Investigation into DCT Feature Selection for Visual Lip-Based Biometric Authentication

Investigation into DCT Feature Selection for Visual Lip-Based Biometric Authentication

C Wright, D Stewart, P Miller, F Campbell-West

Centre for Secure Information Technologies (CSIT)
Queen's University Belfast
{cmclarnon03, Dw.Stewart, p.miller, f.h.campbellwest} @qub.ac.uk

Abstract

This paper investigated using lip movements as a behavioural biometric for person authentication. The system was trained, evaluated and tested using the XM2VTS dataset, following the Lausanne Protocol configuration II. Features were selected from the DCT coefficients of the greyscale lip image. This paper investigated the number of DCT coefficients selected, the selection process, and static and dynamic feature combinations. Using a Gaussian Mixture Model - Universal Background Model framework an Equal Error Rate of 2.20% was achieved during evaluation and on an unseen test set a False Acceptance Rate of 1.7% and False Rejection Rate of 3.0% was achieved. This compares favourably with face authentication results on the same dataset whilst not being susceptible to spoofing attacks.

Keywords: Authentication, Biometrics, GMM-UBM, XM2VTS, DCT

1 Introduction

It is widely recognised that passwords are not enough to use as a sole means of authentication, which has been made apparent by many high profile hacking cases. This has resulted in biometric authentication becoming increasingly popular. Physiological-based biometric systems have been incorporated into the most common mobile platforms, i.e. Android's Face unlock and iPhones fingerprint scanner, and have been hacked using replay attacks and spoofing [Racoma, 2012], [Kleinman, 2014]. Behavioural biometrics are potentially more difficult to crack but are also more complex to capture, model and authenticate robustly.

In this area 'Speaker Verification' is acknowledged as the ability to authenticate a person's claimed identity from their voice [Campbell, 1997]. Gaussian Mixture Model–Universal Background (GMM-UBM) systems are commonly used with speaker verification systems, [Hautamäki et al., 2015]. The set up involves using creating a GMM to model each individual, and another large GMM that represents the whole population – the UBM. When authenticating a person's claimed identity, a likelihood is calculated with respect to their individual model and another with respect to the UBM. A resulting score can be calculated using these likelihoods.

[Cetingul et al., 2006] researched lip motion features for speaker and speech recognition using Hidden Markov Models (HMMs). Features evaluated include dense motion features within a bounding box around the lips, and features created from lip shape (contours) and motion. The MVGL-AVD database consisting of 50 individuals was used. The best recorded result for speaker recognition was found during the cross validation of the system using motion features with an Equal Error Rate (EER) of 5.2%.

[Faraj and Bigun, 2006] investigated a combination of audio and visual features from the lips for person authentication. Experiments used Gaussian Mixture Models (GMM), the XM2VTS database and the Lausanne Protocol, configuration I. Results reported an EER of 22% on visual features alone.

Whilst lip features have shown promise in previous published work there has been no accompanying comparative results for other modalities on the same dataset, e.g. face. This paper investigates using the GMM-UBM framework to model lip movements as a behavioural biometric. The XM2VTS dataset and the Lausanne

 ISBN 978-0-9934207-0-2

Protocol, configuration II was followed [Luettin and Maître, 1998] in an attempt to rigorously benchmark this system and allow for comparison. Discrete Cosine Transform (DCT) coefficients of greyscale lip images were used as visual lip features. As with existing speaker verification systems a likelihood value was calculated using both models of the claimed individual and the UBM.

2 DCT Features

The DCT coefficients were chosen as they can capture lip appearance in a compact form. Investigating gender recognition [Stewart et al., 2013] used DCT coefficients as a feature, the positive results demonstrate that they captured speaker specific appearance and dynamics. However there has been no rigorous investigation of DCT-based features for speaker authentication which is the aim of this work.

