Liveness Verification Using Deep Neural Network Based Visual Speech Recognition

Philip McShane[1], Darryl Stewart[1,2]
[1]EEECS Queens University Belfast
[2]Liopa Ltd
Belfast, UK

Abstract

We present a novel approach to liveness verification based on visual speech recognition within a challenge-based framework which has the potential to be used on mobile devices to prevent replay or spoof attacks during Facebased liveness verification. The system uses visual speech recognition and determines liveness based on the Levenshtein Distance between a randomly generated challenge phrase and the hypothesis utterances from the visual speech recognizer. A Deep learning-based approach to visual speech recognition is used to improve upon the state of the art for the use of visual speech recognition for liveness verification. A comparison of a number of different deep neural networks is presented. The performance of these deep neural networks on a visual speech recognition task (digit recognition) is used to determine which is the most suitable for use in a liveness verification system. Experiments are performed using Long short-term memory (LSTM), bi-directional long short-term memory (BLSTM) and time delay neural networks (TDNN) with a Gaussian mixture model (GMM) being used to provide a baseline for comparison. The best performing of these is then used for determining the performance of a visual speech-based liveness verification system on a variety of phrase lengths.

1. Introduction

Alternatives to the use of passwords are increasingly being considered as means of securing access to electronic devices such as laptops and phones. The most common approaches towards user authentication for gaining access to these devices make use of passwords, user IDs, identification cards and PINs. These techniques have a number of limitations: Passwords and PINs can be guessed, stolen or illicitly acquired by surveillance or brute force attack. There have been many high-profile hacks emanating from password breaches in recent times. These hacks allow malicious individuals to gain access to a system using the credentials of a valid user without the user being present.

In order to enhance security, alternatives to the passwordbased approaches have been considered and these have primarily been focused on forms of Biometric authentication. A number of different biometrics have been proposed, with the most popular involving recognition of the Face [1], Voice [1] or Fingerprint [1], [2]. These systems, while more secure than passwords, also have some limitations. Fingerprint scanning systems are accurate, fast and robust, however, they can be susceptible to forms of ‘spoofing’ whereby false fingerprints, can be used to fool the sensor [2]. A further limitation is the additional cost of having a dedicated fingerprint sensor within the device means that few devices have offered fingerprint scanning as an authentication process.

Speech recognition systems can be deployed inexpensively and universally to all mobile device types as they use only the standard microphone in the device. Voice has been shown to be highly accurate and reasonably robust in quiet environments. The performance can be affected by the presence of loud and/or time-varying background noises. Furthermore, in some environments, it may be considered inappropriate or indiscrete to speak clearly into a microphone.

Face recognition has been shown to be highly accurate and can be robust to changes in the user's environment, appearance, variations in pose and illumination conditions. A key concern with face recognition systems is that they may be susceptible to spoofing attacks where an unauthorized user holds a photograph in front of the camera and gains access as the person in the photo [3]. These forms of attack are more likely to be successful in the unsupervised, remote access use cases involving mobile devices. The security of remote unsupervised face recognition systems would be significantly improved by ensuring that “liveness” detection is included in the authentication process, thereby ensuring that the authorized user is present and responds intelligently when prompted for input by the system. In this paper, a means of liveness verification based on visual phrase verification algorithm which uses a visual speech recognition system within a challenge-based verification framework. Specifically, the process of verification involves the user being challenged to say

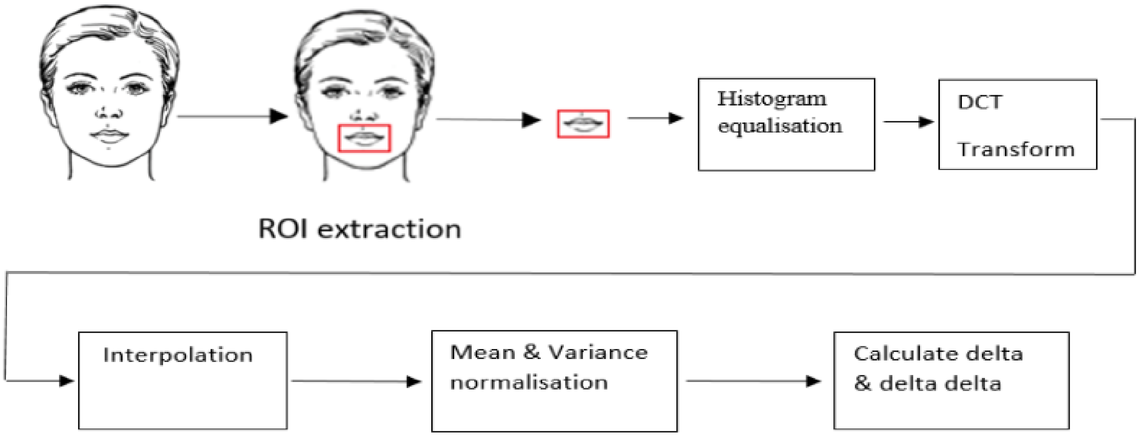


Figure 1. The feature extraction process

a randomly generated string of digits which they will then speak into the phone's camera.

