Learning Multimodal Temporal Representation for Dubbing Detection in Broadcast Media

Nam Le[1,2], Jean-Marc Odobez[1,2]
[1] Idiap Research Institute, Martigny, Switzerland
[2] École Polytechnique Fédéral de Lausanne, Switzerland
{nle, odobez}@idiap.ch

ABSTRACT

Person discovery in the absence of prior identity knowledge requires accurate association of visual and auditory cues. In broadcast data, multimodal analysis faces additional challenges due to narrated voices over muted scenes or dubbing in different languages. To address these challenges, we define and analyze the problem of dubbing detection in broadcast data, which has not been explored before. We propose a method to represent the temporal relationship between the auditory and visual streams. This method consists of canonical correlation analysis to learn a joint multimodal space, and long short term memory (LSTM) networks to model cross-modality temporal dependencies. Our contributions also include the introduction of a newly acquired dataset of face-speech segments from TV data, which we have made publicly available. The proposed method achieves promising performance on this real world dataset as compared to several baselines.

Keywords

Multimodal, Person Diarization, Recurrent Neural Networks.

1. INTRODUCTION

Making large multimedia corpora easily accessible through search, retrieval and fast browsing tools is a crucial task given the daily production of broadcast TV and internet content. As the retrieval of information on people in videos is of high interest for users, research efforts have been devoted to unsupervised segmentation of videos into homogeneous segments according to person identity, like speaker diarization [21, 17, 29], face diarization [5, 35], and audio-visual (AV) person diarization [10, 25, 16, 8]. Combined with names extracted from overlaid text, AV person diarization makes it possible to identify people in videos [9].

Solving the AV person diarization and naming tasks requires associating visual person tracks or overlaid names with auditory voices, which has several difficulties. Firstly, the visible person may not be the current speaker. This issue occurs when anchors or invited speakers are commenting on video footage displaying famous people who might be talking or when several persons appear to be talking in the background. This cannot be solved with existing systems which visually detect talking faces [4, 1, 26, 8] to reinforce the AV association. Secondly, another recurrent issue in international TV is dubbing. The problem is common when an interviewed person, shown in the video, is speaking in a different language than that of the target audience, and is dubbed by a narrator. Such situations are problematic for person diarization, as they often lead to associating a face track (or a face cluster) or an overlaid person name with the wrong voice, potentially creating multiple ambiguities in person diarization and naming (e.g. if the same voice is dubbing several persons).

In this paper, we focus on dubbing detection in broadcast data, which involves modelling the synchrony of audio and lip motion. This task can be used to handle the two issues mentioned above, by detecting *which of the talking persons (if any) actually produces* the audio discourse. Although it is related to several research problems (AV speech recognition, voice over detection, spoofing in AV biometry), to the best or our knowledge, this dubbing problem has not been addressed previously. To initiate further research, we acquired the DW-dubbing corpus comprising 4722 segments of 2 seconds with the corresponding face track and audio. In addition, from a methodological perspective, we propose to exploit the recently revived LSTM networks to model the joint dynamics of synchronized AV segments in a multimodal space obtained via canonical correlation analysis (CCA). Experiments demonstrate the benefit of our method over several baselines. In summary, our contributions are:

- We address for the first time the problem of dubbing detection in broadcast data;
- We propose a method relying on LSTM and multimodal feature extraction, which achieves promising result on this problem;
- We make publicly available a dubbing dataset collected from TV news for future research.

2. RELATED WORK

Dubbing detection shares some similarities to several problems discussed below along with the related works.

Talking faces. Person diarization and naming require matching audio segments with face tracks of talking people. To detect talking people, mean squared intensity differences [4] or motion entropy [1] within mouth regions were typically used, potentially combined with head motion [26]. Such

MM '16, October 15 - 19, 2016, Amsterdam, Netherlands

ISBN 978-1-4503-3603-1/16/10...$15.00

DOI: http://dx.doi.org/10.1145/2964284.2967211

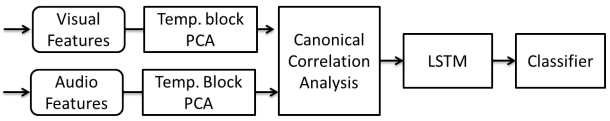

Figure 1: System overview: feature extraction, dimensionality reduction over concatenated feature from block of frames, mutual information extraction via Canonical-Correlation Analysis (CCA), and temporal modelling with LSTM.

visual-based approaches could be further enhanced using multimodal contextual information, like audio segment-face track overlap duration, or face size and position [8]. It is interesting to see that none of the existing systems relied on temporal models for this task. Also, when several persons are seen talking, or in dubbing situations, audio and video need to be jointly considered.

