Audiovisual Synchrony Detection with Optimized Audio Features

Sami Sieranoja, Md Sahidullah, Tomi Kinnunen, Jukka Komulainen, Abdenour Hadid

▶ **To cite this version:**

Sami Sieranoja, Md Sahidullah, Tomi Kinnunen, Jukka Komulainen, Abdenour Hadid. Audiovisual Synchrony Detection with Optimized Audio Features. ICSIP 2018 - 3rd International Conference on Signal and Image Processing, Jul 2018, Shenzhen, China. hal-01889918

HAL Id: hal-01889918
https://inria.hal.science/hal-01889918v1

Submitted on 8 Oct 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Audiovisual Synchrony Detection with Optimized Audio Features

Sami Sieranoja, Md Sahidullah, Tomi Kinnunen
School of Computing
University of Eastern Finland, Joensuu, Finland

Jukka Komulainen, Abdenour Hadid
Center for Machine Vision and Signal Analysis (CMVS)
University of Oulu, Oulu, Finland

***Abstract*—Audiovisual speech synchrony detection is an important part of talking-face verification systems. Prior work has primarily focused on visual features and joint-space models, while standard mel-frequency cepstral coefficients (MFCCs) have been commonly used to present speech. We focus more closely on audio by studying the impact of context window length for delta feature computation and comparing MFCCs with simpler energy-based features in lip-sync detection. We select state-of-the-art hand-crafted lip-sync visual features, space-time auto-correlation of gradients (STACOG), and canonical correlation analysis (CCA), for joint-space modeling. To enhance joint space modeling, we adopt deep CCA (DCCA), a nonlinear extension of CCA. Our results on the XM2VTS data indicate substantially enhanced audiovisual speech synchrony detection, with an equal error rate (EER) of 3.68%. Further analysis reveals that failed lip region localization and beardedness of the subjects constitutes most of the errors. Thus, the lip motion description is the bottleneck, while the use of novel audio features or joint-modeling techniques is unlikely to boost lip-sync detection accuracy further.**

***Keywords*-Audiovisual Synchrony, Presentation Attack Detection, Multimodal Processing, Feature Extraction, Mel-Frequency Cepstral Coefficients (MFCCs).**

I. Introduction

Nowadays, most mobile devices are equipped with a microphone and a front-facing video camera, enabling non-intrusive audiovisual user authentication. Although integrated face and voice modalities can increase recognition accuracy, both are highly vulnerable to *presentation attacks* (spoofing attacks) [1]. For instance, presentation of pre-recorded audio clip (replay) together with a still photograph is enough to circumvent talking face verification relying on late fusion [2]. Multi-biometrics by itself is not inherently robust to spoofing attacks since successful spoofing of just one modality may compromise the entire system [3].

One approach to counter audiovisual presentation attacks is to independently validate face [4] and voice [5] liveness. Another approach is to determine whether the content and timing of the captured audible and visual speech match (see Fig. 1). Such *audiovisual speech synchrony detection* can be performed using both text-independent [2], [6], [7], [8], [9], and text-dependent [10], [11] methods. The former are effective in detecting attacks whereby the attacker uses separate audio and video recordings (or photo) of the target person. These methods, however, are powerless under pre-recorded video replay attacks with synchronized audiovisual speech.

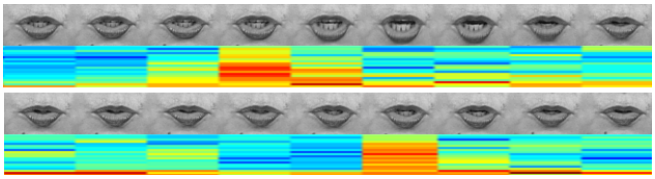


Figure 1. Synchronous lip region frames and audio spectrograms extracted from phrase *Joe took*. Audiovisual synchrony detection seeks to detect discrepancy of the two modalities (an indication of a spoofing attack).

Text-dependent synchrony assessment methods tackle this issue by utilizing *challenge-response* approach by prompting the user a randomly selected sentence [10], [11] (challenge) and then verifies whether the preassigned utterance can be recognized in both modalities within a specific time window (response).

Audiovisual speech synchrony studies largely focus on visual features and joint-space models, leaving an open question whether, and how much, improvement could be obtained by optimization of the audio features. Standard *mel-frequency cepstral coefficients* (MFCCs) are commonly used (with some exceptions [9]). Speech activity detection (SAD) is another audio front-end component whose usefulness has not been directly addressed in prior work. For these reasons, it is difficult to tell what are the bottlenecks in the existing lip-sync detection pipelines, *i.e.* the used audio or visual features, or the joint modeling and synchrony detection of the two modalities.