Visual speech recognition will be performed on the video and if the visual recognition system is confident that the video contains lip movements which match the challenge phrase, then the ‘liveness’ of the user will be verified. The challenge phrases are randomly generated at each verification attempt to limit the possibility of replay attacks using previously recorded videos. For practical use, this approach to liveness verification would be combined with other biometric authentication processes such as face recognition in order to improve the overall security and robustness of the biometric system and would not inconvenience the user significantly beyond the standard face capture process. Visual speech recognition has been the focus of extensive research in recent years and has matured to the point that it can be used robustly for limited vocabulary tasks [4], [5]. Prior research on the use of visual speech recognition for biometric applications have focused on the use of the visual information combined with audio [6], [7] and most of the research has focused on using visual speech as an alternative means of verifying the user's identity not for verifying liveness. Evano and Besacier [8] investigated liveness verification based upon an analysis of the synchronicity of visual and audio features and reported an Equal Error Rate of 14.5% using the XM2VTS dataset. In [10] a liveness verification system based on only using visual information was proposed that is based on speech recognition with an SVM (support vector machine) to recognize digits that had been individually segmented. A speech recognition rate of 68% was reported on the XM2VTS dataset, using the approach in [10] with only the visual modality. In this paper, the aim is to show an improvement over previous works through the use of deep learning.

2. Visual speech recognition

2.1. Feature extraction

Visual speech recognition aims to determine the text spoken by an individual based on the movement of their lips. When a visual speech recognition system receives a video the first step to performing recognition is first to determine where in the images the lips are located and to extract the lip region to be used as the region of interest (ROI). The region of interest for a number of speakers from the dataset can be seen in Figure 3, from this it can be seen that the lips and the area around the lips are used to make up the ROI that is used as the input to our system. For the system that is used in this paper, the Dlib image processing library was used [9]. Dlib provides a facial landmark detector that has been used to locate and extract the ROI from each video frame, this process is described in [9]. Once the frames have been extracted from the video histogram equalization is applied to it in order to make the system more robust to changes in illumination. The frame is then converted to grayscale and then down sampled to a 16x16 image. After this, a DCT (discrete cosine transform) transform is applied to the frame. The DCT transform was chosen as it was shown to give good performance in [5]. A triangular mask is then applied to the result of the DCT transform and from this the 15 lowest frequency coefficients are selected with the DC component being removed, leaving 14 DCTs. The DC component is removed as initial experiments showed that the system performed better when the DC component was not present. Mean and variance normalization is then applied to the feature vectors. The number of features is increased through cubic spline interpolation to 100 fps, as this was found to increase the performance of the visual speech recognizer. From the 14 DCTs, differential and acceleration coefficients are calculated. These are then concatenated with the 14 DCTs to give a feature vector of 42 coefficients.

2.2. Lip modeling

Deep learning approaches have shown promise in solving problems in areas such as computer vision [11], [12], audio speech recognition [13] and natural language processing [14]. In order to create a visual speech recognition system that is capable of performing to a level comparable with audio-based speech recognition software, a deep learning-based approach was chosen. By incorporating such an approach, the aim is to produce a system that would be suitable for real-world applications. A number of different deep neural networks have been proposed for use in a DNN/HMM system [13] [15] [16]. In order to determine which neural network would be the most suitable for use in this work, a number of different experiments were conducted using the different architectures. Long short-term memory (LSTM) is a kind of recurrent neural network. These are contrasted with feed forward neural networks. In a feed-forward neural network, the network is trained by passing through the training data and adjusting the weights in order to minimize the objective function. When trained the network classifies the inputs without reference to the prior inputs to the system. In contrast to this within a recurrent neural network, the prior inputs are looped through the network allowing information about the previous step to be used. An LSTM makes use of self-contained memory units. These memory units are comprised of 3 gates and an internal state within a cell unit, the structure of the cell can be seen in Figure 2. These gates consist of an input, output and forget gate. The input and output gates control the extent to which information passes into and out of the cell. The forget gate is important to the function of the cell as a unit capable of learning long-term dependencies. The internal state allows the cells to deal with the problem of exploding/shrinking gradients through a process is known as constant error carousel [15]. The internal state of the cell is a node with a linear activation function and a fixed weight. As the internal state has a self-connected recurrent edge with a fixed weight error is able to flow across time steps without exploding or vanishing this is known as constant error carousel. The forget gate is used to flush the internal state of the cell allowing long-term dependencies to be discarded. The operation of these memory units can be expressed as:

