Wright and Stewart *EURASIP Journal on Information Security* (2020) 2020:3
https://doi.org/10.1186/s13635-020-0102-6

EURASIP Journal on
Information Security

RESEARCH Open Access

Understanding visual lip-based biometric authentication for mobile devices

Carrie Wright* and Darryl William Stewart

Abstract

This paper explores the suitability of lip-based authentication as a behavioural biometric for mobile devices. Lip-based biometric authentication is the process of verifying an individual based on visual information taken from the lips while speaking. It is particularly suited to mobile devices because it contains unique information; its potential for liveness over existing popular biometrics such as face and fingerprint and lip movements can be captured using a device's front-facing camera, requiring no dedicated hardware. Despite its potential, research and progress into lip-based biometric authentication has been significantly slower than other biometrics such as face, fingerprints, or iris.

This paper investigates a state-of-the-art approach using a deep Siamese network, trained with the triplet loss for one-shot lip-based biometric authentication with real-world challenges. The proposed system, LipAuth, is rigourously examined with real-world data and challenges that could be expected on lip-based solution deployed on a mobile device. The work in this paper shows for the first time how a lip-based authentication system performs beyond a closed-set protocol, benchmarking a new open-set protocol with an equal error rates of 1.65% on the XM2VTS dataset. New datasets, qFace and FAVLIPS, were collected for the work in this paper, which push the field forward by enabling systematic testing of the content and quantities of data needed for lip-based biometric authentication and highlight problematic areas for future work. The FAVLIPS dataset was designed to mimic some of the hardest challenges that could be expected in a deployment scenario and include varied spoken content, miming and a wide range of challenging lighting conditions. The datasets captured for this work are available to other university research groups on request.

Keywords: Lip-based, Biometric, Authentication, One-shot-learning, Open-set, Real-world, FAVLIPS, qFace, XM2VTS, Siamese network, Triplet loss

Introduction

In today's society, mobile devices such as phones, ts and laptops are considered essential for both personal and business purposes and the risks of passwords as a sole means of authentication is widely recognised. These devices can provide a gateway to gaining access to private and confidential data and online services such as social media, financial services and ecommerce services. Secure authentication before gaining access to personal devices is essential.

Biometric authentication is the process of verifying the claimed identification of a person based on an innate human characteristic or trait. Figure 1 gives an overview of the 2 stages of biometric authentication which first includes an enrolment stage, users can then authenticate themselves against the enrolment data.

Biometric traits can be physiological or behavioural. Physiological biometrics such as face or fingerprint have already been successfully rolled out in many state-of-the-art devices, both of these examples have been spoofed in high profile media cases. Behavioural biometrics capture a pattern or behaviour such as signature or voice verification. Behavioural biometrics can be more difficult to spoof; however, they can also be more difficult to model and authenticate robustly. Within biometric authentication, liveness detection refers to being able to detect if a human is live and present during the authentication process. If liveness is successfully incorporated within a biometric system it could prevent face recognition systems from being spoofed using photographs or artificial

*Correspondence: cwright32@qub.ac.uk
ECIT, Queen's University Belfast, Northern Ireland Science Park, Queen's Road, Queen's Island, Northern Ireland

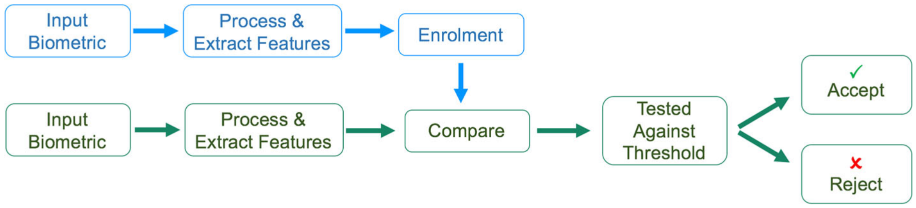

Fig. 1 Biometric authentication overview, involving an enrolment stage shown in blue, and authentication stage shown in green

fingerprints being successful. Liveness detection is naturally easier to build into a behavioural biometric system as the behaviour requested can be altered.

Lip-Based Biometric Authentication (LBBA) is the process of authenticating a person based on their visual lip movements while speaking. LBBA has great potential for mobile devices; it is a behavioural biometric in which liveness could be easily incorporated by randomising the requested spoken content, and it can be captured using a device's front-facing camera.

Despite its suitability, LBBA research has been sporadic. Much of the LBBA research has stemmed from speech and speaker verification, where early studies [1–4] suggested lip-movements alone were not sui as a solo biometric.

Due to the recent limited success in selecting suitable lip-based features for speaker verification, work in [4] explored behavioural lip motion and intensity features for visual speaker verification and speech recognition. [4] used a closed-set protocol and 50-person dataset for all training and evaluation. HMMs were used to model the behavioural features and the best recorded result using visual information only was an equal error rate (EER) of 1.6% on the evaluation set. Following this, additional research [5–8] went on to further confirm lip-movements do contain unique information. However, in all these works training and testing were completed using single, small private datasets containing 9, 20, 43, and 40 individuals, respectively, and closed-set protocols. In a closed-set protocol, all users are known in advance and enrolled during training. An open-set protocol differs as it enrols new users during evaluation and testing stages, producing more realistic results of how a system would actually perform in deployment.

Lip information has also been used in identification. Similarly, identification is also a 2-step approach, as in Fig. 1. However, identification differs from authentication in that it is the ability to identify an individual from a predefined group of users. Work in [9] generated lip features using information about the lip area, height and width of the lip contours, oral cavity pixels and visible teeth. For the work in [9], they collected a private 20 person data set for all training and testing. Their best recorded result for identification was reported as 94.7% accuracy.

While evidence clearly shows the potential for visual information taken from the lips to be used for authentication, each paper discussed so far used a different approach to feature selection and extraction. Results have been reported on single datasets that have ranged in size and availability and closed-set protocols. Using larger datasets that are widely available or multiple datasets would not only enable comparison of results but also give a greater understanding to the strengths and weaknesses of the proposed algorithms.

Work in [10] researched LBBA using the popular XM2VTS dataset, containing 295 individuals and closed-set Lausanne protocol. They reported an EER of 2.2% during evaluation and a FAR of 1.7% at a FRR of 3% on the test set. The work in [10] used Discrete Cosine Transform (DCT) coefficient features modelled with GMMs, where enrolment required 4 videos for each user speaking a total of 80 digits from multiple sessions. While not unfeasible, the required amount of enrolment data is not ideal for a biometric authentication solution.

Work in [11] presented a preliminary study of LBBA with deep Siamese networks. They achieved state-of-the-art on the XM2VTS dataset and closed-set Lausanne protocol for LBBA with an EER of 1.03%. The proposed solution, referred to as LipAuth, enabled a one-shot-learning solution which allows new users to enrol with a single 20-digit video.

This paper extends [11] with a more rigourous investigation into the suitability of LipAuth for mobile devices. The work in this paper proposes a new, more realistic open-set protocol for the XM2VTS and testing with 2 new real-world datasets. The datasets, qFace and FAVLIPS were designed and collected for this work. They were captured on a mobile device to enable testing of LipAuth's potential for liveness, to discover how much data is required for enrolment and authentication, to test a series of real-world challenges such as miming, varied content and a range of lighting conditions. Furthermore, this paper shows the LipAuth model can be improved with the addition of real-world data to better handle the challenging lighting conditions. The final experiments in this paper show the results of the FAVLIPS dataset on with the more traditional approach to LBBA as proposed in [10].

For the first time, this paper tests LBBA beyond a closed-set protocol and it's actual potential for mobile devices. To the best of the authors knowledge this is the first time LBBA has been investigated with varied content and real-world challenges. One of the aims of this paper is to establish the challenges that cause problems for LBBA in the real-world. This paper proposes a way to improve LipAuth using real-world data so the one-shot embedding is more robust to varied lighting conditions.

Methods

The LipAuth system was trained using a Siamese network and the triplet loss function. Figure 2 shows an overview of how this was implemented. The LipAuth model in each branch of the Siamese network was inspired by work in LipNet [12]. The LipNet model was designed for visual speech recognition and contains 3× Spatio-Temporal Convolutional Neural Network (STCNN) layers each directly followed with a max pooling layer, and 2× Bi-directional Gated Recurrent Unit (Bi-GRU) layers each with 128 neurons. The LipNet architecture has shown it can learn from video data containing only the lips and mouth area; however, the LipNet weights were optimised for speaker independant visual lip reading.