In this work, we study the use of energy and delta features that we believe to be useful for audio-visual synchrony detection. We fix the visual features to *space-time auto-correlation of gradients* (STACOG) and the joint-space analysis to *canonical correlation analysis* (CCA), that have formed the state-of-the-art in audiovisual speech synchrony detection [7], [12]. To gain further insight into the importance of feature choice versus joint-space modeling, we consider *deep* CCA (DCCA), a nonlinear extension of CCA [13]. Finally, we analyze the misclassified test cases with an aim to *explain* the reasons behind these errors in audio-visual synchrony detection task.

II. Related work

Audiovisual speech synchrony detection studies focus primarily on back-end synchrony measures between audio and video. MFCCs [14] are commonly used to present

speech [7], [15], [16], [17], with a few alternatives such as frame energy [18]. For video, *discrete cosine transform* (DCT) [6], lip measurements [16] and *multi-channel gradient model* (MCGM) [19] are commonly used. In [7], STACOG were found to outperform DCT features in measuring audio and visual speech correlation. Very recently, also *convolutional neural network* (CNN) architectures have been proposed for both audio and visual speech feature extraction [20], [21]. As a back-end, CCA is the default choice [9], with alternatives such as *coinertia analysis* (CoIA) [8], *generalized bimodal linear prediction* [16], *kernel* CCA [19] and *deep neural networks* (DNNs) [17], [20], [21], [22].

III. Methodology

A. Video features

STACOG [23] is a motion feature extraction method that encodes the local auto-correlations of space-time gradients in a video for capturing the local geometric characteristics of moving objects. Originally, STACOG has been successfully used for e.g. hand gesture and human action recognition where it has demonstrated superior performance and computation efficiency over similar methods [23].

Following up [7], [12], we use `dlib`[1] to determine eye and mouth locations in every video frame, following the strategy of [24] to obtain a good approximation of the whole mouth region. The resulting rectangular mouth image is resized to 70×40 pixels from 1584-dimensional STACOG features using default parameters[2]. The features represent dynamics between three consecutive video frames (excluding endpoints).

B. Audio features

MFCCs [14] are computed from short-term discrete Fourier transform (DFT). If we denote the DFT power spectrum of one frame as a vector $\mathbf{s} \in \mathbb{R}^{N/2+1}$, where N is the size of discrete Fourier transform (after zero padding), MFCC vector is $\mathbf{c} = \mathbf{D}\log(\mathbf{H}\mathbf{s})$, where $\mathbf{H} \in \mathbb{R}^{Q\times(N/2+1)}$ is a filterbank matrix containing Q filter responses, $\log(\cdot)$ operates element-wise and $\mathbf{D}$ is a DCT matrix.

Prior to MFCC extraction, we downsample the audio to 8 kHz as most voiced speech energy lies below 4 kHz. Speech activity detection [25] is optionally used to eliminate silence at the beginning and end of utterances. Then 20 MFCCs are extracted using $Q = 20$ filters from 40 ms frames with no frame overlap[3]. To measure local spectral dynamics [26], we append MFCCs with their *delta* features. If $\mathbf{c}_t$ denotes the MFCCs of t^{th} frame, the corresponding delta vector using a time context of $f = 2\tau + 1$ frames is $\Delta\mathbf{c}_t = \sum_{m=-\tau}^{\tau} m\mathbf{c}_{t+m}$. Double deltas, $\Delta^2\mathbf{c}_t$, are computed by applying the same operator to the deltas. In the experiments we study the effect of τ and deltas/double deltas.

[1] http://dlib.net/

[2] https://staff.aist.go.jp/takumi.kobayashi/codes.html#STACOG

[3] This somewhat unconventionally long speech window is chosen to match the 40ms associated with the 25fps video frame rate (1/25fps=40ms).