$$ft = \sigma g(Wfxt + Ufht{-}1 + bf) \qquad (1)$$

$$it = \sigma g(W\ ixt + Uiht{-}1 + bi) \qquad (2)$$

$$ot = \sigma g(W\ oxt + Uoht{-}1 + bc) \qquad (3)$$

$$ct = ft{\circ}ct{-}1 + it{\circ}\sigma c(W\ cxt + Ucht{-}1 + bc) \qquad (4)$$

$$ht = ot{\circ}\sigma h(ct) \qquad (5)$$

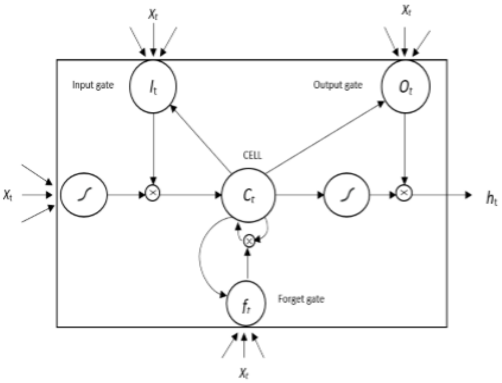


Figure 2. Long short-term memory cell

Where f is the forget gates activation vector, i is the input gates activation vector, o is the input gates activation vector, h is the output vector of the LSTM with t indicating the timestep. U and W are the weight matrices for the input and recurrent connections, ° is the Hadamard product, σg is the sigmoid activation function, σh and σc are the hyperbolic tangent function, Ct is the cell and b is the bias vector. The connections between the gates and the cell can be seen in Figure 2 with the flow of information through the memory unit being indicated by the arrows. Bidirectional long short-term memory (BLSTM) is a variation on the LSTM network that encodes information from both past and future. A BLSTM network is trained as separate RNNs with both a left context, that is from past to future and a right context, that is information from future to past. It may appear that this would not be suitable for speech recognition where normally only the left context is available. However, speech recognition does not require that an output is produced after every input but rather after each segment which allows the system to make use of the right context. Time delay neural networks (TDNN) [16] [17] are a kind of neural network originally developed for use in audio-based speech recognition [16]. Its suitability for use in speech recognition is based on its time shift invariance. This is important for processing speech as result of differences in pronunciation of sub-word units, accents or other factors that can affect the temporal characteristics of speech. As their name suggests TDNN add delays to their neural network architecture in order to achieve time shift invariance. TDNN are implemented as a feedforward neural network with the delays being added by making a copy of the features and their outgoing connections prior to updating the original value [17]. The number of steps that are saved is known as the delay. The features and their respective delays are fully connected to the next layer, these units are referred to as the receptive fields. The delays modify the activation of a unit by multiplying the inputs

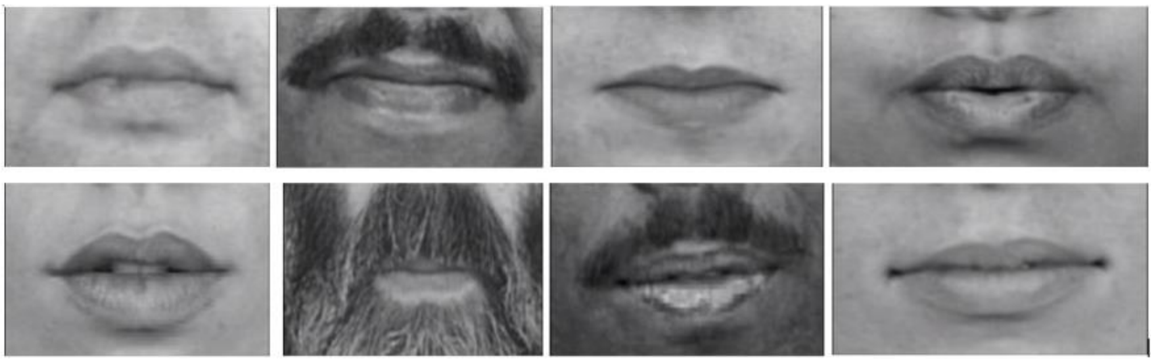


Figure 3. Cropped images from XM2VTS dataset

by the layers N delay steps. This adds a temporal dimension to the network making it suitable for processing time series data such as speech.