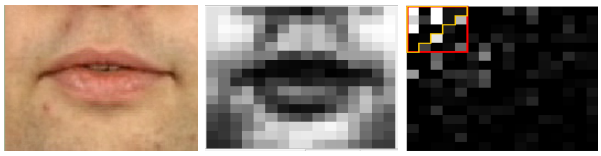

Figure 1: Individual 0, Session 1, video 1, frame 1. From left to right: Pre-processed cropped image, Grey scale resized image, DCT coefficients of the frame with both square and triangular mask

Although we are aware that DCT coefficients can be useful for modelling lip appearance, we do not know how many DCT coefficients are required to effectively model identities and we do not know the relative utility of static DCT coefficients compared to their derivative features. Furthermore when extracting DCT coefficients as features from the full DCT coefficient matrix, two common masks can be applied, square or triangular. It has been shown for lip-based speech recognition that a triangular mask offers better performance, [Stewart et al., 2013], and we will seek to establish if the same is true when modelling identities.

Figure 1 shows how the video data is processed during these experiments. The first image shows the cropped Region Of Interest (ROI). Next each frame is converted to grey scale. After histogram equalisation the frame is resized to a 16 by 16 pixel image, this is shown in the second image in figure 1. The third image shows a visual representation of the 2D DCT coefficients of the frame overlaid with both square and triangular masks.

The masks were used to extract the required number of DCT coefficients, k. The extracted DCT coefficients of the j^{th} frame are represented by $\mathbf{x}_j$, a $k \times 1$ column vector. The frames are stacked to make a $k \times n$ matrix, where n is the number of frames in a video, $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$. Normalisation was used to reduce the effects of inter-session variability using:

$$\overline{x}_j^i = \frac{(x_j^i - \mu^i)}{\sigma^i}, \tag{1}$$

where μ^i is the mean of the i^{th} DCT coefficient across all frames, similarly σ^i is the standard deviation. The resulting normalised feature for the entire video input is therefore: $\overline{\mathbf{X}} = \{\overline{\mathbf{x}}_1, \overline{\mathbf{x}}_2, ..., \overline{\mathbf{x}}_n\}$. The same steps were used to prepare the videos for all steps in the training, evaluation and testing.

3 GMM Modelling

GMMs were used to represent both the UBM and the individual models. During an attempted login the system will test the input against the claimed individual model and against the UBM. Using $\overline{\mathbf{X}}$, we want to compute how alike it is to the features that created a model, λ. The likelihood $p(\overline{\mathbf{X}}|\lambda)$, is calculated using:

$$p(\overline{\mathbf{X}}|\lambda) = \prod_{j=1}^{n} \sum_{i=1}^{M} \omega_i \; p_i\left(\overline{\mathbf{x}}, \; j\right) \tag{2}$$

where M is the number of unimodal gaussians, n is the number of frames, $\overline{\mathbf{x}}_j$ is the normalised j^{th} frame, and the mixture weights, ω_i must satisfy the constraint $\sum_{i=1}^{M} \omega_i = 1$. During training the objective is to maximise $p(\overline{\mathbf{X}}|\lambda)$, where $p_i\left(\overline{\mathbf{x}}_j\right)$ is the likelihood of the j^{th} frame to the i^{th} unimodal gaussian, and the i^{th} unimodal gaussian is parameterised by a mean, $\boldsymbol{\mu}_i$, and a covariance matrix $\boldsymbol{\Sigma}_i$ as described in 3:

 ISBN 978-0-9934207-0-2

$$p_i\left(\overline{\mathbf{x}}_j\right) = \frac{1}{(2\pi)^{\frac{k}{2}}|\boldsymbol{\Sigma}_i|^{\frac{1}{2}}} exp\left\{-\frac{1}{2}\left(\overline{\mathbf{x}}_j - \boldsymbol{\mu}_i\right)^T(\boldsymbol{\Sigma}_i)^{-1}\left(\overline{\mathbf{x}}_j - \boldsymbol{\mu}_i\right)\right\} \tag{3}$$

