pool for choosing negative examples was significantly greater, it was decided to match the number of anchor-positives and use 1,200 anchor-negatives.

During evaluation the anchor videos were selected from the training data, and the positive examples were selected from the evaluation data. With 4 anchors per person and 2 evaluation videos, there were 8 positive videos per individual. This produced $200 \times 8 = 1,600$ positive examples during evaluation. As it is a closed-set evaluation no new anchor videos were added. Imposter tests during evaluation involved testing all 800 anchor videos against all available evaluation videos, where imposters-only contribute $(8 \times 25 \times 800\text{anchors}) = 160,000$ imposter tests, and returning clients contribute $(2 \times 200 \times 800\text{anchors}) - 1,600 = 318,400$ imposter tests. This results in **478,400** imposter tests and **1,600** client tests during evaluation.

During testing there are 2 videos per individual. As in evaluation, these 400 videos will produce the same number of anchor-positive examples (1,600) and contribute 318,400 imposter tests. The new imposters-only will contribute an additional $(8 \times 70 \times 800\text{anchors}) = 448,000$ imposter tests. This results in **761,400** imposter tests and **1,600** client tests during the test stage.

4 Experiments and Discussion

For each branch of the siamese network the model architecture selected was inspired by work in LipNet [2]. LipNet is a deep ANN designed for visual speech recognition. The LipNet model contains 3× STCNN layers each directly followed with a max pooling layer, and 2× Bi-GRU layers each with 128 neurons. LipNet's architecture has successfully shown it can handle video data containing only the lips and mouth area, however it was optimised for speaker independant visual lip reading. The aim of the model and optimised weights developed for this work, LipAuth, is to model the uniqueness within the lip movements. Preliminary experiments using only a subset of the available training data were carried out to fine tune the model hyperparameters. This includes dropout, the loss function and initialization. Following the preliminary experiments, LipAuth was trained using all available XM2VTS training data as in the Lausanne protocol.

4.1 Preliminary Experiments

Access was provided to a Nvidia-GPU graphics card, GRID M60-8Q, with 8 GB RAM. The GPU did not provide enough memory for batch training so the model was trained by showing it a single triplet at a time and updating the weights after each triplet. An epoch was considered finished after a full pass of all triplets.

A subsample of the available training and evaluation data was selected and used for preliminary experiments to fine tune the model hyperparameters. If the model struggled to fit the subset of training an evaluation data then it suggests it would not have been suitable for a larger dataset.

For the preliminary experiments the first 30 individuals from the training set were used as a training set, along with the first 10 individuals in the evaluation set. The subset of the data produced 180 anchor-positive pairs for training. The aim of the preliminary experiments was:

- Confirm how LipAuth should be initialized. The LipNet weights might not be useful given LipNet was optimised for a different task.
- Select between the BCE and triplet loss functions.
- Decide if dropout should be removed from the architecture as the aim of LipAuth is to model the uniqueness within individual's lip movements.

Results from the preliminary experiments are shown in Fig. 3. Training data for both loss functions was selected beginning with all 180 anchor-positive pairs, then selecting a negative for each anchor. With BCE the pairs are shown to the model one at a time, and 1 triplet at a time with the triplet loss. Semi-Hard and Easy (SH+E) negative examples were chosen.

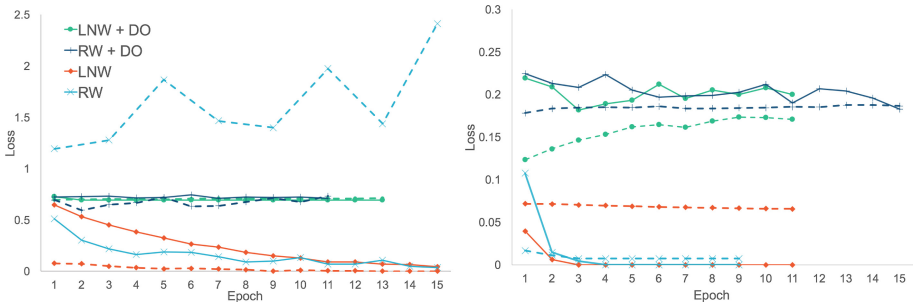

Fig. 3. Loss for the models trained for the preliminary experiments. On the left the figure shows the results from the BCE loss function, and on the right is the triplet loss function. Solid lines show the loss on the training data and dashed lines shows the loss on the evaluation set. LNW = Initialized with LipNet weights, RW = Initialized with Randomised Weights, DO = dropout.