3. Comparison of visual speech recognition models

3.1. XM2VTS Dataset

For the experiments, the XM2VTS dataset [23] was chosen. The XM2VTS dataset was created for research into the multimodal identification of human faces. It has also been used for biometrics research into the use of voice for the purposes of liveness verifications [8], [10]. It has been chosen in preference to other possible datasets, as a liveness verification system is used as an addition to other biometric systems, such as face recognition, and by using a dataset used for that task a better understanding of how the proposed system would enhance the security of existing biometric systems is shown.

The XM2VTS dataset is a multi-model dataset comprised of 295 speakers saying the phrases “zero one two three four five six seven eight nine”, “five zero six nine two eight one three seven four” and “Joe took father’s green shoe bench out”. The focus of our experiments is on digit recognition by visual speech recognition so only the digit string phrases have been used. The data is split between training and testing data based on the Lausanne protocol [24]. This protocol divides the dataset into training and test for the training and testing of biometric systems. The protocol specifies two distinct configurations for the dataset. We use Configuration II of the protocol as the starting point for selecting our training and test data. Specifically, we selected the videos from the 70 speakers in the test partition as our test data for the comparison of the various deep neural networks. The videos from the speakers that are not in the test set are used when training the recognizer. As none of the videos from the speakers present in the training set were used for our experiments the results reported indicate how the system would perform under speaker independent conditions and are therefore a good indication of how the system would perform when presented with data from new speakers, as would occur when such a system would be deployed for practical use. The videos used to evaluate the performance of the different neural networks all contain twenty-digit strings. All of the training data consists of the same phrase being spoken by a number of different speakers. The phrase being spoken in the videos is “zero one two three four five six seven eight nine five zero six nine two eight one three seven four”. None of the speakers are seen by the models during training to ensure that the results obtained reflect the performance of a speaker independent system. This is important as it is unlikely that speakers in the training data would also be the users of the system, so the system must be able to perform visual speech recognition on unseen speakers to an acceptable level. All of the videos used have been cropped to only contain the ROI, this may be seen in Figure 4, where images from a number of different speakers are shown.

3.2. Experiments on visual speech recognition models

In order to determine the best neural network to use for this work a number of different neural networks were trained. Prior to this, a GMM/HMM model was trained to provide a baseline to which the performance of the DNN/HMM models could be compared. GMM/HMM model have been widely used in audio-based speech recognition and much of the work to date on visual speech recognition has made use of a GMM/HMM model. As such it provides a good basis upon which to evaluate if improvements in visual speech recognition performance can be achieved with deep learning. The Kaldi speech recognition toolkit was used as the basis of the system, while the is an audio-based speech recognition toolkit it was adapted for use with visual features for our experiments. It supports a number of different deep neural networks making it useful for our experiments. As noted the experiments are conducted using a test set that is comprised of speakers that the system has not seen before as such to perform successfully the deep neural networks we need to have properly leaned to deal with visual speech as overfitting to the training set should cause poor performance on the test set used.

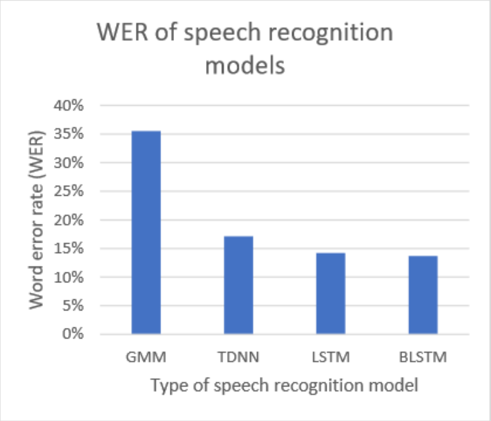


Figure 4. Performance of speech recognition models

The GMM/HMM model achieved a word error rate of 36.5% as shown in Figure 5. Has can be seen from Figure 5 all of the DNN models give significantly better performance than the GMM model. For the TDNN model, a performance of 17.1% WER is achieved showing an improvement over the GMM performance of a greater than 50% reduction in word error rate. With the LSTM model a word error rate of 14.2% showing improvement over both the GMM and TDNN models. The best performance is obtained with the BLSTM model were a performance of 13.6% was obtained. Our experiments have shown that a deep learning-based approach achieves a significant increase in performance over GMM based visual speech recognition. The consideration of a number of different deep neural networks has shown that a recurrent neural network will give better performance than a feedforward network and that a further increase is possible where the deep neural network can make use of both the left and right context.

3.3. The DNN system for liveness verification

From our work on comparing deep neural networks, we selected the BLSTM as the best performing system, for use in the development of a liveness verification system. Further details of the system used for the liveness verification experiments will be discussed in this section. we have employed a hybrid system for performing visual speech recognition for liveness verification. The term hybrid refers to a speech recognition system in which a DNN (deep neural network) and HMM (hidden Markov model) are combined [18]. The DNN is used to provide the posterior probability estimates for the HMM states. The HMM models the long-term dependencies needed to take account of the temporal dimension of speech. For this work, we employed a DNN-HMM trained on DCT features. The use of DNN-HMM recognizers has shown significant improvement in the performance of speech recognition systems over prior approaches [13], [18]. The architecture of the DNN can be seen in Figure 2.