4 Classification

We want to compute the likelihood that the feature extracted from the video input, $\overline{\mathbf{X}}$, was generated by the claimed identity, and the likelihood that the feature was not generated by the claimed identity. If we denote the likelihood of $\overline{\mathbf{X}}$ being generated by the claimed identity as $p(\overline{\mathbf{X}}\,|\lambda_{hyp})$, where λ_{hyp} represents the mean vector and covariance matrix parameters of the hypothesised Gaussian model, and the likelihood of $\overline{\mathbf{X}}$ of being generated by anybody else as $p(\overline{\mathbf{X}}|\lambda_{UBM})$, where λ_{UBM} represents the mean vector and covariance matrix parameters of the UBM Gaussian model. Then a log-likelihood ratio can be calculated using equation 4:

$$\Lambda(\overline{\mathbf{X}}) = \log\ p(\overline{\mathbf{X}}\,|\lambda_{hyp}) - \log p(\overline{\mathbf{X}}|\lambda_{UBM}) \tag{4}$$

The log-likelihood generated from equation 4 can then be tested against the threshold and the identity accepted or rejected as shown in the modular diagram in figure 2.

5 Experiments

The aim of these experiments was to investigate the feature representation that produced the lowest EER when varying:

- The mask type used to select the number of DCT coefficients. The right most image in figure 1 shows both triangular and square masks.
- The number of DCT coefficients. Square masks were tested from a 3 by 3 mask producing a feature vector containing 9 DCT coefficients to a 7 by 7 mask producing 49 DCT coefficients. For the triangular masks the range went from a 4 by 4 mask producing 10 DCT coefficients to a 9 by 9 mask producing 45 DCT coefficients.
- The 'type' of feature, ie. Static / Dynamic. This work looked into the performance of static, dynamic and combinations of both to help find the optimum feature representation.

The dynamic features included the first and second order derivatives of the static DCT coefficients with respect to time, known as delta, Δ, and deltadelta, $\Delta\Delta$, features. After testing the features separately, all combinations of the 3 features were tested. The features are combined by concatenating the feature vectors. For example if 15 DCT coefficients had been selected, when combining static and Δ, the $\Delta\Delta$, DCT coefficients were concatenated after the static making a total of 30 DCT coefficients.

5.1 System Overview

Figure 2 shows a modular diagram showing how the system was used during testing. After a video was read in the features were extracted and normalised as described in section 2. The features are used to get a log-likelihood from the claimed GMM and the UBM as described in section 3, and a log-likelihood ratio was obtained using equation 4. The ratio log-likelihood will be either accepted or rejected based on the threshold set during the evaluation stage. This is how the system would be used in deployment.

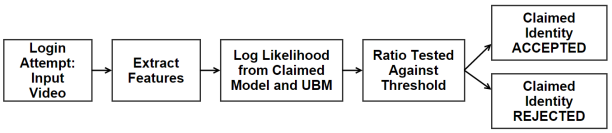

Figure 2: General Modular Diagram of the System

5.2 Database and Protocol

Experiments were carried out on the XM2VTS [Messer et al., 1999]. The XM2VTS is a large audio-visual database containing video recordings of 295 individuals during 4 sessions. Each session contains 2 videos per person and the sessions were recorded over 4 months. For these experiments only digit sequences spoken during recording sessions were used. Pre-processed video data was also used as the audio was removed and the video was cropped to only contain mouth region, for preprocessing steps see [Seymour et al., 2008].

For these experiments the Lausanne Protocol [Luettin and Maître, 1998], configuration II was strictly followed. The Lausanne Protocol is a closed-set verification protocol [Bourlai et al., 2005], because the population of clients does not change the system does not need to account for new users in the evaluation and testing. As shown in figure 3, the protocol divides the dataset into Training, Evaluation and Test data.

- The Training data consists of video from the first 2 sessions for 200 individuals.
- The Evaluation data consists of video data from the third session for the same 200 individuals. Plus all video data for a separate 25 individuals not used in training. See section 5.4 for more details on how these videos are used to represent returning clients and imposters.
- The Test data is made up of the 2 videos from the fourth session of the 200 individuals used for training, plus all video data available for a separate 70 individuals not used in the training.

Data from 3 individuals was removed from our tests as some vidos were found to be corrupt. The IDs of the individuals removed are 342, 272 and 313. Figure 3 shows the number of individuals used in these experiments after the corrupt videos were removed.