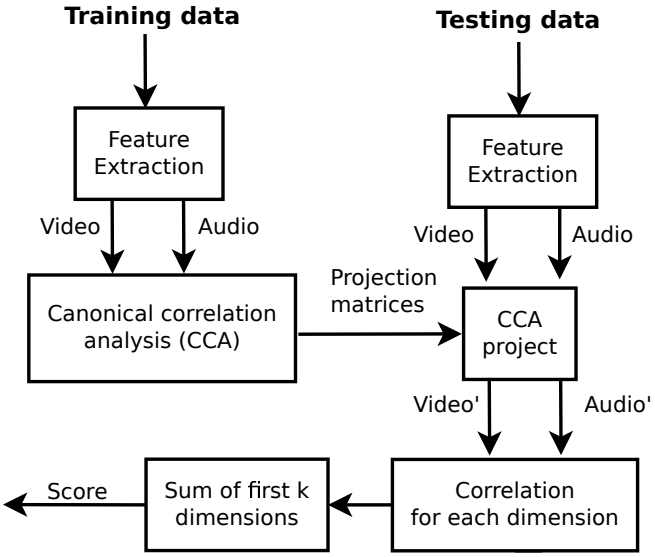


Figure 2. Joint-space model using CCA

C. Joint-space modeling

As STACOG features are produced at 25fps starting from the third video frame, the first two audio feature vectors are excluded to synchronize the two modalities. To measure the degree of synchrony between speech and video features, we use the same CCA [27] joint-modeling and synchrony measures as described in [7], [12]. Given two multidimensional random variables $\boldsymbol{X}$ and $\boldsymbol{Y}$, CCA finds orthogonal linear projections $\boldsymbol{w}_i$ and $\boldsymbol{z}_i$ that are maximally correlated and ordered from highest to lowest correlating:

$$(\boldsymbol{w}_i^*, \boldsymbol{z}_i^*) = \arg\max_{(\boldsymbol{w}_i, \boldsymbol{z}_i)} \text{corr}(\boldsymbol{X}\boldsymbol{w}_i, \boldsymbol{Y}\boldsymbol{z}_i) \qquad (1)$$

Training consists of computing CCA projection matrices ($\boldsymbol{X}$ and $\boldsymbol{Y}$) for both modalities (Fig. 2). In the test phase, these are used to project audio and video features to a common space. For each test video, correlation between audio and video is calculated for each dimension of the projected features. The synchrony score is the sum of the K highest correlating dimensions for all the frames:

$$S_{W,Z}(X,Y) = \frac{1}{K}\sum_{k=1}^{K} |\text{corr}(\boldsymbol{X}\boldsymbol{w}_k, \boldsymbol{Y}\boldsymbol{z}_k)|, \qquad (2)$$

where K is a control parameter.

For more advanced joint-space modeling, we consider *deep* CCA (DCCA) [13]. It works the same way as CCA, but $\boldsymbol{X}\boldsymbol{w}_i$ and $\boldsymbol{Y}\boldsymbol{z}_i$ (in Eq. 1) are replaced with *nonlinear* functions $f_X(X;\boldsymbol{\theta}_w)$ and $f_Y(Y;\boldsymbol{\theta}_z)$, where for $v \in \{x, y\}$, f_v is a deep feedforward neural net of L layers with parameters $\boldsymbol{\theta}_v$ consisting of the weights and biases of all the layers. The network is trained with backpropagation on the gradient of correlation; for more details, see [13].

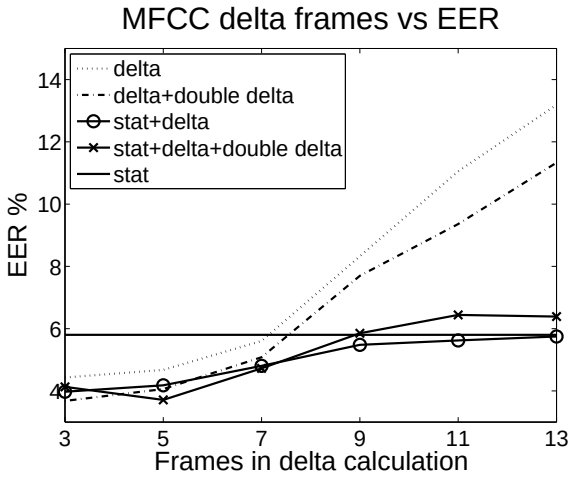


Figure 3. Effect of window size for MFCC delta calculation. Best result is obtained with three frames (equal to STACOG). The baseline system has the same computation pipeline as [12], but with revised MFCC extractor and EER estimation.