Siamese network overview

A Siamese network was used to learn the similarity between inputs of the same person. This is done by training a network using 3 inputs at a time, where 2 inputs belong to the same person (an anchor-positive pair) and the third input (negative) is from a different person to the anchor. The network is trained as illustrated in Fig. 2, where a branch is created with an identical model for each input. Each LipAuth model creates an embedding of the input and the triplet loss function minimises the distance between the anchor and positive embeddings and maximises the distance between anchor and negative embeddings. The model weights are then updated in all 3 branches identically. The duplicated model weights ensure that if identical inputs are passed to the network they will be mapped to the same feature embedding.

Artificial neural network layers

STCNN layers

STCNN is a variation on 2D convolution used to process video data [13]. STCNNs differ from CNNs because they include an additional summation over time. Given 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:

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

where x_{ctij} is the pixel at location i, j in the cth channel of the video frame at timestep t, and $w_{c'ct'i'j'}$ indexes the STCNN layer weights. Equation 1 ignores bias, assumes 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 unit layers

GRU layers are a type of RNN [14] used in the LipNet architecture and are formulated as:

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

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

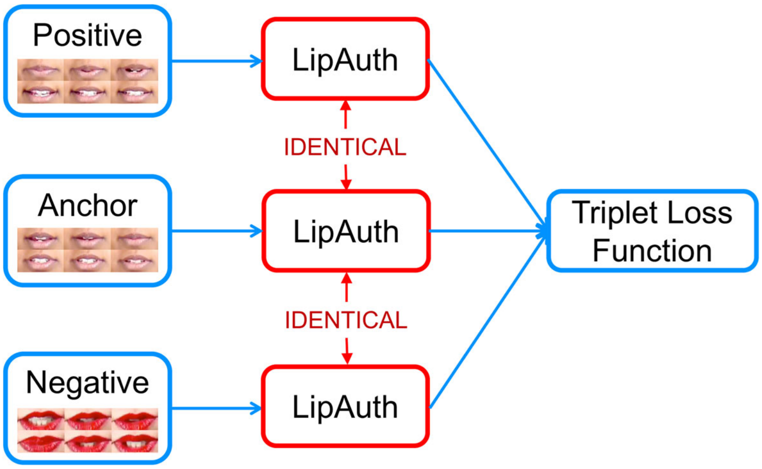

Fig. 2 LipAuth training overview. LipAuth was trained using a Siamese network architecture as shown here. Training data is organised in triplets, where each element of the triplet is passed to a branch containing an identical LipAuth model. The triplet loss updates the LipAuth weights identically in each branch

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

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

where $\mathbf{x}^{\langle t\rangle}$ is the output of the STCNN, the previous timestep's activations are $\mathbf{c}^{\langle t-1\rangle}$ and $\sigma(\mathbf{z}) = \frac{1}{(1+e^{(-z)})}$. The reset and update gate parameters are denoted by $[\mathbf{W}_r, \mathbf{b}_r]$ and $[\mathbf{W}_u, \mathbf{b}_u]$, respectively. A bi-GRU [15] not only takes advantage of previous frames but also can use information contained in all the frames in a video. The 2 bi-GRU layers used in LipAuth produce a many-to-many and many-to-one mapping, respectively. The output the of second bi-GRU layer is a 1D vector embedding representing an entire videos position in 256D space.

Triplet loss function

As seen in Fig. 2, the triplet loss function requires training data organised into triplets, where 1 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:

$$J(y, \hat{y}) = -\frac{1}{m}\sum_{i=1}^{m} \max\left(d(\mathbf{A}, \mathbf{P}) - d(\mathbf{A}, \mathbf{N}) + \alpha, \quad 0\right) \tag{6}$$

where α is the margin which sets the minimum euclidean distance between the positive and negative input that the network tries to satisfy. The loss will not be affected by a training triplet if it is *too easy*, resulting in it having no contribution to the weight updates. When choosing the training triplets there are 3 possible categories:

1. *Easy Triplets* contain a positive which is very similar to the anchor and a negative which is very unlike the anchor; therefore: $d(\mathbf{A}, \mathbf{P}) + \alpha < d(\mathbf{A}, \mathbf{N})$ so the loss will be 0 and the model weights will not be updated.
2. *Semi-Hard Triplets* contain a positive which is closer to the anchor than the negative but will create a positive loss because the negative remains within the margin: $d(\mathbf{A}, \mathbf{P}) < d(\mathbf{A}, \mathbf{N}) < d(\mathbf{A}, \mathbf{P}) + \alpha$.
3. *Hard Triplets* contain a negative example which is more similar to the anchor than the positive example: $d(\mathbf{A}, \mathbf{P}) > d(\mathbf{A}, \mathbf{N})$.

The choice of triplets used for training can have an important effect on training. Work in [16] used randomly sampled triplets when training a model for character recognition and achieved state-of-the-art results. Within face recognition research, [17] found semi-hard triplets optimal. [17] reported when triplets were randomly sampled only a few contributed to the loss and the model took longer to converge, and using only hard triplets caused the model to fail to converge. Work in [18] trained a Siamese network on person re-identification and found a 2:1 ratio of semi-hard triplets, followed by fine tuning the final layers using only hard pairs was optimal.

Training LipAuth

For this work LipAuth was trained twice, once using the closed-set protocol and again using the open-set protocol. LipAuth was trained using semi-hard and easy triplets with a learning rate of 1×10^{-5} and no dropout. These hyperparameters were fine tuned in preliminary work in [19]. The closed- and open-set models were trained for 99 and 113 epochs, respectively, beyond which the evaluation performance plateaued.

Datasets

This work used the existing XM2VTS dataset [20] as it is large and popular and has an accompanying closed-set protocol [21] which enables comparison with other algorithms. In order to test the proposed algorithm for LBBA under more realistic conditions, a new open-set protocol is defined for XM2VTS in this work. Additional data was captured to enable more thorough testing of LBBA for mobile devices. Data was collected in 2 parts: qFace and FAVLIPS, and are available to other university groups for research upon request.

XM2VTS

The XM2VTS dataset [20] is well-known and widely used for audio-visual speaker and speech recognition. The dataset contains 2360 videos from 295 individuals uttering the same digit sequence 8 times, recorded over 4 sessions. Sessions contain each individual repeating a digit sequence twice and a phonetically rich sentence, with approximately 1 month between sessions. The month duration between each session enabled variation in individuals appearance, such as changes in facial hair, lipstick and facial blemishes. With authentication being the focus of this work, it is important that natural changes in appearance such as these are represented in the dataset. In all XM2VTS videos, the speakers are recorded in a well lit up room sitting infront of a blue background and videos are recorded at 25 frames per second (fps). The digit sequence uttered is '01234567895069281374'.

XM2VTS protocols

The Lausanne protocol [21] is a closed-set authentication protocol for XM2VTS. A closed-set authentication protocol requires knowing the whole population of clients in advance [22], meaning no new users are added to the system during cross validation and testing. An open-set protocol differs as it takes new users into consideration during both cross validation and testing. An open-set protocol imitates a real-world scenario, providing a more realistic idea of how a system would actually perform in deployment. It would be unrealistic to retrain a new system with each new user in deployment; thus, it is essential to know how the performance is affected by new users. This makes the open-set protocol a harder challenge. To the best of the authors knowledge, there is no prior open-set protocol currently available for XM2VTS. Figure 3 shows the distributions of individuals for the existing closed-set protocol (Fig. 3a), and the new proposed open-set protocol (Fig. 3b).

qFace

qFace is an audio-visual dataset designed and collected for this work. qFace's design was specifically structured to enable a series of systematic tests for a model trained on XM2VTS. qFace contains real-world data of varied content and duration collected on a mobile device.

qFace was collected using the front-facing camera on a Nexus 7 Android tablet. The dataset contains 10 users saying 10 different digit sequences 8 times each, producing 800 videos. The dataset is made up of 7 males and 3 females and contains a wide range of ages and facial hair. All participants within the dataset are from the UK and Ireland. Users were asked to centre their face in the screen and read out-loud the digit sequences. As can be seen in Fig. 4, minimal instructions on angle and distance to the screen were given, ensuring a more natural and realistic dataset for mobile-based authentication. Full face audio-visual videos were captured at 30 fps and saved as mp4. Recordings for 9 individuals were completed in a single session to avoid significant changes in appearance and behaviour. Due to a technical issue, Jack's recordings were completed in 2 sessions with a single day between sessions. Recording the dataset in a well-lit environment and within a single session limited changes in appearance over time so that changes in results can be attributed to the content and quantity of data instead.