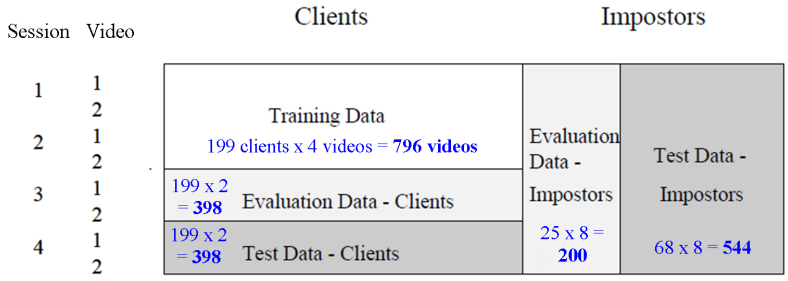

Figure 3: Partitioning of the XM2VTS database according to the protocol Configuration II

5.3 Training

The UBM was trained with all the designated Training data as specified in figure 3, 796 videos from 4 individuals. General guidelines for unconstrained speech suggest 512-2048 Mixtures for the UBM, where lower-order mixtures are more common with constrained speech such as digits and fixed vocabulary [Bimbot et al., 2004]. For the experiments in this work all UBMs were trained with 256 mixtures as all video data contains digits. Individual GMMs were created for each of the 199 individuals in the training data and each model was created using 32 mixtures, likewise 32 mixtures has been used in speaker recognition for individual models, [Stewart et al., 2013].

5.4 Evaluation

Evaluation was carried out in order to select a threshold before running the system on the unseen Test data, using the Evaluation data specified in figure 3. All 598 videos were tested against all 199 individual models, producing $598 \times 199 = 118,604$ attempted logins, with 398 registered user attempts and 118,206 imposter attempts.

Kevin Murphy's toolbox [Murphy, 2001] was used to create the GMM's and retrieve the log-likelihoods. System performance on the Evaluation data was measured by calculating the False Acceptance Rate (FAR- Imposters who can login as another) and the False Rejection Rate (FRR - individuals who cannot in as themselves) and using this to calculate the EER. The EER is the point when the FAR = FRR. The threshold for this system was set to the EER.

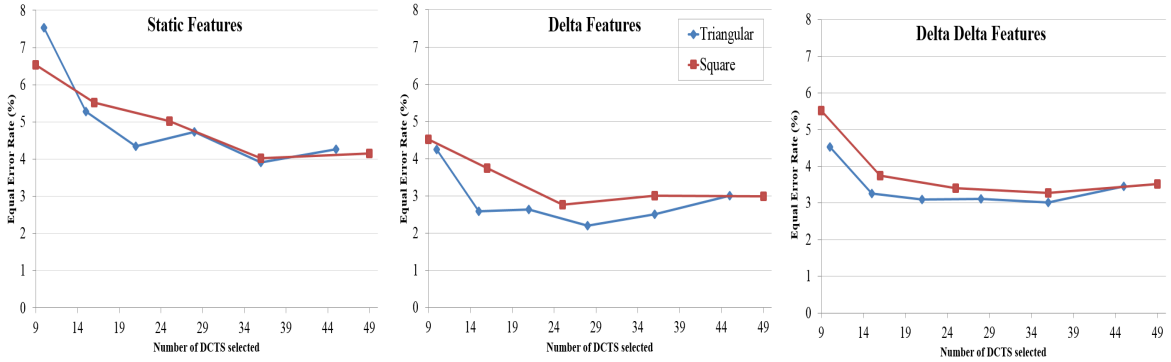

Figure 4: Evaluation results for Static, Δ, ΔΔ features.

The graphs in figure 4 show the EER against the number of DCT coefficients. The graphs show the static, Δ and ΔΔ features. It can be seen that the triangular feature selection outperforms the square feature selection. All 3 graphs show that as the number of DCT coefficients increases the EER is reduced, and it appears that no additional information is gained by using more than 28 DCT coefficients. It can also be seen that the Static features produce the highest EER, and the Δ features produced the lowest EER.