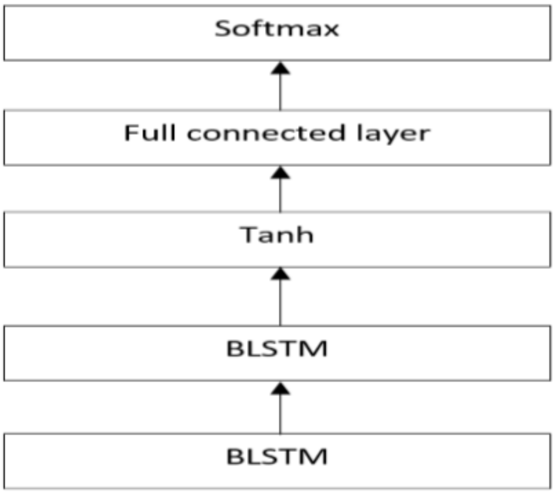


Figure 5. DNN Architecture

Prior to training the DNN a DBN (deep belief network) of stack RBMs (restricted Boltzmann machines) was trained. This process is used to initialize the parameters of the hidden layers in the DNN. This is done via a greedy layer-wise procedure with each RBM trained and then stacked to produce the DBN. The RBM's are trained via approximate stochastic gradient descent. After this pre-training step, the DNN is trained using sMBR (state level minimum Bayes risk) sequence discriminative training as this is suggested as the best criteria for sequence discriminative training in [20], [21]. The output of the network is a pseudo-likelihood of a specific sub-word unit having been spoken give the observed speech.

$$WER = \frac{S + D + I}{N} \tag{6}$$

Where S is the number of substitution errors found in the hypothesis phrase, I is the insertion errors, D is deletions errors and N is the total number of words in the correct transcription. S, D and I are determined through the use of dynamic programming during the calculation of the Levenshtein distance between the correct transcription of the spoken utterance and the hypothesis phrase.

Ideally, when a speaker says the challenge phrase the result of visual speech recognition would be a perfect match, but visual speech recognition is not yet perfect and typically may operate at WERs of between 10% - 40% depending on the user and the quality of the video provided. Therefore, given that a recognition system will operate at a certain average WER, it seems plausible that if a challenge phrase of sufficient length is compared to the output of the recognizer and the Levenshtein distance is within an Acceptable Levenshtein Distance (ALD) threshold then it could be postulated that the challenge phrase was probably spoken as opposed to a random phrase and therefore liveness could be verified. Given this setup, the probability of a successful spoofing attack using a video containing the correct number of random digits can be expressed as in Equation 7.

$$P = \frac{1}{w}^{w-\varepsilon} \cdot \frac{v}{w}^{\varepsilon} \cdot \binom{W}{\varepsilon} \qquad (7)$$

Where P is the probability of a match being found with a challenge phrase containing w words chosen from a vocabulary of v+1 word types and where the system allows ε errors. Taking a specific example, the probability of a random digit string video being used successfully for a spoof attack where w=20 and e=12 is 3.4×10−10. Therefore, while ideally the ε would be kept as low as possible, even where the recognizer is not completely accurate the probability of a successful random spoof attack will be extremely small when the phrase length is sufficiently long.

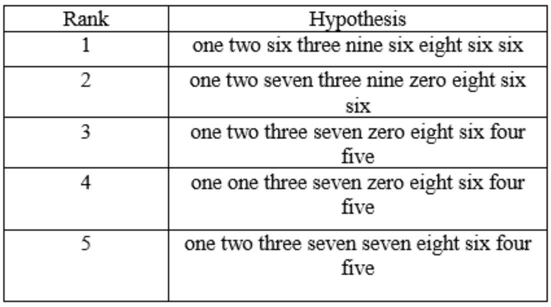
5. Lattice-based Phrase Verification

Aside from the single highest-likelihood hypothesis phrase, it is also possible to generate an N-best list of phrase hypotheses ranked according to their likelihoods from what is known as the recognizer's search lattice. The N-best list typically includes phrases which are plausible slight variations of the highest ranked hypothesis. For example, if a user was challenged to say the following phrase:

"one two seven three nine zero eight six four"

As can be seen in this example, the second-ranked hypothesis contains fewer errors than the best hypothesis and it is not unusual for the correct transcript or a close match to it to be found elsewhere within the N-best list rather than at the very top. The maximum length of an N-best list is primarily determined by the beam width and other pruning parameters during recognition (potentially several thousand hypotheses) but in practice, the correct phrase is generally found close to the top and in our experiments always within the top 50 phrases. Therefore, we allow the system to perform phrase verification with each of the hypothesis phrases in the top 100 phrases and if any of the phrases are verified based on the search of the Nbest list, then the liveness is determined to be positive. This potentially allows the ALD threshold to be reduced slightly leading to a reduction in False Rejection Errors (FRR).