This dataset was designed to verify if an authentication system trained with the XM2VTS dataset can be ported for use in real-world applications using video data captured on a mobile device. Furthermore, to verify if liveness can be easily incorporated by varying digit content. In order to test a range of possible enrolment and authentication content we asked the individuals in qFace to repeat digit sequences of varying length and content 8 times, Table 1 shows the digit sequences spoken.

The first digit sequence in Table 1 is the same digit sequence used in all videos in the XM2VTS. This means a direct comparison can be made using both the XM2VTS open-set protocol test set. Digit sequences 2 to 5 are subsets of digit sequence 1. Sequences 6 to 10 match the length of the previous sequences but the digit sequence was randomised. It is worth noting that unlike sequences 2 to 5, sequences 7 to 10 are not subsets of sequence 6.

qFace is a small dataset, designed and collected to fulfil the specific needs of this work. However, qFace allows for a comparison of real-world data to XM2VTS, and enables a series of systematic experiments to carefully monitor the effects of not only the number of digits spoken, but specific digit content also. The qFace dataset was designed to be used solely as an additional test set.

FAVLIPS

The FAVLIPS dataset contains video data and basic information from 42 individuals captured over 4 months for person authentication. The aim of creating this dataset was to capture real-world data, collected from the same people over a long period of time. This mimics the 4 sessions from the XM2VTS collection process, with the addition of the individuals having control of their own recordings. In this type of authentication, a certain level of cooperation can be expected from individuals, so minimal instructions were provided to users and they were asked to record themselves as they would if they were using it to log into a personal application.

The 42 individuals recorded themselves speaking and miming a series of digit sequences and phonetically rich sentences using a custom application on a Nexus 7 Android tablet over 4 sessions, with approximately 4 weeks between sessions. This time period between sessions allowed for changes in appearance and behaviour, making the authentication challenge more realistic. Each

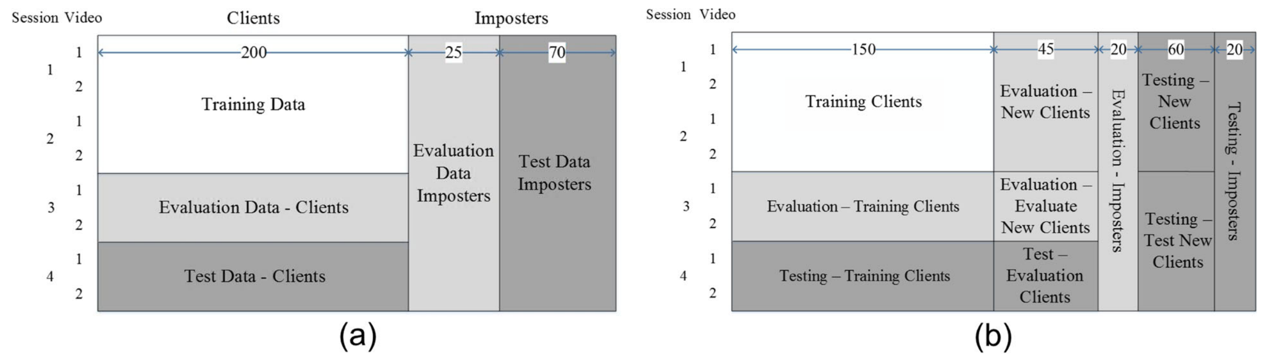

Fig. 3 XM2VTS protocols. **a** Existing closed-set Lausanne protocol, configuration II, as defined in [21]. **b** New open-set protocol as defined for this work

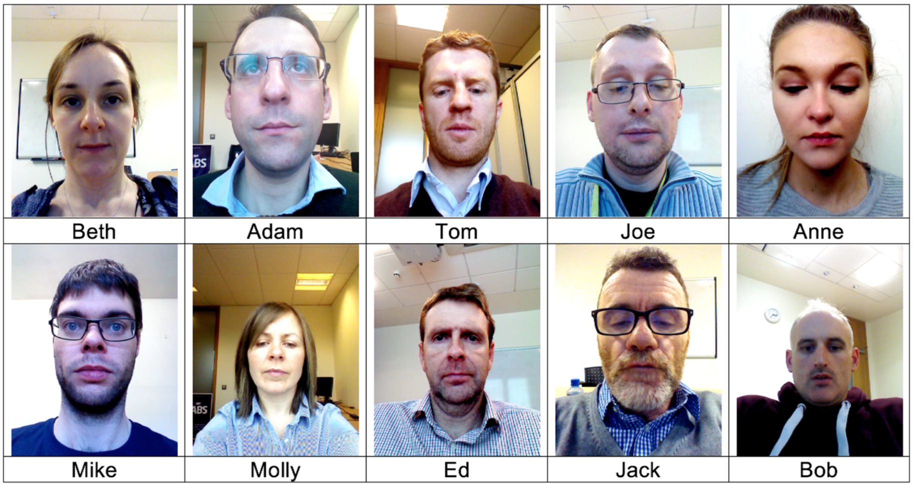

Fig. 4 Sample of the qFace dataset: 10 users consisting of 7 males and 3 females, with each user's reference name. Real-world data collected on handheld mobile device

individual made a total of 54 recordings. During every session individuals were asked to utter the following:

- Ten digits in series
- A randomised 10-digit sequence (same for all users)
- Mime both of the 10-digit sequences
- Utter the randomised 10-digit sequence under 3 additional lighting conditions: light from the side, light behind, and light infront. Figure 5 illustrates the neutral lighting and the 3 additional lighting conditions.
- Randomly selected phonetically rich sentences from 450 possible sentences taken from the TIMIT dataset [23]

The FAVLIPS dataset contains 2268 videos from the 42 individuals. This is made up of 11 females and 31 males, from 6 different countries and ages ranging from 19 to 59 years. Although FAVLIPS is a diverse dataset it is still considered small in comparison to [24–26]; however, it fills a need for testing a LBBA with real-world data. The content captured provides a wide range of real-world testing situations such as miming and lighting that could be expected if deployed on a mobile device. In addition, it enables performance measurements on varied content which is essential for potential liveness checks.

Table 1 The digit sequences uttered 8 times, by each individual in qFace

	Content of digit sequence	No. of digits	Type
1	01234567895069281374	20	XM2VTS
2	5069281374	10	Subset
3	69281374	8	Subset
4	692813	6	Subset
5	2813	4	Subset
6	91763284058540263917	20	Random
7	6873021594	10	Random
8	53216970	8	Random
9	280914	6	Random
10	4795	4	Random

Sequence 1 is the same 20-digit phrase as XM2VTS, sequences 2–5 are subsets of sequence 1. Sequences 6–10 are random. Sequences 7–10 are not subsets of sequence 6

Dataset preprocessing

For this work all datasets were cropped to only contain the lips and surrounding mouth area using an open-source library, DLib [27]. The cropped RGB video data was passed directly to the LipAuth model.

Results

The aim of the work in this paper is to rigorously test LipAuth, a LBBA solution that uses one-shot-learning, with real-world data to better understand its suitability for mobile devices. The experiments include comparison closed-set protocol to the new open-set protocol, benchmarking real-world datasets, varied enrolment and authentication data, and a range of potential challenging situations that would be possible with LBBA on a mobile device. These include miming, randomised digits and sentences for liveness checks and varied illumination. The next experiment presents a proof-of-concept that shows how using a Siamese network can transfer knowledge of challenging conditions to new unseen individuals. The final experiments show the performance of the FAVLIPS dataset on a more traditional approach to LBBA which used handcrafted features, to help provide additional insight on the amount of overtraining seen within the LipAuth model.

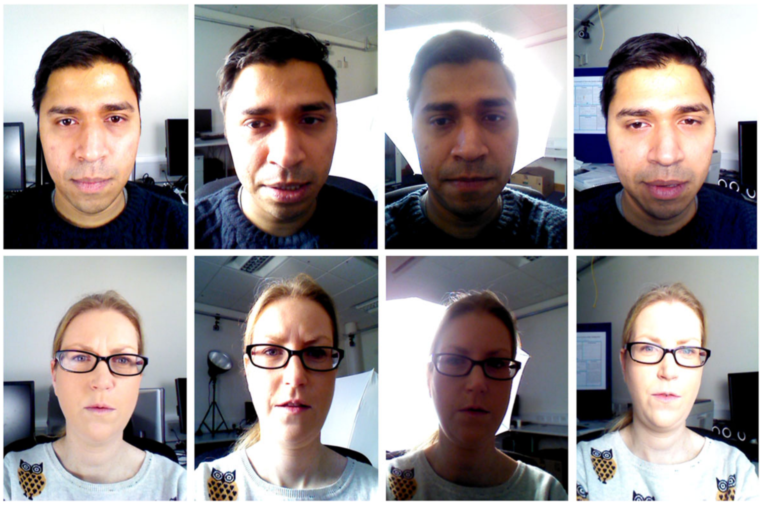

Fig. 5 A sample of 2 participants from FAVLIPS showing the range of lighting conditions captured. Lighting condition from left to right: neutral, side, behind, and front

New open-set protocol

The XM2VTS dataset and closed-set Lausanne protocol were employed to enable comparisons with other algorithms; however, an open-set protocol is essential in order to test robustness and potential for real-world application and facilitate comparisons with other datasets in the future. Two separate LipAuth models were trained with the training data available in each protocol as previously described. Table 2 includes EER on both evaluation and test sets and the FAR at a 1% FRR.