	No. DCT coefficients			
Features	15	21	28	36
Static	5.28	4.34	4.73	3.91
Δ	2.59	2.63	**2.20**	2.51
ΔΔ	3.25	3.09	3.14	3.02
Static & Δ	3.59	3.27	3.27	3.52
Static & ΔΔ	4.02	3.98	4.02	3.52
Δ & ΔΔ	3.10	3.52	2.94	3.27
Static, Δ & ΔΔ	3.02	3.69	3.52	3.94

Table 1: Equal Error Rate (%) on Evaluation set: Showing highest performing number of DCT coefficients, selected used a triangular mask.

Following this, experiments were then run to test combinations of static, Δ and ΔΔ features using triangular feature selection and 15, 21, 28 and 36 DCT coefficients. Combining the static and dynamic information did not appear to add additional information to the features as the Δ alone produced the lowest EER, 2.20%. Results can be seen in table 1.

5.5 Testing

From the evaluation, the set up producing the lowest EER was found to be triangular features using 28 Δ DCT coefficients. The optimum threshold for this system was then calculated based on the EER and applied for testing the unseen data. Before running the unseen Test data on the system using the threshold calculated, practice dictates that the system is retrained using both the Training and Evaluation data [Hastie et al., 2009]. These experiments investigated both this practice and running the unseen Test data on the system without retraining. In theory a system would be trained with all available data before deployment and a threshold calculated based on the data it was trained on. If we retrain the models with the Evaluation data it would be expected the threshold calculated in the evaluation would no longer be optimum therefore the unseen data would be expected to not perform as well.

The 942 videos were tested against all the 199 individual models. This produced $942 \times 199 = 187,458$ attempted logins, with 398 registered user attempted logins and 187,060 imposter attacks. The results for this set up can seen in the top row in table 2.

	FRR	FAR
Models Not Retrained	3.02%	1.68%
Models Retrained	1.76%	4.21%

Table 2: False Rejection Rate (FRR) & False Acceptance Rate (FAR) for unseen test data.

Note the performance (in terms of both FAR and FRR) from evaluation for these same models with the same operating threshold was 2.20%. As seen in table 2, a FRR of 3.02% and FAR of 1.67% was obtained.

On the bottom row of table 2 we can see the results for the Test data after evaluation, when the models have been retrained to contain all the Training and Evaluation data. The table shows a FRR of 1.76% and a FAR of 4.21%.

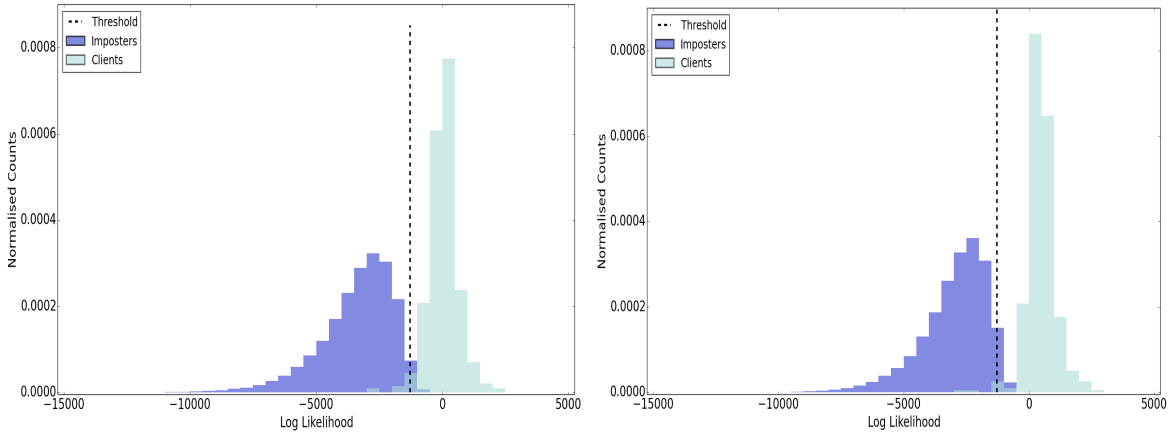

Figure 5: Histogram showing the Client & Imposter Log Likelihoods. Left to right: Not retrained, Retrained

The image on the left in figure 5 shows a histogram of the normalised log-likelihoods of the Test data and the threshold is marked on with a dashed line. There is small overlap in clients and imposters around the threshold. We can see that no matter what threshold was chosen in this experimental setup there will always be FAR or FRR, but the chosen threshold does appear to minimise both the FAR and FRR for unseen data. The image on the right in the figure shows the normalised log-likelihoods for the retrained setup. By comparing the histograms in figure 5 we can see how the threshold set for models trained with less data is no longer optimum with increased training data. This means that if the model is retrained with more data a new threshold should be calculated as in the evaluation stage of these experiments. The overlap of imposters and clients also appears to have reduced in this histogram, indicating that the system improves with more Training data.