Table 1. 5-best list



Rank	Hypothesis
1	one two six three nine six eight six six
2	one two seven three nine zero eight six six
3	one two three seven zero eight six four five
4	one one three seven zero eight six four five
5	one two three seven seven eight six four five

then the resulting 5-best list might be as in **Error! Reference source not found..**

6. XM2VTS segmentation

In order to allow for experiments to be performed on videos containing shorter phrase lengths, it was necessary to segment the XM2VTS dataset. The two 10-digit sequences were combined within one video to give the 20-digit phrase "zero one two three four five six seven eight nine five zero six nine two eight one three seven four". Only the 20-digit videos were used during training of the recognizer. Using this model, a word accuracy of 86.3% was obtained using the 20-words videos. To allow for investigation into the effect of varying the length of challenge phrases, we segmented the videos in the test set to generate new videos from the test data which contained digits strings of 6, 10 and 15 digits. This was achieved by segmenting the 20-digit videos into videos containing shorter phrases based upon word boundaries for each digit in the video. These were obtained by performing forced alignment of the audio from the videos using a highly accurate (99% word accuracy) audio-based speech recognizer. A variety of phrases were generated using these boundaries by moving a window of size w one word at a time over the 20-digit phrase.

As a result of generating the videos based on this approach, it was possible to expand the number of phrases that were used in our experiments. The variety of phrases can be seen by looking at the first few 10-digit videos generated from the original 20-digit videos were "zero one two three four five six seven eight nine", "one two three four five six seven eight nine five", "two three four five six seven eight nine five zero" etc. While running the experiments, each video was tested as a possible spoof attack case and as a valid user test. The spoof attacks were set up as 1000 random challenge phrases of the correct length containing digit strings that did not match the actual content of the video were created. This simulates the possibility of an attack where the attacker poses a video of the correct user saying a phrase different to the one the system prompts the user to say.

7. Liveness Verification Experiments

Experiments were conducted using the visual speech recognizer on videos containing different length phrases. For practical use, a shorter phrase is preferable as it would take less time for a user to say, however, a longer phrase might be desirable where a stronger level of security is required.

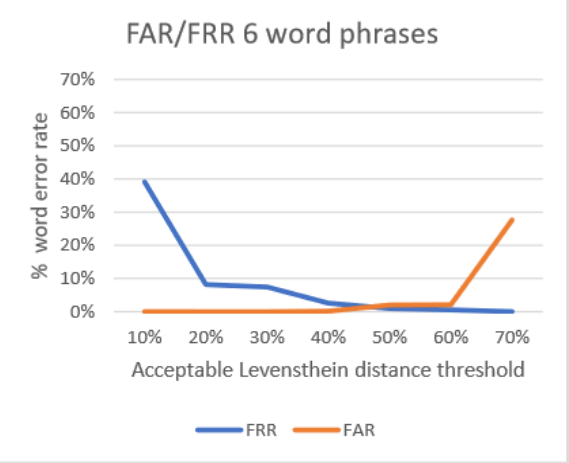


Figure 6. FAR/FRR for 6-word phrases

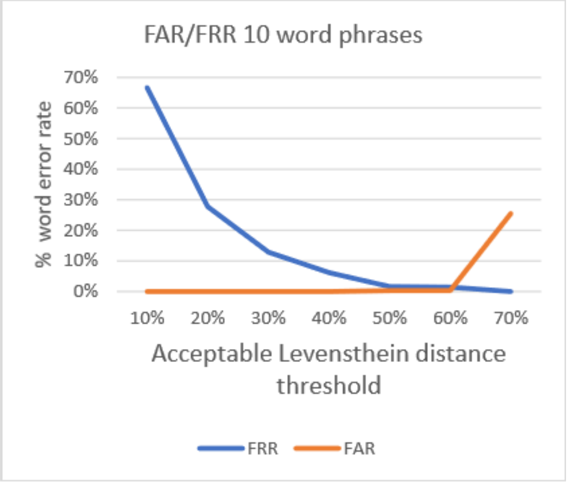


Figure 7. FAR/FRR for 10-word phrases

The average durations for the videos of 6, 10, 15 and 20 words in length were 2, 4, 6 and 8 seconds respectively. The results of the experiments can be seen in charts 1-4.