Table 2 shows that the closed-set protocol performed slightly better than the open-set protocol throughout. This is not unexpected as the open-set protocol is a harder challenge with completely new users who must be enrolled during evaluation and testing. Upon further analysis of the test set errors when the threshold was set to the EER, the 1.03% achieved in the closed-set protocol equates to 15 attempted login tests where the individual could not login as themselves, and these 15 attempted logins were all from 3 individuals. From reviewing the 3 individuals, there were 2 male and 1 female, no notable facial hair or obvious physical characteristics were the cause. The corresponding 1.65% EER on the open-set protocol resulted from 40 attempted logins. The increased number of failed attempted logins was due to an increase in the number of tests in the new protocol, and the 0.62% increase in EER. There was an overlap in problematic individuals between the protocols, with approximately a third of the closed-set failed client tests also failing during the open-set test set. The open-set protocol test set contains 60 completely new individuals who had no representation in training, these 60 individuals produced over 2500 of the client-attempted login tests. In the open-set protocol test set results, 30 of the 40 failed attempted logins were from these new users. While, the majority of new clients were not problematic, this suggests that the LipAuth model could be overfit to the training data. In Table 3, the open-set EER is split into repeating clients who had a data representation in training and completely new clients.

There is a notable difference in Table 3 between the EER for new clients and clients with a representation of their data used to train LipAuth. This confirms the LipAuth model trained with the open-set protocol is overfit to the training data. Although the EER on new clients is higher than the repeating clients, at 2.80%, it shows the model can handle new unseen individuals reasonably well. The LipAuth model trained with the open-set protocol is used to test the real-world data sets below.

Table 2 Closed-set Lausanne protocol and open-set protocol results on the XM2VTS test set. Results are FRR at a 1% FAR

	Closed-set	Open-set
Evaluation EER	0.93%	1.21%
Test set EER	1.03%	1.65%
Test set at 1% FAR	1.07%	1.83%

Table 3 Results for the XM2VTS test set and open-set protocol on LipAuth

		EER breakdown
Full test set		1.65%
	Repeating clients	0.97%
	New clients	2.80%

The test set is split into existing clients who are represented in training and those who are not. This provides a measure of how much LipAuth is overfit to the training set

One-shot-learning

One of the biggest advantages of the Siamese network over existing traditional approaches such as [10, 28] is the one-shot-learning solution. One-shot-learning for LBBA is the ability to authenticate from a single enrolment video. This experiment was used in order to know if authentication improves with more enrolment data. For this experiment, to generate anchor videos longer than 20 digits, videos were concatenated to produce 40, 60, and 80 digit enrolment videos. The XM2VTS and open-set protocol was used for this experiment. The results are shown in Table 4.

Table 4 shows that there is very little difference between using more than 20 digits for enrolment. The Siamese network remained stable with a single 20 digit enrolment video, and no additional information was gained by adding more data to enrolment. For all future LipAuth experiments in this paper, a single video is used for enrolment.

Dataset direct comparison between XM2VTS and qFace

The XM2VTS dataset can be considered an 'ideal' dataset as it was recorded in consistant illumination with a steady camcorder and uniform recording conditions. LipAuth was trained using only XM2VTS and performs well on data from the same distribution as the training data, however showed signs of overfitting. The qFace dataset contains an identical 20-digit sequence (Table 1, Sequence 1) to the XM2VTS dataset to enable a measure of LipAuths performance on data captured on a mobile device.

Table 5 shows the performance of the qFace 'like-for-like' sequence to XM2VTS. The higher EER on qFace shows the embedding learned on the XM2VTS data did not transfer well to real-world data. This is expected as new clients in the XM2VTS test set were still from the same distribution as the training data. Training the LipAuth model using only XM2VTS data means the embedding learned has never had to account for the variations seen in real-world data.

This 'like-for-like' experiment with qFace involved 1,600 attempted logins, 160 of which are true client tests. The 6.25% contributed to 12 failed attempted logins, 8 of which resulted from 2 videos from a single person, Jack, who failed to log in against all 4 of their enrolment videos. The remaining missed clients are due to 1 individual, Molly, who was unable to log in against 1 of her enrolment videos given 4 attempts, despite being able to log in against all other enrolment videos with the same authentication videos. After inspection of the original videos, Molly seemed to not move her mouth much while speaking in the problematic enrolment video, in her other enrolment and test videos more movements were seen. The qFace dataset contained 7 males, 5 of which had notable facial hair. While Jack was the only participant with a clearly defined beard this did not appear to be the problem, his 2 problematic authentication videos appeared to contain an extreme change in illumination. In all Jack's enrolment videos, and 2 of his authentication videos, the videos appeared slightly darker. Figure 6 shows a sample of his 'normal' videos on the left, and one of the problematic videos on the right.

t-Distributed Stochastic Neighbour Embedding (t-SNE) [29] was used on the qFace embeddings calculated in this experiment in order to reduced the number of dimensions from 256 to 2 for visualisation. t-SNE uses a non-linear dimensionality reduction method to transform data from a high dimensional space and project it into a low-dimensional space, while maintaining a relationship between the structure of the points in the high dimensional space and the distance between them in the low-dimensional space. Figure 7 shows the resulting plot. Interestingly, it can be seen that the points for each individual are clearly clustered. Each point represents the embedding for a single video, where enrolment videos are shown with a border and authentication videos without. From Fig. 7 Jack's outlying videos can be seen to lie much further from the rest of his points.

qFace: testing digit quantity and content

The qFace dataset contains multiple digit sequences of varying number of digits from 20 to 4 digits. The specific digit content was planned to enable systematic testing of the digit order and quantity of sequences after training with the XM2VTS dataset. Videos were captured in as few sessions as possible so errors could be attributed to the enrolment and authentication content rather than major changes in individual's appearance or changes in environment.

Table 4 One-shot-learning enrolment data results

Num digits	EER
20	1.65%
40	1.94%
60	1.51%
80	1.79%

Results on XM2VTS, varying the number of digits used in enrolment against the EER test set

Table 5 Dataset comparison results: the EER showing the comparison between qFace and XM2VTS

Test set	EER
XM2VTS: full test set	1.65%
XM2VTS: new clients	2.80%
qFace	6.25%

All enrolment and authentication contained the same digit sequence content. Results compare the XM2VTS open-set protocol results with matching qFace sequence 1 from Table 1

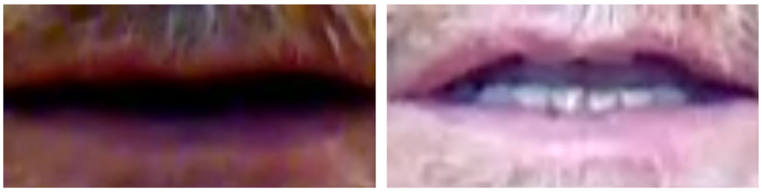

Fig. 6 qFace—Jack. Sample frame from two 20-digit videos in which the XM2VTS sequence was spoken. Video on the left represents the majority of Jack's data, while the sample on the right was from 1 of the 2 problematic videos for Jack

Matched enrolment and authentication

This experiment was carried out to see how the number of digits in a video affects performance. The experiment uses a 'like-for-like' number of digits within sequences for enrolment and authentication. This is important for user experience.

The one-shot-learning means using a single enrolment and authentication sequence which is already a desirable trait for a biometric for mobile devices, but if 20 digits provide no additional benefit over 4 digits, then why request 20 digits from users? In this experiment both the random and subset sequences were used. The first 4 videos for each sequence were used as individual enrolment videos, and the final videos as the authentication sequence. When testing against a 20-digit enrolment video, 20-digit authentication videos were used, for 10-digit anchor videos, 10 digit authentication sequences were used and so forth. For each of the 10 individuals this produces 8 enrolment videos and 8 authentication videos, meaning each experiment contained 8 enrolment videos × 8 authentication videos × people = 640 client scores and 5760 imposter scores. Results are shown in Fig. 8 .