Figure 3 shows the training loss (solid lines) and evaluation loss (dashed lines) for the preliminary experiments. The figure on the left shows models trained with the BCE loss function, and triplet loss function on the right. It can be clearly seen for both loss functions, including dropout prevents the model from making useful updates as the loss remains fairly constant. The model initialized with LipNet weights converged and did not appear to overfit, however, the model initialized with randomised weights overfit the training data. With the triplet loss function the model initialized with LipNet weights overfit the training data and as the training loss dropped to zero it remained overfit. The model trained with the triplet loss, with zero dropout and randomly initialized weights performed the best in the preliminary experiments, the model converged quickly and it did not overfit, these hyperparameters were used to train LipAuth with all available training data.

4.2 Results and Discussions

LipAuth was trained using SH+E triplets with a learning rate of 1×10^{-5}, the triplet loss function, zero dropout, and LipNet architecture. There were 1,600 available anchor-positive pairs for training, all anchor positives were used to generate triplets.

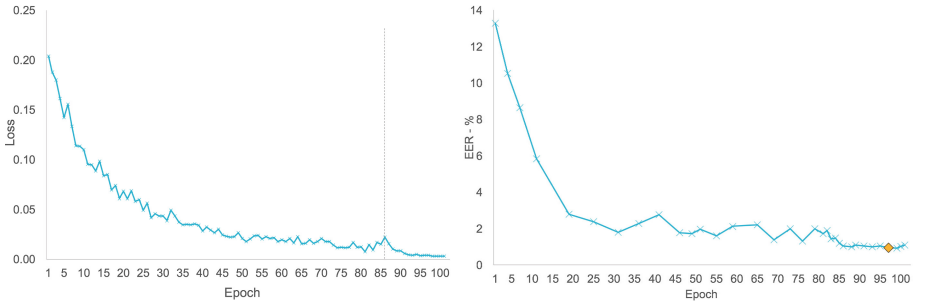

Fig. 4. Training LipAuth. Left figure shows the training loss. The dashed line marks where the learning rate was dropped from 1×10^{-5} to 1×10^{-6}, at epoch 86. The figure on the right shows the EER on the evaluation set. The yellow diamond marks the lowest EER achieved on the evaluation set, 0.93% at epoch 99. (Color figure online)

Figure 4 shows a plot of the training loss on the left, and on the right the evaluation EER set during training. From the figure, the model converged to approximately 2% by epoch 30, after this the model was trained for a further 55 epochs and the evaluation EER did not drop below 1.18%. At epoch 85 the learning rate was reduced to 1×10^{-6} as the training loss appeared to have plateaued. The model was trained for a further 20 epochs at this lower learning rate and achieved a minimum EER of 0.93% at epoch 99. The weights from epoch 99 were used for the LipAuth model to calculate results on the unseen test set.

The LipAuth model produced an EER on the unseen test set of 1.03% and a FAR of 1.07% at a 1% FRR. If the threshold was set to only accept 1% of imposters (FAR=1%), then the FRR=1.09%. This 1.09% FRR equates to 15 returning individual tests that did not successfully log in, 14 of which were from 2 individuals. Each individual had 4 anchor videos and 2 authentication tests, producing 8 attempted logins per person. The 2 problematic individuals: 366 and 369, could not login 7 times each. From reviewing the videos, one individual was male, the other female and there was no facial hair or striking differences between sessions. However, there did appear to be some jitter from the tracking and cropping of these videos caused by the users moving their head while speaking.

Figure 5 shows the FRR against the FAR as the threshold is varied. If the requested authentication application required an extremely secure setup where no imposters could login (0% FAR), then the FRR would be 7.41%.