As the threshold was set during evaluation, the top row in table 2 shows the true results on unseen data after training and evaluation. These results show even with limited Training data the system successfully authenticated 96.98% of registered clients and successfully prevented access to 98.32% of the imposters.

These results compare very favourably with previous lip based authentication results even though some were on much smaller datasets, [Cetingul et al., 2006]. On the Evaluation data we achieved an EER of 2.2%, producing a predicted performance of 97.8%. This is an improvement on [Cetingul et al., 2006] et al who recorded an error rate of 5.2%. Faraj et al [Faraj and Bigun, 2006] recorded a performance of 78% for the visual features on their own. The EER was used to calculate a threshold which was used to test unseen Test data, this gives a more accurate result on how the system would work in a deployment scenario. These tests produced a FAR of 1.7% and a FRR of 3.0%.

The DCT-based features compare well with results recorded in [Bhattacharjee and Sarmah, 2012], where a 4.55% EER was achieved with a GMM-UBM system and audio features for authentication.

The performance of the system using these features also compares very well with the face recognition system by [Brady et al., 2007] who recorded an error rate of 2.5% on the Evaluation data using the same database and protocol, and the facial recognition system presented by IDIAP in [Messer et al., 2003]. IDIAP used a GMM system and DCT features of the full face image and achieved an EER of 2.45% on the Evaluation data

and on the Test set a FAR and FRR of 1.35% and 0.75% respectively.

Upon further analysis of the specific test cases which caused the FRR errors to occur. The 3.02% of FRR errors equated to 12 attempted logins from only 9 individuals from the 199 registered individuals in the system. Of these, only 3 individuals failed to be authenticated as themselves on both of their test videos. Therefore only 1.5% of individuals could not be authenticated successfully if at least two attempts were considered.

Figure 6 illustrates the data for the 3 problematic individuals. Upon close inspection, the most obvious reason for individual 79 not being authenticated appears to be inconsistent registration of the lip ROI which led to slight rotation of the Test dataset frames compared to the Training frames. Similarly, it is inconsistent ROI extraction which appears to have caused the error for individual 264. In this case the individuals facial hair may have caused the poor lip tracking. For individual 191 the error appears to be caused by a significant change in facial hair prior to the test.

Figure 6: From top down, individual 79, 191, 264. From left to right: Frames 1-4 are from each video included in model, Frames 5-6 are from each video that failed to login

6 Conclusion

This work provided a rigorous investigation of the effectiveness of DCT-based features for modelling speaker's lips within a GMM-UBM verification framework. In particular, we investigated the performance of different numbers of DCT coefficients, different selection of masks and different DCT-based feature types. The types included the static DCT coefficients and their first and second order derivatives known as Δ and $\Delta\Delta$ features. The largest available dataset for such experiments was used, including 292 individuals, namely the XM2VTS database along with the robust Lausanne Protocol configuration II. We showed for the first time that:

- Δ features produced the best feature representation over static, $\Delta\Delta$ and multiple combinations
- 28 DCT coefficients were found to be optimal for the feature
- Triangular mask used in feature selection is better than a square mask

On the Evaluation set an EER of 2.2% was obtained, producing a predicted performance of 97.8%. The EER was used to calculate a threshold which was used to test unseen data, this gives amore accurate result on how the system would work in a deployment scenario. These tests produced a FAR of 1.7% and a FRR of 3.0%.

These results compare very favourably with previous works in verification and authentication using alternative features and models, and compared to facial recognition systems using the same database and protocol.

Our analysis of the errors indicates that the system performance can be affected by poor and inconsistent lip ROI tracking. We will be investigating this further in our future work.