In Fig. 8, the 10 digit enrolment and authentication sequence out performed all other sequence lengths with an EER of 5.00%; however, the 20 and 8 digit sequences scored similar EERs of 5.94% and 5.49%, respectively.

On closer inspection of the mistakes made by the 10-digit setup, the 5.00% EER equated to 32 out of 640 client tests where the client could not log in as themselves and 287 out of 5760 cases where an imposter was able to log in as another. Of the 32 problematic client scores, 16 were from a single individual, Beth. Beth had 8 rejected attempted logins from a single video that was unable to log in against any anchor video. On manual inspection of the video and other videos by Beth that were not problematic, it is unclear why it caused an issue. In the imposter tests, the same video by Beth was not able to log in to any imposter models either, and all of Beth's anchor videos were not susceptible to any imposter tests. Looking at the imposter scores, Adam's anchor videos were repetitively fooled by Joe, Mike and Molly, producing over half the imposter logins. The majority of the remaining mistakes involved Joe, Molly and Jack. Figure 9 gives a sample frame containing only the mouth region from each person in the qFace dataset. These results show

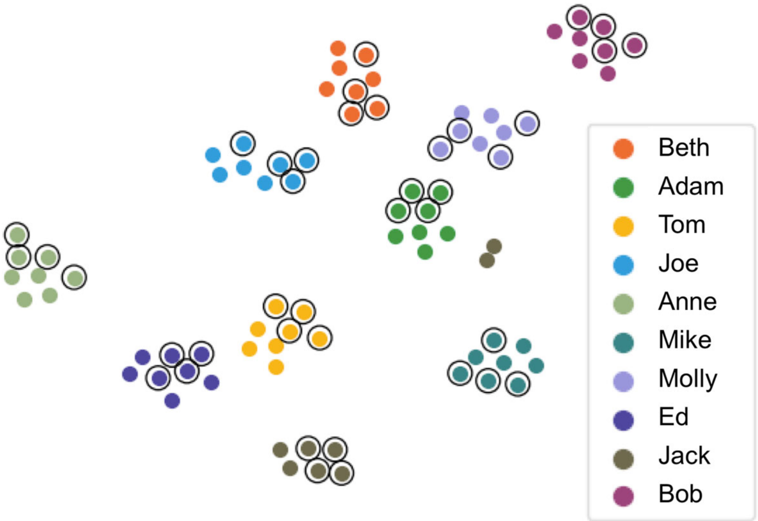

Fig. 7 Results using t-SNE to reduce videos from the qFace dataset from 256 to 2 dimensions. Each individual is represented by a unique colour. Enrolment videos are shown with a border and authentication videos without

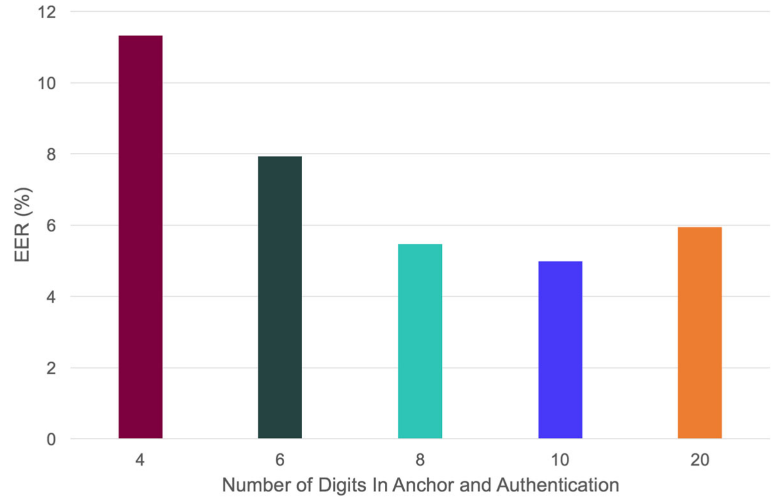

Fig. 8 Figure shows the performance using LipAuth trained on XM2VTS data tested on the qFace dataset for varied length sequences. Results reported on authentication sequences of matched number of digits to the Anchor video

that the LipAuth embedding is not specifically making mistakes based on gender or facial hair. Figure 9 also highlights the real-world element of the qFace dataset, as all volunteers were asked to record themselves in neutral lighting; however, a range of lighting conditions is observed.

Number of digits within authentication

The next experiment was created in order to take a closer look specifically at the number of digits within an authentication sequence. Following results from Figure 8, the number of digits in authentication was investigated using enrolment videos of length 8, 10 and 20 digits. A similar setup was implemented using the same anchor videos and authentication sequences, but this time multiple length authentication sequences were tested against each enrolment video, as shown in Table 6.

From Table 6, it can be clearly seen that each authentication sequence performed best when tested against an anchor video of the same length. It can also be seen with the current LipAuth weights and one-shot-learning approach that 4 and 6 digit sequences perform relatively poorly and it could be that, given their length, they contain less unique information in comparison to the longer sequences tested.

Mistmatched digit content

The previous experiments using the qFace dataset used all digit sequences of the same length. For this experiment the 20- and 10-digit videos were split into the 2 different sequences recorded as described in Table 1. Each length of sequence has a subset of the digits used in the XM2VTS sequence and a randomised digit sequence.

For this experiment, each of the 4 sequences used were tested as an enrolment video against matched digit content and an other sequence of the same length. This will show if the LipAuth embeddings favour the XM2VTS sequence it was trained with, and the effects of changing the content for a matched length sequence at test time. The results are shown in Table 7.

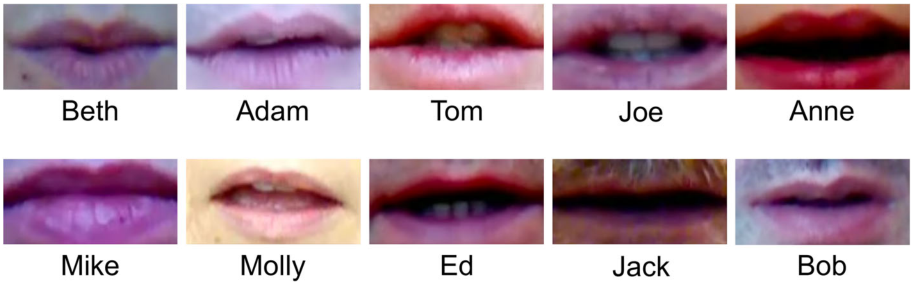

Fig. 9 qFace: sample frames from a 10-digit sequence. The frames were passed to the LipAuth model to obtain the embedding. The figure also highlights the real-world element of this dataset as all videos are considered 'neutral lighting'

Table 6 Number of digits in an enrolment video (Enrol) against number of digits in authentication video (Auth)

	Enrol	8	10	20
Auth	4	19.89%	17. 29%	17.41%
	6	10.63%	11. 27%	12.96%
	8	5.47%	9. 04%	10.73%
	10	9.77%	*5.00%*	8.49%
	20	10.24%	8.44%	*5.94%*

Results show EER on qFace dataset, using the Siamese network

Table 7 shows 4 different enrolment videos against authentication sequences of matched number of digits, where italic highlights matched content as well. The results on the 10-digit sequences from Table 7 appear in general better than the 20-digit results and suggest that content appears to have little effect. The lowest EER produced in this experiment involved enrolling with a random 10-digit sequence and authenticating with a different 10-digit sequence, achieving 2.50%. This is lower than the 2.80% achieved on the new clients in the XM2VTS test set.

When enrolled with the XM2VTS sequence both 20 digit authentication sequence and the content of the authentication sequence had little effect. However, when enrolled and authenticated with the 20-digit random sequence, the performance dropped to 2.88% which is comparable to the new clients in the XM2VTS test set.

FAVLIPS: varied real-world content

The FAVLIPs dataset was designed and collected over 4-month-long sessions, to test the performance of LBBA on mobile devices under a series of challenging conditions including miming, random sentences and varied lighting conditions.

It is already known from the previous experiments that the LipAuth model is overfit to the XM2VTS dataset; however, the performance on qFace suggested the embedding learned from the XM2VTS performed fairly well on controlled mobile data . The FAVLIPs dataset was not collected in a controlled manor, providing a much more difficult challenge.