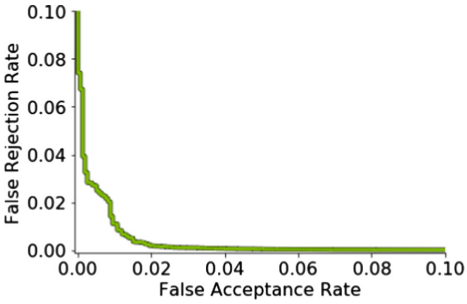

Fig. 5. Results of the LipAuth model on the unseen test set. FRR against the FAR as the threshold for logging in varies.

These results set a new state-of-the-art for lip-based authentication on the XM2VTS dataset. The previous state-of-the-art [23] achieved 2.2% EER on the evaluation set and a FAR = 1.7% at a FRR = 3.0% using the same dataset and protocol. If the LipAuth threshold is set to a FRR = 3% it achieves 0.25% FAR. The LipAuth results also compare very favourably to work in [12], who used an ANN with LSTM layers for lip-based identification and achieved an accuracy of 96.01%. A single layer ANN was trained in [20], where they reported 71% accuracy, and 93.25% accuracy on lips and face combined. In addition, both these works used significantly smaller datasets of 57 and 20 individuals respectively. See Table 1 for comparison of results.

Table 1. The table shows the results of previous authentication/identification publications using visual lip-based information only and dataset used.

Paper	Dataset	Task	Metric	Result
[18]	XM2VTS, 295	Authentication	EER	14%
[13]	M2VTS, 36	Authentication	EER	19.7%
[6]	private, 43	Authentication	FAR@ 3% FRR	14.5%
[17]	private, 40	Authentication	HTER	1.26%
[23]	XM2VTS, 295	Authentication	FAR@ 3% FRR	1.7%
[12]	private, 57	Identification	Accuracy	96.01%
[20]	OuluVS, 20	Identification	Accuracy	71%
LipAuth	XM2VTS, 295	Authentication	**FAR@ 3% FRR**	**0.25%**
LipAuth	XM2VTS, 295	Authentication	**EER**	**1.03%**

5 Conclusions and Future Work

The work in this paper showed for the first time that ANNs can be successfully applied to LBBA, in particular recurrent convolutional neural networks with

STCNN and Bi-GRU layers. This work explored the application of a siamese network architecture, trained end-to-end for one-shot LBBA at test time. The network trained, LipAuth, and handled variable length RGB video data. Preliminary experiments showed the effects of dropout, pre-trained weights and the training loss function. It was discovered that selecting training data in triplets, where a triplet involves an anchor-positive-negative, and trained with the triplet loss function outperformed training on pairs with the binary cross-entropy loss function. The final LipAuth network was randomly initialized and contained zero dropout.

LipAuth was trained using XM2VTS data. On the Lausanne protocol evaluation and test set LipAuth achieved state-of-the-art performance with an EER = 0.93%, and FAR of 1.07% at a 1% FRR, outperforming the GMM-UBM setup which was previously state-of-the-art for lip-based authentication using the same dataset and protocol. In addition, LipAuth was designed to implement a one-shot-learning approach meaning only one previous example is required for enrollment, whereas the previous state-of-art approach was dependent on multiple videos for each individual.

Although these results show promise, future work is needed to better understand how the performance would be affected with real-world challenges. Future work includes testing with an open-set protocol, exploring real-world datasets with a variety of lighting conditions and diverse content.

References

1. Ahmed, E., Jones, M., Marks, T.K.: An improved deep learning architecture for person re-identification. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3908–3916, June 2015. https://doi.org/10.1109/CVPR.2015.7299016
2. Assael, Y.M., Shillingford, B., Whiteson, S., de Freitas, N.: Lipnet: sentence-level lipreading. CoRR abs/1611.01599 (2016). http://arxiv.org/abs/1611.01599
3. Brand, J.: Visual speech for speaker recognition and robust face detection. Ph.D. thesis, University of Wales, Swansea, UK (2001)
4. Cetingul, H.E., Yemez, Y., Erzin, E., Tekalp, A.M.: Discriminative analysis of lip motion features for speaker identification and speech-reading. Trans. Img. Proc. **15**(10), 2879–2891 (2006). https://doi.org/10.1109/TIP.2006.877528
5. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. CoRR abs/1412.3555 (2014)
6. Faraj, M., Bigun, J.: Motion features from lip movement for person authentication. In: 2006 18th International Conference on Pattern Recognition, ICPR 2006, vol. 3, pp. 1059–1062 (2006). https://doi.org/10.1109/ICPR.2006.814
7. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional LSTM networks. In: Proceedings 2005 IEEE International Joint Conference on Neural Networks, vol. 4, pp. 2047–2052, July 2005. https://doi.org/10.1109/IJCNN.2005.1556215
8. Kazemi, V., Sullivan, J.: One millisecond face alignment with an ensemble of regression trees. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1867–1874, June 2014