References

[Bhattacharjee and Sarmah, 2012] Bhattacharjee, U. and Sarmah, K. (2012). Gmm-ubm based speaker verification in multilingual environments. *Internation Journal of Computer Science*.

[Bimbot et al., 2004] Bimbot, F., Bonastre, J.-F., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., Merlin, T., Ortega-García, J., Petrovska-Delacrétaz, D., and Reynolds, D. A. (2004). A tutorial on text-independent speaker verification. *EURASIP J. Appl. Signal Process.*, 2004:430–451.

[Bourlai et al., 2005] Bourlai, T., Messer, K., and Kittler, J. (2005). Scenario based performance optimisation in face verification using smart cards. In *Audio-and Video-Based Biometric Person Authentication*, pages 289–300. Springer.

[Brady et al., 2007] Brady, K., Brandstein, M., Quatieri, T., and Dunn, B. (2007). An evaluation of audio-visual person recognition on the xm2vts corpus using the lausanne protocols. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 4, pages IV–237. IEEE.

[Campbell, 1997] Campbell, J.P., J. (1997). Speaker recognition: a tutorial. *Proceedings of the IEEE*, 85(9):1437–1462.

[Cetingul et al., 2006] Cetingul, H. E., Yemez, Y., Erzin, E., and Tekalp, A. M. (2006). Discriminative analysis of lip motion features for speaker identification and speech-reading. *Trans. Img. Proc.*, 15(10):2879–2891.

[Faraj and Bigun, 2006] Faraj, M. and Bigun, J. (2006). Motion features from lip movement for person authentication. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 3, pages 1059–1062.

[Hastie et al., 2009] Hastie, T. J., Tibshirani, R. J., and Friedman, J. H. (2009). *The elements of statistical learning : data mining, inference, and prediction*. Springer series in statistics. Springer, New York. Autres impressions : 2011 (corr.), 2013 (7e corr.).

[Hautamäki et al., 2015] Hautamäki, R. G., Kinnunen, T., Hautamäki, V., and Laukkanenn, A.-M. (2015). Automatic versus human speaker verification: The case of voice mimicry. *Speech Communication*.

[Kleinman, 2014] Kleinman, Z. (2014). Politician's fingerprint 'cloned from photos' by hacker. http://www.bbc.co.uk/news/technology-30623611.

[Luettin and Maître, 1998] Luettin, J. and Maître, G. (1998). Evaluation protocol for the extended M2VTS database (XM2VTSDB). Idiap-Com Idiap-Com-05-1998, IDIAP.

[Messer et al., 2003] Messer, K., Kittler, J., Sadeghi, M., Marcel, S., Marcel, C., Bengio, S., Cardinaux, F., Sanderson, C., Czyz, J., Vandendorpe, L., Srisuk, S., Petrou, M., Kurutach, W., Kadyrov, A., Paredes, R., Kadyrov, E., Kepenekci, B., Tek, F., Akar, G. B., Mavity, N., and Deravi, F. (2003). Face verification competition on the xm2vts database. In *In 4th Int. Conf. Audio and Video Based Biometric Person Authentication*, pages 964–974.

[Messer et al., 1999] Messer, K., Matas, J., Kittler, J., and Jonsson, K. (1999). Xm2vtsdb: The extended m2vts database. In *In Second International Conference on Audio and Video-based Biometric Person Authentication*, pages 72–77.

[Murphy, 2001] Murphy, K. P. (2001). The bayes net toolbox for matlab. In *Computing Science and Statistics*.

[Racoma, 2012] Racoma, J. A. (2012). Android jelly bean face unlock 'liveness' check easily hacked with photo editing. http://www.androidauthority.com/android-jelly-bean-face-unlock-blink-hacking-105556.

[Seymour et al., 2008] Seymour, R., Stewart, D., and Ming, J. (2008). Comparison of image transform-based features for visual speech recognition in clean and corrupted videos. *J. Image Video Process.*, 2008:14:1–14:9.

[Stewart et al., 2013] Stewart, D., Pass, A., and Zhang, J. (2013). Gender classification via lips: Static and dynamic features. *IET Biometrics*, 2(1):28–34.