The setup of this experiment mimics as closely as possible the previous experiments. The same enrolment videos were used throughout this experiment, which includes the 10-digit sequences from the first 3 sessions, this means each individual had 6 separate enrolment videos. All enrolment videos were recorded in neutral lighting and 6 different categories of authentication sequences were tested:

1. Spoken Digit Sequences with 2 authentication sequences per individual containing different 10-digit sequences. Recorded in neutral lighting
2. Mimed Digit Sequences with 2 authentication sequences per individual, with the same content as the previous spoken digit sequences. Recorded in neutral lighting
3. Sentences with 5 random authentication sentences per individual. Recorded in neutral lighting
4. Light Front with 1 authentication sequence per individual. The content spoken in the sequence is the same 10-digit sequence as one of the digit sequences spoken aloud and mimed
5. Light Side with 1 authentication sequence per individual. The content spoken in the sequence is the same 10-digit sequence as the light front sequence.
6. Light Behind with 1 authentication sequence per individual. The content spoken in the sequence is the same 10-digit sequence as the light front sequence.

The results shown in Fig. 10 show LipAuth trained using XM2VTS data only, is not portable to real-world applications. Despite the extremely poor performance, there was no notable difference between spoken digits, mimed digits and sentences for authentication. However, as the lighting conditions vary, the performance significantly degrades.

Possible solutions to improve these results could include preprocessing of the data, filtering out videos that contain extreme lighting conditions or retraining LipAuth. The quality of one-shot-learning is entirely dependent on the quality of the embedding, and the quality of the embedding is dependent on the quality of the training data. When training a Siamese network for LBBA, ideally the training data should contain a wide range of challenging real-world conditions. It could be expected that an embedding trained with enough real-world data, could be

Table 7 qFace results showing the EER of the separate 10 and 20 digit sequences

	Enrol	20-XM2VTS	20-Random	10-Subset	10-Random
Auth	20-XM2VTS	*6.25%*	8.12%	-	-
	20-Random	6.36%	*2.88%*	-	-
	10-Subset	-	-	*4.50%*	4.60%
	10-Random	-	-	2.50%	*5.83%*

Results show 4 different enrolment sequences against authentication sequences of matched length and varied content. Results in italics mark the results when the enrolment and authentication sequences matched

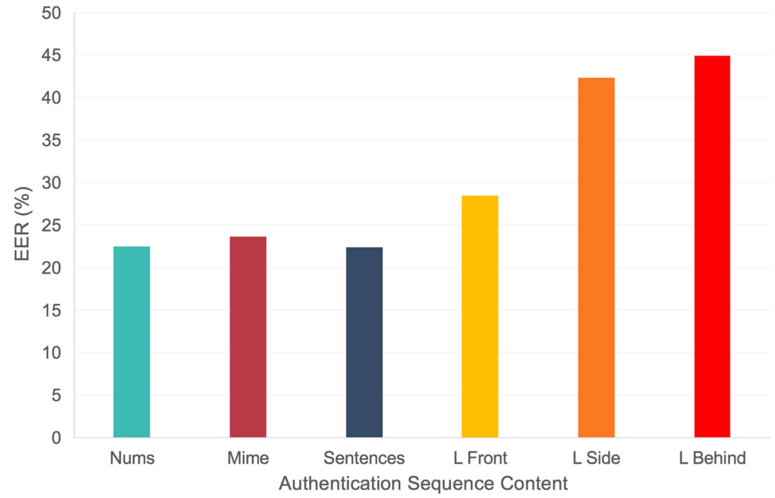

Fig. 10 Results from LipAuth trained with XM2VTS data on the FAVLIPS dataset

robust to these challenges and, therefore, more suited to deployment on mobile devices.

Retraining LipAuth weights with real-world content

In theory, the LipAuth embedding could be improved using real-world data, learning from other people's videos so that it improves on challenging conditions for new individuals. To test this theory LipAuth was retrained with videos from the FAVLIPS dataset. The FAVLIPS dataset contains 52 recordings from each of the 42 individuals. All data for 22 of the individuals was added to the pool of training data. The remaining 20 individuals still will not have a representation seen during training and will be completely new unseen test subjects. If each individual contributes 52 videos, then each individual contributes 1326 anchor-positive pairs for generating triplets. With 22 individuals that total 29,172 possible anchor-positive pairs. It would not be expected that 22 individuals alone could provide a robust embedding that will generalise well to the whole population, so updating the weights will include XM2VTS and FAVLIPS data. Combining XM2VTS and FAVLIPS training data produces over 30,000 anchor-positive pairs for selecting triplets. Given the computational constraints, 10% the available training data was randomly selected of the semi-hard and easy triplets on a 5 to 1 of FAVLIPS to XM2VTS ratio to train each epoch.

Two models were generated using FAVLIPS data, one model involved updating the LipAuth weights from the open-set protocol, and second trained LipAuth from scratch with randomly initialised weights. When updating the existing weights the model trained for an additional 40 epochs before the loss plateaued. When training from scratch the model was trained for 60 epochs before the training loss significantly decreased and the model showed signs of overfitting. The results of the updated weights against each lighting condition are presented in Table 8.

Table 8 shows a side by side comparison of the LipAuth model trained with just XM2VTS data, weights updated with training data that contained real-world and XM2VTS data and trained from scratch with real-world and XM2VTS data. Compared to the baseline model, both models show FAVLIPS data improved significantly with additional real-world training data. The best performance on the four different lighting conditions are shown in italics. Neither of the models trained with FAVLIPS data outperformed the other on all lighting conditions, though both setups do show significant improvements for all lighting conditions. Every lighting condition improved by over 10%, with the light behind test case improving by more than 20%. These results were obtained using one-shot-learning with anchor videos recorded in neutral lighting conditions, up to 4 months before the authentication videos, and the model has not seen any representation of the individuals tested. This confirms that a Siamese network can learn and improve on challenging conditions from other individuals' data.

Table 8 Comparison of results from the 3 different LipAuth model weights presented in this section, against lighting conditions for the 20 unseen individuals from FAVLIPS

	XM2VTS only	Updated weights	Retrained
XM2VTS: Evaluation Set	1.21%	1.95%	5.60%
FAVLIPS: Neutral Nums	22.43%	13.79%	*10.83%*
FAVLIPS: Light Front	28.44%	*17.50%*	20.54%
FAVLIPS: Light Side	42.29%	36.67%	*30.00%*
FAVLIPS: Light Behind	44.91%	*24.17%*	29.12%

Comparison with traditional approach for verification of network overtraining

The experiments in this work have shown that the LipAuth model is overfit to the training data and the large decrease in performance on real-world data was attributed mostly to this. While, in an attempt to rectify this problem the LipAuth model was retrained containing a sample of the real-world data and the performance was shown to improve, it could be argued that perhaps the challenging nature or sparsity of the real-world data was a greater problem than the overtraining. In order to test this theory, the FAVLIPS dataset was tested using the LBBA approach from work in [10], using DCT coefficient features modelled with GMMs. The setup proposed in [10] was selected as it did not use deep learning and was the previous state-of-the-art using the XM2VTS and Lausanne protocol prior to LipAuth. Work in [10] used DCT coefficient features selected from 4 separate 20-digit videos to train a 32-mixture GMM per person during enrolment. Table 9 shows the results of varying the amount data used during enrolment on the XM2VTS dataset. This experiment was carried out to provide intuition into the amount of data required during enrolment for the FAVLIPS dataset with the traditional approach.

Results in Table 9 for the traditional approach show a strong correlation between the EER and the amount of data used during enrolment. As the enrolment data is decreased the performance significantly deteriorates, whereas the LipAuth embedding in remains more stable.

Given the need for more enrolment data, the FAVLIPS dataset was tested using the traditional LBBA approach with the following 3 different enrolment setups:

1. Model trained with digit sequences only—6 videos
2. Model trained with 15 phonetically rich sentences—15 videos
3. Model trained with Session 1 data only—12 videos

While the varied amount of enrolment data prevents a direct comparison between LipAuth and the traditional approach, it does provide a more realistic testing scenario for the traditional approach as more than 1 enrolment video would be required.

Table 9 Number of digits used in enrolment for LipAuth (as in Table 4) and traditional approach using DCT coefficient features and GMMs (described in [10])

	EER on	LipAuth	Traditional approach
Num enrolment digits			
20		1.65%	16.49%
40		1.94%	7.93%
60		1.51%	2.64%
80		1.79%	1.88%

These results show the EER on the XM2VTS open-set test set. Results from Table 4 are included for comparison

Results in Fig. 11 show some improvements on the FAVLIPS dataset over the LipAuth model. However, the authentication sequences captured under the most challenging lighting conditions prove problematic for both approaches. The figure shows that as the amount of data used during enrolment increased so did the performance on the challenging authentication data, but this came at the expense of the performance on the other authentication sequences.