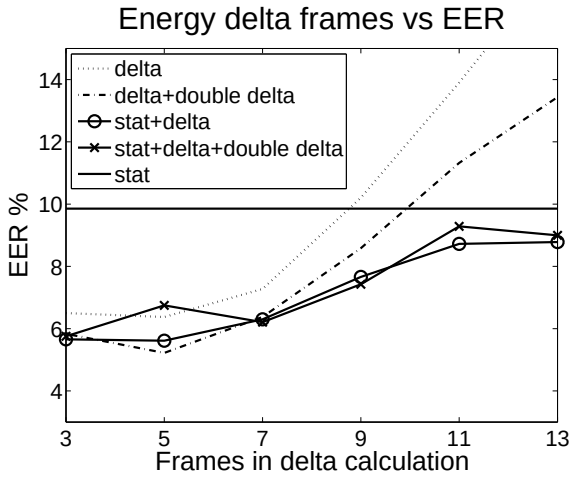


Figure 4. Effect of window size for energy delta calculation. Best result is obtained with five frames.

IV. Experimental setup

A. Dataset and evaluation protocol

To provide comparable results with [7] and [12], we use the XM2VTS dataset [28] to evaluate the lip-sync detectors. XM2VTS contains audiovisual sequences of 295 subjects recorded in four sessions. The audio was recorded at 32 kHz and video at 25 fps. The same sentence ("*Joe took fathers green shoe bench out*") is repeated in all videos. *Logical access attack scenarios* were created by combining audio and video from different sessions of the same person uttering the same sentence, thus high level of synchrony (but not perfect) is present in the audiovisual speech.

B. Evaluation

We measure the performance of audiovisual synchrony detection based on the score $S_{W,Z}(X,Y)$ using *equal error rate* (EER), the operating threshold with equal miss and false alarm rates[4]. We follow exactly the same evaluation protocol as in [7], [12], where the dataset is split into two equal size subject-disjoint halves. The lip-sync detection models are trained on real videos of one group and the resulting model is evaluated on the other group. The process is repeated by alternating the role of the two folds and the reported EER is the average of the two tests. The parameter K is optimized for both folds separately, with optimal K for first fold used for scoring the second fold, and vice versa.

V. Results

A. Effect of audio feature configuration

The results for audio feature optimization with CCA back-end are illustrated in Figs. 3 and 4, correspondingly for MFCCs and energy features (sum of mel filterbank output).

[4]Computed using BOSARIS, https://sites.google.com/site/bosaristoolkit/.

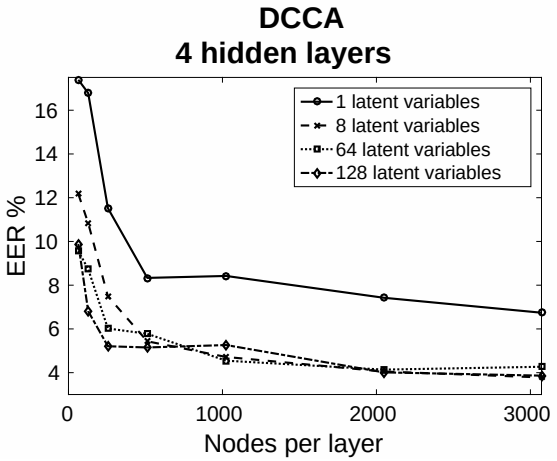
Table I
Audio feature optimization. $f = 2\tau + 1$ is the number of frames (context window size) for delta calculation.

	CCA	DCCA
Feature Configuration	EER (%)	EER (%)
(Baseline) MFCC(f=9)-stat+Δ	5.48	5.17
MFCC-stat	5.80	5.60
MFCC(f=3)-stat+Δ	3.97	3.76
MFCC(f=3)-stat+Δ+Δ^2	4.13	3.56
MFCC(f=3)-Δ+Δ^2	**3.68**	**3.52**
Energy-stat	9.86	12.29
Energy(f=3)-stat+Δ	5.66	7.29
Energy(f=3)-stat+Δ+Δ^2	5.75	6.19
Energy(f=3)-Δ+Δ^2	5.83	6.02

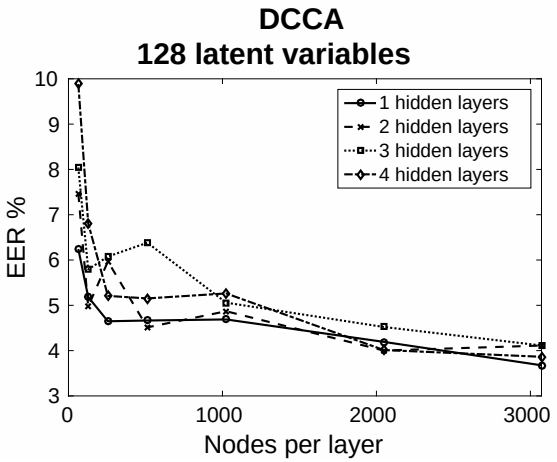
A summary is provided in Table I along with the baseline method [12] in the first row. The best result of EER 3.68% is obtained using MFCCs with deltas and double deltas and three-frame context, though performance with 5-frame context is similar. The trends in Figs. 3 and 4 are consistent — less frames in delta computation is better. The energy features are behind MFCCs (5.22%>3.68%) as one might expect, but provide nevertheless low EERs. This suggest that most of the detection accuracy (even with MFCCs) results from detecting the synchrony of the amount of sound energy and the lip motion. Inclusion of speech activity detection had very little effect (possibly due to short utterances); with MFCCs with deltas, EER changed from 4.07% to 3.97%, and with delta plus double deltas from 3.70% to 3.68%.