9. Kittler, J., Li, Y.P., Matas, J., Sánchez, M.U.R.: Lip-shape dependent face verification. In: Bigün, J., Chollet, G., Borgefors, G. (eds.) AVBPA 1997. LNCS, vol. 1206, pp. 61–68. Springer, Heidelberg (1997). https://doi.org/10.1007/BFb0015980
10. Koch, G., Zemel, R., Salakhutdinov, R.: Siamese neural networks for one-shot image recognition. In: ICML deep learning workshop, vol. 2 (2015)
11. Lu, L., et al.: Lip reading-based user authentication through acoustic sensing on smartphones. IEEE/ACM Trans. Networking **27**(1), 447–460 (2019). https://doi.org/10.1109/TNET.2019.2891733
12. Lu, Z., Wu, X., He, R.: Person identification from lip texture analysis. In: 2016 IEEE International Conference on Digital Signal Processing (DSP), pp. 472–476, October 2016. https://doi.org/10.1109/ICDSP.2016.7868602
13. Lucey, S.: An evaluation of visual speech features for the tasks of speech and speaker recognition. In: Kittler, J., Nixon, M.S. (eds.) AVBPA 2003. LNCS, vol. 2688, pp. 260–267. Springer, Heidelberg (2003). https://doi.org/10.1007/3-540-44887-X_31
14. Luettin, J., Maître, G.: Evaluation protocol for the extended M2VTS database (XM2VTSDB). Idiap-Com Idiap-Com-05-1998, IDIAP (1998)
15. Messer, K., Matas, J., Kittler, J., Jonsson, K.: Xm2vtsdb: the extended m2vts database. In: In Second International Conference on Audio and Video-based Biometric Person Authentication, pp. 72–77 (1999)
16. Morikawa, S., Ito, S., Ito, M., Fukumi, M.: Personal authentication by lips EMG using dry electrode and CNN. In: 2018 IEEE International Conference on Internet of Things and Intelligence System (IOTAIS), pp. 180–183, November 2018. https://doi.org/10.1109/IOTAIS.2018.8600859
17. Nakata, T., Kashima, M., Sato, K., Watanabe, M.: Lip-sync personal authentication system using movement feature of lip. In: 2013 International Conference on Biometrics and Kansei Engineering (ICBAKE), pp. 273–276, July 2013. https://doi.org/10.1109/ICBAKE.2013.53
18. Sanchez, M.U.R.: Aspects of facial biometrics for verification of personal identity. Ph.D. thesis, University of Surrey, Guilford, UK (2000)
19. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: a unified embedding for face recognition and clustering. CoRR abs/1503.03832 (2015). http://arxiv.org/abs/1503.03832
20. Shang, D., Zhang, X., Xu, X.: Face and lip-reading authentication system based on android smart phones. In: 2018 Chinese Automation Congress (CAC), pp. 4178–4182, November 2018. https://doi.org/10.1109/CAC.2018.8623298
21. Shi, X., Wang, S., Lai, J.: Visual speaker authentication by ensemble learning over static and dynamic lip details. In: 2016 IEEE International Conference on Image Processing (ICIP), pp. 3942–3946, September 2016. https://doi.org/10.1109/ICIP.2016.7533099
22. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: closing the gap to human-level performance in face verification. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2014
23. Wright, C., Stewart, D., Miller, P., Campbell-West, F.: Investigation into DCT feature selection for visual lip-based biometric authentication. In: Dahyot, R., Lacey, G., Dawson-Howe, K., Pitié, F., Moloney, D. (eds.) Irish Machine Vision & Image Processing Conference Proceedings 2015, pp. 11–18. Irish Pattern Recognition & Classification Society, Dublin, Ireland (2015), winner of Best Student Paper Award