One of the most significant differences between LipAuth and this traditional approach to LBBA is enrolment. While the features can be fine tuned for the traditional approach using one dataset and applied to another, it cannot learn an embedding from another data set that can be applied to new data. Results in Table 8 showed that when a small amount of real-world data was used in training LipAuth, the results across new enrolling persons improved. With the traditional approach every new person enrolling will always need to provide a large profile of diverse video data for training their own model, as it cannot learn from other people's data. Figure 11 also demonstrated that adding more diverse data may reduce the model performance on less challenging authentication sequences.

Discussion

The work in this paper provided a rigorous investigation into LBBA in the real-world. A new open-set protocol for the XM2VTS was proposed and benchmarked for LBBA achieving an EER of 1.65% and FRR of 1.83% at a 1% FAR. This 1.65% EER is close to the closed-set 1.03% existing state-of-the-art [11], despite the open-set protocol being a significantly harder challenge. This is this first time LBBA has been investigated beyond a closed-set scenario. Furthermore, results showed the one-shot-learning solution was stable for a single enrolment video, which has many benefits over existing approaches presented in [10, 28, 30] which require multiple enrolment examples.

The qFace dataset was collected and used as an additional test set to enable systematic testing of varied digit content and length of sequences for both enrolment and authentication. The results showed for the first time LBBA with real-world data. The results on the qFace dataset were unpredictable as the LipAuth embeddings were not robust to natural deviations within real-world data. Results for 20-digit sequences ranged from EER of 2.88 to 8.12%, and EER for 10-digit sequences from 2.50 to 5.83%. The qFace dataset did perform best when digit sequences contained at least 8 digits, and the number of digits in an authentication sequence matched the number of digits in enrolment.

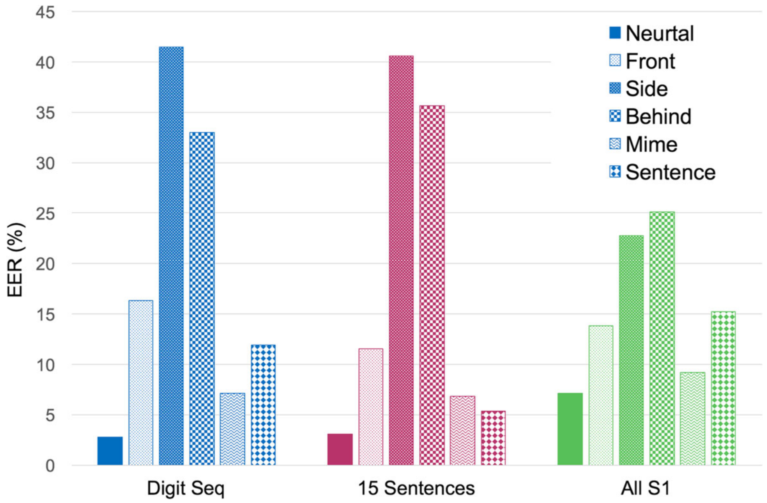

Fig. 11 Results for the FAVLIPS dataset using traditional LBBA approach from [10]. EER against 3 different enrolment models per person: using digit sequences only (blue), sentences only (pink), and all data from Session 1 (green). Each model was tested against 6 different authentication sequences, shown in different textures

The FAVLIPS dataset presented some of the hardest challenges that could be expected for LBBA. Enrolment and authentication data were separated by up to 4 months and individuals had complete control over their own recordings. FAVLIPS was used to compare the results when the authentication sequence were spoken, mimed, sentences and 3 additional lighting conditions. The benchmarking results on the neutral lighting FAVLIPS sequences showed spoken digit sequences, sentences and miming all achieved and EER of 22.50% ± 1%. Despite the extremely poor performance, it was noted that there was no significant difference between all 3 of these challenges. However, significant declines in performance were seen with the variations in the lighting. These results, while significantly worse than the results on XM2VTS in this paper, are comparable to the results reported for LBBA on the XM2VTS in other works [2, 3], where they reported 19.7% and 22%, respectively.

Experiments in this work showed how the LipAuth model could be improved to better handle real-world data for LBBA on mobile devices. Results showed that a drop in EER from 44.91 to 24.17% (relative reduction of 46%) could be achieved on the hardest lighting challenge when only a small amount of real-world data was added to the LipAuth training data.

The final experiments in this work showed the performance of the LipAuth model compared to a more traditional approach to LBBA as proposed in [10]. Using a single enrolment video the LipAuth model achieved 1.65% EER on the XM2VTS, whereas the traditional approach obtained a 16.49% EER on the experiment. The results showed that the traditional approach required significantly more enrolment data to get comparable results on the same test set. The FAVLIPS dataset was tested using the traditional approach and these results highlighted that despite using 12 times as many enrolment videos the traditional approach had difficulty with the real-world challenges too. It could be concluded from these final experiments that the performance worsening noticed for the LipAuth model was not completely due to the overtraining of the LipAuth weights.

Conclusions

While recent research has shown promising results for LBBA on the highly controlled video recordings of the XM2VTS dataset, work in this paper demonstrates the challenges still faced for real-world deployment on mobile devices for models such as LipAuth. The challenges investigated here included the length and content of enrolment and authentication sequences as well as illumination effects. It was found that varying content had little impact on performance which is crucial for liveness checks, while illumination was shown to present the greatest challenge. To demonstrate one path forward, this paper showed that training the LipAuth model on examples of varied lighting leads to significant improvements in performance (46% relative reduction in errors). However, to push the field forward and more inline with other biometrics such as face or fingerprint, a larger training set of data expected from mobile devices is needed. Future work could also explore preprocessing techniques such as [31, 32] to correct for these illumination effects. Accounting for these illumination effects could be key to developing LBBA systems suitable for deployment on mobile devices.

Open-set protocol for XM2VTS

Training

Open-Set Protocol for XM2VTS: The ID Numbers of the 150 Training Clients.

003, 004, 005, 009, 012, 016, 017, 019, 020, 021, 022, 024, 025, 027, 029, 030, 033, 035, 036, 037, 038, 040, 042, 045, 049, 050, 051, 052, 053, 055, 058, 060, 061, 064, 066, 069, 071, 073, 078, 079, 080, 085, 089, 090, 099, 101, 102, 103, 105, 110, 112, 115, 116, 121, 123, 125, 126, 129, 132, 133, 137, 138, 140, 145, 148, 150, 152, 154, 159, 163, 164, 165, 166, 167, 168, 169, 173, 181, 182, 183, 188, 191, 193, 196, 198, 206, 207, 208, 209, 210, 211, 219, 221, 222, 227, 229, 231, 232, 235, 237, 240, 244, 246, 248, 253, 255, 258, 259, 261, 264, 267, 269, 270, 275, 278, 279, 282, 285, 287, 288, 289, 290, 292, 293, 295, 305, 310, 312, 316, 321, 322, 324, 325, 328, 329, 330, 332, 336, 337, 338, 339, 340, 357, 358, 360, 364, 365, 369, 370, 371

Evaluation

Open-Set Protocol for XM2VTS: The ID Numbers of the 45 New Evaluation Clients.

006, 013, 018, 026, 032, 034, 041, 054, 056, 065, 068, 072, 074, 075, 082, 091, 092, 108, 113, 124, 135, 136, 141, 146, 179, 180, 197, 213, 216, 218, 224, 228, 233, 236, 243, 266, 274, 281, 319, 320, 334, 342, 359, 362, 366

Open-Set Protocol for XM2VTS: The ID Numbers of the 20 Evaluation Imposters.

000, 002, 046, 057, 083, 093, 104, 120, 143, 157, 177, 187, 189, 203, 212, 215, 242, 276, 301, 314

Testing

Open-Set Protocol for XM2VTS: The ID Numbers of the 60 New Test Clients.

001, 007, 010, 028, 031, 043, 044, 047, 062, 067, 070, 081, 086, 088, 095, 098, 107, 109, 119, 122, 127, 128, 130, 134, 149, 153, 155, 158, 160, 171, 172, 174, 175, 176, 178, 185, 190, 199, 200, 201, 202, 225, 234, 241, 249, 250, 263, 272, 280, 283, 284, 300, 315, 313, 317, 318, 323, 331, 333, 335

Open-Set Protocol for XM2VTS: The ID Numbers of the 20 Test Imposters.

008, 011, 023, 039, 048, 059, 087, 096, 111, 114, 131, 142, 147, 161, 170, 226, 271, 286, 241, 367

Acknowledgements

The authors would like to acknowledge and thank all volunteers who participated in the data collection process.