B. Deep CCA back-end

The effect of DCCA parameters on EER is shown in Fig. 5. To run this test in reasonable time, we use a fast training setup with a minibatch size of 5000 and 10 training epochs. Keeping audio feature type fixed to MFCC (f=3)-stat+Δ+Δ^2, we vary the number of hidden layers and nodes per layer. The results indicate that more nodes per layer and more latent variables per layer leads to better results. Based on this observation, we fix the DCCA configuration to 128



(a)



(b)

Figure 5. Effect of varying the number of latent variables, layers and nodes per layer in DCCA.

latent variables, five hidden layers, 2048 nodes per layer, sigmoid activations and perform more extensive training with full batch size. The results included in Table I indicate that DCCA performs better than CCA with MFCC features, having the best result of EER 3.52% and also performing better with other MFCC feature combinations.

DCCA performs worse in combination with the energy features. This might be due to a conflict between the high dimensional visual ($D = 1584$) STACOG features and the low dimensional ($D = 1, 2, 3$) energy features leading to overfit. This interpretation is backed by additional experiments with DCCA and energy features where we increased the DCCA regularization parameter from the default 10^{-4} to 10^{-2}. This improved the results for the 1-dimensional energy-stat features from EER 12.29% to 9.14%, and for Energy-stat+Δ+Δ^2 from 6.19% to 5.88%.

C. Error case analysis

The obtained error rates are reasonably low already with classic CCA back-end. To gain further insight into the remaining issues to be solved in audio-visual synchrony detection, we study the errors of CCA with MFCCs with deltas and double deltas (with f=3). We fix the detection threshold to the EER operating point (at 3.68%). The cases that contained synchronous video, but were falsely rejected as asynchronous, 27% contained incorrectly detected mouth positions and 67% of subjects had a beard or moustache — see Fig. 6. In the opposite case of asynchronous video wrongly classified as synchronous, there were no above normal levels of beardedness. In summary, video of a person with a beard or a moustache would be more likely wrongly classified as a spoofing attack. Possible reasons include; (1) as beard covers a part of the mouth, it conceals some lip movements that would be visible on a shaved face; (2) as bearded persons are a minority in the dataset, the amount of training data to model their features might be insufficient.

Figure 6. Sample faces from synchronous videos wrongly classified as asynchronous.

VI. Conclusion

We studied MFCC delta configurations to provide a suitable counterpart for the visual STACOG features, leading to audiovisual speech synchrony detection with an EER 3.68%, surpassing the selected baseline [12]. We additionally used simpler energy measure that yielded an EER of 5.22%, suggesting that most of the recognition accuracy is based on detecting if the amount of speech energy produced is in synchrony with the lip movements. Finally, we found that beardedness of subjects and failed lip region localization explains most errors. Therefore, in conclusion, the lip motion description seems to be the bottleneck, while the use of other audio features or advanced joint-modeling, such as [17], [22], may not increase accuracy further.

Acknowledgments

The project was partially funded by Academy of Finland and the Finnish Foundation for Technology Promotion.

REFERENCES

[1] S. Marcel, M. Nixon, and S. Li, *Handbook of Biometric Anti-Spoofing: Trusted Biometrics Under Spoofing Attacks*. Springer, 2014. 1

[2] H. Bredin and G. Chollet, "Making talking-face authentication robust to deliberate imposture," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2008, pp. 1693–1696. 1

[3] R. Rodrigues, L. Ling, and V. Govindaraju, "Robustness of multimodal biometric fusion methods against spoof attacks," *Journal of Visual Languages and Computing (JVLC)*, vol. 20, no. 3, pp. 169–179, 2009. 1

[4] Z. Boulkenafet, J. Komulainen, and A. Hadid, "Face spoofing detection using colour texture analysis," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 8, pp. 1818–1830, 2016. 1