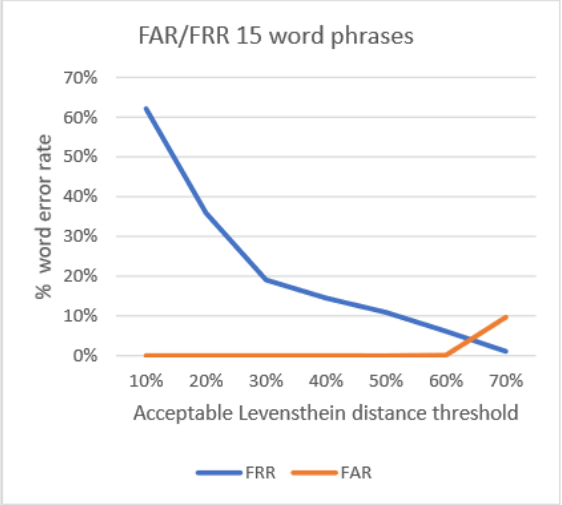


Figure 8. FAR/FRR for 15-word phrases

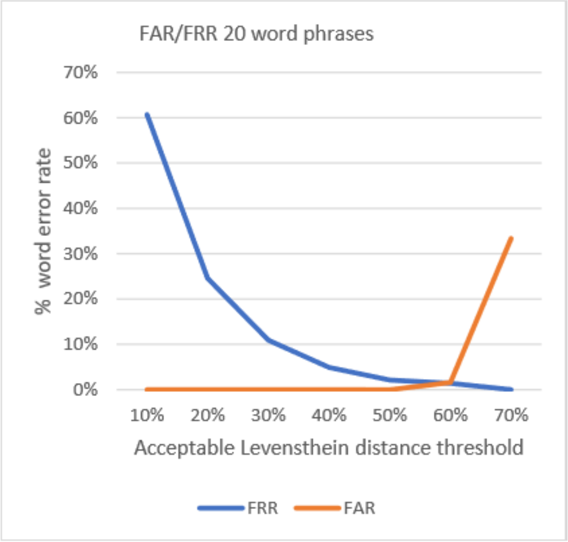


Figure 9. FAR/FRR for 20-word phrases

These charts show the FRR and FAR (false acceptance rate) when the ALD threshold is set to different values. It is shown that the FRR stays below one percent even with ALD thresholds as high as 40%. As can be seen in Figure 7 there are two lines, one indicates the false rejection rate of the system when there are 6-word phrases and the other is the false rejection rate. When the ALD threshold is applied at 20%, 8 % of real phrases would be considered imposters. From Figure 7 it can be seen that where the ALD threshold is increased the number of false rejections is reduced while increasing the number of false acceptances. It should also be noted that the false acceptance rate does not increase above 1% until the threshold is set above 40% for the six-word phrases. From the other Figures, it can be seen that where a longer phrase length is used the number of false acceptances does not rise above 1% until a much higher ALD threshold is used. As the phrases get shorter the impact of error the recognition becomes greater on the overall performance of the liveness verification system. There is, therefore, a trade-off to be made between shorter phrase which is more convenient for the user and longer phrase lengths that provide a greater level of security. From our experiments, a ten-word phrase would appear to be the best choice between a shorter phrase length to enhance usability and a longer phrase that would make the system less likely to allow false acceptance.

7.1. Future work

In this work, the use of deep learning based visual speech recognition as the basis for challenge-based liveness verification has been investigated. The performance of the system on a variety of phrase lengths has been shown and the appropriate ALD thresholds for the different phrase lengths are indicated. Future work will look at improving the performance of the visual speech recognition system and how to make it more robust to noise that such system would encounter when used in real-world conditions. The performance of the system shown in this paper would make it suitable for use as a real-world application and the identification of potential areas where the variety of conditions encountered in such a setting could result in a degradation in the performance of the system would be useful for guiding further work in this area.

8. Conclusion

In this paper, we resented a novel approach to liveness verification based on visual speech recognition. We identified a need for this based on the use of biometric security systems being deployed on mobile devices and the danger of unauthorized users by passing the security through the use of spoofing attacks. A number of different deep neural networks were investigated for their performance on the visual speech recognition task of recognizing digits. We have shown that a deep learning model outperforms a GMM model as well as producing results that improve upon the state of the art in liveness verification through the use of visual speech recognition. Following that we have demonstrated a novel architecture for ant- spoofing that employs Levenshtein distance. We demonstrated that this phrase verification system has a number of parameters which can be tuned based upon the acceptable false acceptance rate / false rejection rate. We have also shown that it can operate at a high degree of confidence when the phrase length is at least 10 words long and an appropriate ALD threshold is chosen.