Authors' contributions

The authors conceived, designed, and performed experiments. Data for this work was collected by author 1. Author 1 was the primary author of the manuscript. All authors read and approved the final manuscript.

Funding

Not applicable

Availability of data and materials

The datasets generated and/or analysed during the current study are available from the authors for other university research groups upon reasonable request.

Competing interests

The authors declare that they have no competing interests.

Received: 30 July 2019 Accepted: 23 January 2020
Published online: 12 March 2020

References

1. M. U. R. Sanchez, *Aspects of Facial Biometrics for Verification of Personal Identity. PhD thesis*. (University of Surrey, Guilford, UK, 2000)
2. S. Lucey, in *Audio-and Video-Based Biometric Person Authentication*, ed. by J. Kittler, M. S. Nixon. An evaluation of visual speech features for the tasks of speech and speaker recognition (Springer, Berlin Heidelberg, 2003), pp. 260–267
3. M. I. Faraj, J. Bigun, in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference On*. Motion features from lip movement for person authentication, vol. 3 (IEEE, New York, 2006), pp. 1059–1062. https://doi.org/10.1109/ICPR.2006.814
4. H. E. Cetingul, Y. Yemez, E. Erzin, A. M. Tekalp, Discriminative analysis of lip motion features for speaker identification and speech-reading. Trans. Img. Proc. **15**(10) (2006). Piscataway. https://doi.org/10.1109/TIP.2006.877528
5. A. G. de la Cuesta, J. Zhang, P. Miller, in *Machine Vision and Image Processing Conference 2008. IMVIP '08 International*. Biometric identification using motion history images of a speaker's lip movements (IPRCS, Ireland, 2008), pp. 83–88. https://doi.org/10.1109/IMVIP.2008.13
6. G. Chetty, M. Wagner, Biometric person authentication with liveness detection based on audio-visual fusion. Int. J. Biometrics Int. J. Biometrics. **1**, 463–478 (2009). Inderscience Publishers, Geneva. https://doi.org/10.1504/IJBM.2009.027306
7. T. Nakata, M. Kashima, K. Sato, M. Watanabe, in *Biometrics and Kansei Engineering (ICBAKE), 2013 International Conference On*. Lip-sync personal authentication system using movement feature of lip (IEEE, New York, 2013), pp. 273–276. https://doi.org/10.1109/ICBAKE.2013.53
8. X. Shi, S. Wang, J. Lai, in *2016 IEEE International Conference on Image Processing (ICIP)*. Visual speaker authentication by ensemble learning over static and dynamic lip details (IEEE, New York, 2016), pp. 3942–3946. https://doi.org/10.1109/ICIP.2016.7533099
9. P. Singh, V. Laxmi, M. S. Gaur, in *2012 5th IAPR International Conference on Biometrics (ICB)*. Speaker identification using optimal lip biometrics (IEEE, New York, 2012), pp. 472–477. https://doi.org/10.1109/ICB.2012.6199795
10. C. Wright, D. Stewart, P. Miller, F. Campbell-West, R. Dahyot, G. Lacey, K. Dawson-Howe, F. Pitié, D. Moloney, in *Irish Machine Vision & Image Processing Conference Proceedings 2015*. Investigation into dct feature selection for visual lip-based biometric authentication (IPRCS, Ireland, 2015), pp. 11–18. Winner of Best Student Paper Award
11. C. Wright, D. Stewart, in *Advances in Visual Computing*, ed. by G. Bebis. One-shot-learning for visual lip-based biometric authentication (Springer, New York, 2019), pp. 405–417
12. Y. M. Assael, B. Shillingford, S. Whiteson, N. de Freitas, Lipnet: Sentence-level lipreading. CoRR abs/1611.01599. **abs/1611.01599**, 1611–01599 (2016)
13. J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, T. Chen, Recent advances in convolutional neural networks. Pattern Recogn. **77**, 354–377 (2018). https://doi.org/10.1016/j.patcog.2017.10.013
14. J. Chung, C. Gulcehre, K. Cho, Y. Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling. CoRR abs/1412.3555. **abs/1412.3555** (2014)
15. A. Graves, J. Schmidhuber, in *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005*. Framewise phoneme classification with bidirectional lstm networks, vol. 4, (2005), pp. 2047–20524. https://doi.org/10.1109/IJCNN.2005.1556215
16. G. Koch, R. Zemel, R. Salakhutdinov, in *ICML Deep Learning Workshop*. Siamese neural networks for one-shot image recognition, vol. 2 (Springer, 2015)
17. F. Schro, D. Kalenichenko, J. Philbin, Facenet: A unified embedding for face recognition and clustering. CoRR abs/1503.03832. **abs/1503.03832**, 1503–03832 (2015)

18. E. Ahmed, M. Jones, T. K. Marks, in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. An improved deep learning architecture for person re-identification, (2015), pp. 3908–3916. https://doi.org/10.1109/CVPR.2015.7299016
19. C. Wright, *Lip-based biometric authentication. PhD thesis,* (Queen's University Belfast, Belfast, UK, 2019)
20. K. Messer, J. Matas, J. Kittler, K. Jonsson, in *In Second International Conference on Audio and Video-based Biometric Person Authentication*. Xm2vtsdb: The extended m2vts database, (1999), pp. 72–77
21. J. Luettin, G. *Maître*, Evaluation protocol for the extended M2VTS database (XM2VTSDB). Idiap-Com Idiap-Com-05-1998, IDIAP. **0** (1998)
22. T. Kanade, A. Jain, N. K. Ratha, *Audio-and Video-Based Biometric Person Authentication: 5th International Conference, AVBPA 2005*. (Proceedings Springer, Hilton Rye Town, NY, USA, 2005, 2005)
23. L. F. Lamel, R. H. Kassel, S. Sene., Speech database development: design and analysis of the acoustic-phonetic corpus. Speech Input/Output Assessment and Speech Database. **2**, 161–170 (1986)
24. T. J. Hazen, K. Saenko, C.-H. La, J. R. Glass, in *Proceedings of the 6th International Conference on Multimodal Interfaces. ICMI '04*. A segment-based audio-visual speech recognizer: Data collection, development, and initial experiments (ACM, New York, NY, USA, 2004), pp. 235–242. http://doi.acm.org/10.1145/1027933.1027972
25. Y. W. Wong, S. I. *Chîng4*, K. P. Seng, L.-M. Ang, S. W. Chin, W. J. Chew, K. H. Lim, A new multi-purpose audio-visual unmc-vier database with multiple variabilities. Patt. Recogn. Letters. **32**(13), 1503–1510 (2011). https://doi.org/doi:10.1016/j.patrec.2011.06.011. Page 22 of 23
26. T. Afouras, J. S. Chung, A. Zisserman, in *arXiv Preprint arXiv*. Lrs3-ted: a large-scale dataset for visual speech recognition, (2018), pp. 1809–00496
27. V. Kazemi, J. Sullivan, in *2014 IEEE Conference on Computer Vision and Pattern Recognition*. One millisecond face alignment with an ensemble of regression trees (IEEE Computer Society, Los Alamitos, 2014), pp. 1867–1874
28. X. X. Shi, S. L. Wang, J. Y. Lai, in *2016 IEEE International Conference on Image Processing (ICIP)*. Visual speaker authentication by ensemble learning over static and dynamic lip details, (2016), pp. 3942–3946. https://doi.org/doi:10.1109/ICIP.2016.7533099
29. L. J. P. van der Maaten, G. E. Hinton, Visualizing data using t-sne. J. Mach. Learn. **9**, 2579–2605 (2008)
30. L. Lu, J. Yu, Y. Chen, H. Liu, Y. Zhu, L. Kong, M. Li, Lip reading-based user authentication through acoustic sensing on smartphones. IEEE/ACM Trans. Networking. **27**(1), 447–460 (2019). https://doi.org/doi:10.1109/TNET.2019.2891733
31. W. Chen, Er. Meng Joo, Wu. Shiqian, Illumination compensation and normalization for robust face recognition using discrete cosine transform in logarithm domain. IEEE Trans. Syst, Man, Cybernet. Part B (Cybernetics). **36**(2), 458–466 (2006). https://doi.org/10.1109/TSMCB.2005.857353
32. A. Baradarani, Q. M. J. Wu, M. Ahmadi, An efficient illumination invariant face recognition framework via illumination enhancement and dd-dtccwt filtering. Pattern Recogn. **46**, 57–72 (2013). https://doi.org/10.1016/j.patcog.2012.06.007

Publisher's Note

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