[5] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Communication*, vol. 66, pp. 130–153, 2015. 1

[6] E. Argones Rúa, H. Bredin, C. García Mateo, G. Chollet, and D. González Jiménez, "Audio-visual speech asynchrony detection using co-inertia analysis and coupled hidden Markov models," *Pattern Analysis and Applications*, vol. 12, no. 3, pp. 271–284, 2009. 1, 2

[7] E. Boutellaa, Z. Boulkenafet, J. Komulainen, and A. Hadid, "Audiovisual synchrony assessment for replay attack detection in talking face biometrics," *Multimedia Tools and Applications*, pp. 1–15, 2016. 1, 2, 3

[8] N. Eveno and L. Besacier, "A speaker independent "liveness" test for audio-visual biometrics," in *INTERSPEECH*, 2005. 1, 2

[9] M. Slaney and M. Covell, "Facesync: A linear operator for measuring synchronization of video facial images and audio tracks," in *Neural Information Processing Systems (NIPS)*, 2000, pp. 814–820. 1, 2

[10] K. Kollreider, H. Fronthaler, M. I. Faraj, and J. Bigun, "Real-time face detection and motion analysis with application in "liveness" assessment," *IEEE Transactions on Information Forensics and Security*, vol. 2, no. 3, pp. 548–558, 2007. 1

[11] A. Melnikov, R. Akhunzyanov, O. Kudashev, and E. Luckyanets, "Audiovisual Liveness Detection," in *International Conference on Image Analysis and Processing (ICIAP)*, 2015. 1

[12] J. Komulainen, I. Anina, J. Holappa, E. Boutellaa, and A. Hadid, "On the robustness of audiovisual liveness detection to visual speech animation," in *IEEE International Conference on Biometrics: Theory, Applications, and Systems (BTAS)*, 2016. 1, 2, 3, 4

[13] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *International Conference on Machine Learning (ICML)*, 2013, pp. 1247–1255. 1, 2

[14] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980. 1, 2

[15] G. Chetty and M. Wagner, "Liveness detection using cross-modal correlations in face-voice person authentication," in *INTERSPEECH*, 2005. 2

[16] K. Kumar, J. Navratil, E. Marcheret, V. Libal, and G. Potamianos, "Robust audio-visual speech synchrony detection by generalized bimodal linear prediction," in *INTERSPEECH*, 2009. 2

[17] A. Aides and H. Aronowitz, "Text-dependent audiovisual synchrony detection for spoofing detection in mobile person recognition," in *INTERSPEECH*, 2016. 2, 4

[18] J. Hershey and J. Movellan, "Audio vision: Using audio-visual synchrony to locate sounds," in *Neural Information Processing Systems (NIPS)*, 1999, pp. 813–819. 2

[19] G. Chetty, "Biometric liveness detection based on cross modal fusion," in *Proc. of 12th International Conference on Information Fusion*, 2009, pp. 2255–2262. 2

[20] J. S. Chung and A. Zisserman, "Out of time: Automated lip sync in the wild," in *Computer Vision – ACCV 2016 Workshops*. Springer, 2017, pp. 251–263. 2

[21] A. Torfi, S. M. Iranmanesh, N. Nasrabadi, and J. Dawson, "3d convolutional neural networks for cross audio-visual matching recognition," *IEEE Access*, vol. 5, pp. 22 081–22 091, 2017. 2

[22] E. Marcheret, G. Potamianos, J. Vopicka, and V. Goel, "Detecting Audio-Visual Synchrony Using Deep Neural Networks," in *INTERSPEECH*, 2015. 2, 4

[23] T. Kobayashi and N. Otsu, "Motion recognition using local auto-correlation of space–time gradients," *Pattern Recognition Letters*, vol. 33, no. 9, pp. 1188–1195, 2012. 2

[24] G. Zhao, M. Pietikäinen, and A. Hadid, "Local spatiotemporal descriptors for visual recognition of spoken phrases," in *ACM International Workshop on Human-centered Multimedia (HCM)*, 2007, pp. 57–66. 2

[25] J. Sohn, N. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, 1999. 2

[26] F. K. Soong and A. E. Rosenberg, "On the use of instantaneous and transitional spectral information in speaker recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 6, pp. 871–879, 1988. 2

[27] D. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural computation*, vol. 16, no. 12, pp. 2639–2664, 2004. 2

[28] K. Messer, "XM2VTSDB: The Extended M2VTS Database," in *AVBPA*, 1999. 3