9. References

[1] A.K. Jain, A. Ross, S. Prabhakar “An introduction to biometric recognition” IEEE Trans. Circuits Systems Video Technol., 14 (1) (2004), pp. 4–20.

[2] C. Roberts, “Biometric attack vectors and defenses,” Computers and Security, vol. 26, no. 1, pp. 14–25, 2007.

[3] B. Biggio, Z. Akhtar, G. Fumera, G. L. Marcialis, F. Roli, "Security evaluation of biometric authentication systems under real spoofing attacks", IET Biometrics, vol. 1, no. 1, pp. 11-24, Mar. 2012.

[4] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition", IEEE Transactions on Multimedia, vol. 2, no. 3, pp. 141-151, 2000.

[5] A. Pass, J. Zhang, and D. Stewart, “An investigation into features for multi-view lipreading,” in Proc. 17th IEEE Int. Conf. Image Process., Sep. 2010, pp. 2417–2420.

[6] G. Chetty and M. Wagner, "Biometric person authentication with liveness detection based on audio-visual fusion", IJBM, vol. 1, no. 4, p. 463, 2009.

[7] M. Alam, M. Bennamoun, R. Togneri, F. Sohel, "A Joint Deep Boltzmann Machine (jDBM) Model for Person Identification Using Mobile Phone Data", Multimedia IEEE Transactions on, vol. 19, pp. 317-326, 2017, ISSN 1520-9210.

[8] N. Eveno, L. Besacier, A speaker independent “liveness” test for audio-visual biometrics, in 9th European Conference on Speech Communication and Technology (Lisbon, 4–8 September 2005.

[9] Kazemi, Vahid, and Josephine Sullivan. "One-millisecond face alignment with an ensemble of regression trees." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1867-1874. 2014.

[10] E Benhaim, H Sahbi, and G Vitte, "Designing relevant features for visual speech recognition, " in Acoustics, Speech, and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013, pp. 2420-242.

[11] Y. LeCun, K. Kavukcuoglu, C. Farabet et al., “Convolutional networks and applications in vision.” in ISCAS, 2010, pp. 253–256.

[12] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, “Deepface: Closing the gap to human-level performance in face verification,” in Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, ser. CVPR ’14. Washington, DC, USA: IEEE Computer Society, 2014, pp. 1701– 1708.

[13] A. Graves, A.-r. Mohamed, and G. Hinton, “Speech recognition with deep recurrent neural networks,” in 2013 IEEE international conference on acoustics, speech and signal processing. IEEE, 2013, pp. 6645–6649.

[14] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, “Natural language processing (almost) from scratch,” Journal of Machine Learning Research, vol. 12, no. Aug, pp. 2493–2537, 2011.

[15] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures", Neural Networks, vol. 18, no. 5-6, pp. 602-610, 2005.

[16] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano and K. Lang, "Phoneme recognition using time-delay neural networks", IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 37, no. 3, pp. 328-339, 1989.

[17] Sugiyama, Masahide & Sawai, Hidefumi & Waibel, A.H. (1991). Review of TDNN (time delay neural network) architectures for speech recognition. 582 - 585 vol.1. 10.1109/ISCAS.1991.176402.

[18] N. Jaitly, P. Nguyen, A. Senior, and V. Vanhoucke, “Application of pre-trained deep neural networks to large vocabulary speech recognition,” in Proc. INTERSPEECH, September 2012.

[19] R. Seymour, D. Stewart, and J. Ming, "Comparison of Image Transform-Based Features for Visual Speech Recognition in Clean and Corrupted Videos", EURASIP Journal on Image and Video Processing, vol. 2008, pp. 1-9, 2008.

[20] B. Kingsbury, “Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling,” in Proc. IEEE, ICASSP, April 2009, pp.3761–3764.

[21] B. Kingsbury, T. N. Sainath, and H. Soltau, “Scalable minimum Bayes risk training of deep neural network acoustic models using distributed Hessian-free optimization,” in Proc. INTERSPEECH, September 2012.

[22] K. Zechner and A. Waibel, “Minimizing word error rate in textual summaries of spoken language”, Proceedings of the 1st North American chapter of the Association for Computational Linguistics Conference, p.186-193, April 29-May 04, 2000.

[23] K. Messer, J. Matas, J. Kittler, and J. Luettin, “Xm2vtsdb: The Extended M2VTS Database,” Proc. Second Int’l Conf. Audio- and Video-Based Biometric Person Authentication (AVBPA ’99), pp. 72-77, 1999.

[24] J. Luettin and G. Maître, “Evaluation Protocol for the XM2FDB Database (Lausanne Protocol)”, Communication 98-05, IDIAP, Martigny, Switzerland, 1998.