AV speech recognition. Modeling the relationship between audio and visual streams can be traced back to early researches in multimodal speech recognition [30]. Typical examples include coupled hidden Markov Model (CHMM) [22] or asynchronous HMM to model anticipation and retention phenomena. More recently, multimodal deep networks [23] showed good performance. However, the approach did not include temporal models: it relied on neural networks applied to groups of frames, whose outputs were averaged over time. Furthermore, the task is limited to the recognition of simple sounds [27] with little noise (head movements, illumination).

AV biometry and synchronization. The dubbing problem somehow resembles AV spoofing detection, where the task is to detect when visual attackers pretend to match with a playback audio. As an early work, cross-modal fusion with latent semantic analysis or canonical correlation analysis were applied, but only tested attacks composed of a single photograph, potentially animated with simple synthetic movements [3]. To deal with real face tracks with different voices, [32] investigated co-inertia or coupled HMM approaches to detected uncorrelated AV signals or unsynchronization. However, the method was applied to a constrained biometric environment, with specific test sentences used as input to the system. In addition, synchrony detection has also been addressed for speaker location & association [11, 7] in scenes with two people, and has focused more on mutual information modeling than temporal aspects. Mutual information was also shown to be important in monologue detection where a system needs to identify real speakers among sets of confusers [24, 15]. However, temporal modeling using HMM to evaluated likelihood of word utterance given joint AV distribution only yielded limited results [24]. Another related problem is to distinguish narrated vs genuine voices in TV news addressed in [19], where only primitive lip features were used without joint modality space or temporal modeling, and the dataset was very small (40 video clips). In contrast to the above works, our approach utilizes both cross-modal correlation analysis and temporal modeling with state-of-the-art LSTM. Furthermore, our dataset is collected from TV with unconstrained settings and unrestricted speech content.

AV modeling with Neural networks. In addition to the AV speech recognition [23], there has been more attention towards using deep neural networks (DNNs) for AV speaker naming with audio and visual streams. [14] used DNNs to jointly learn recognition models from 2 input streams. This work is further extended in the temporal domain with multimodal LSTM by [31]. Nevertheless, these works require identity information and are thus closer to biometric joint recognition than unsynchronization speech detection.

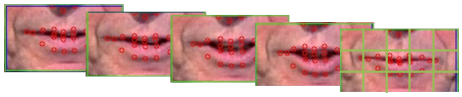

Figure 2: Example of mouth boxes. Mouth region is detected based on landmarks. Features are computed in 3×5 grid and grouped in a block of 5 frames.

3. MULTIMODAL FRAMEWORK

The overview of our system is illustrated in Fig. 1. First, features are extracted per frame for each modality. Subsequently, blocks of frames are concatenated and dimensionality reduction is applied. This is followed by cross-modality correlation modelling, whose outputs are modelled in the temporal domain using an LSTM to get the high level representation used for classification.

3.1 Feature extraction

Our goal is to build a full neural network to represent audio-visual speech. However, in this paper, we rely on standard features which should be sufficient for the task.

Visual stream. First, to obtain face tracks, we rely on the tracking-by-detection method described in [20]. Then, the mouth region is localized within each frame. This is done by detecting landmarks using the DLIB implementation of [18].

To characterise the mouth dynamics, dense optical flow is computed using the OpenCV implementation of [6]. The average flow is subtracted to remove head motion, and the residual flows are quantized into 8 bins based on their angular values, with 1 additional bin for close to static points. The mouth region is divided into 3×5 spatial regions in which flow histograms are computed, resulting in a vector of $3 \times 5 \times 9 = 135$ dimensions.

Audio stream. Every 10ms, we extract from 20ms windows Mel-frequency cepstral coefficient (MFCC) features with 13 coefficients and energy level together with first and second derivatives, resulting in a vector of 42 dimensions.

3.2 Multimodal processing

As often done in gesture recognition [28] and in NN-based AV speech recognition [23], we consider observations over a short interval (0.2s as in [28, 23]) to capture short-term temporal dynamics. Here, a block of 5 visual frames are grouped together (675-dim vector), which corresponds to 20 audio frames (840-dim vector). Principal component analysis (PCA) is applied separately to each modality to keep 95% of the variance, resulting in vectors of $N_{\mathrm{V}} = 100$ (visual) and $N_{\mathrm{A}} = 90$ (audio) dimensions.

Canonical-correlation analysis (CCA). The two modality streams contain different types of information. For example, audio may contain features about identity, semantics, or emotions which are irrelevant for our task and may have little correlation with the visual stream. To capture the synchrony between the two modalities, we use CCA, a powerful multivariate statistical technique. Its principle consists of learning matrices, one for each modality, which project the paired modality samples into a common space where

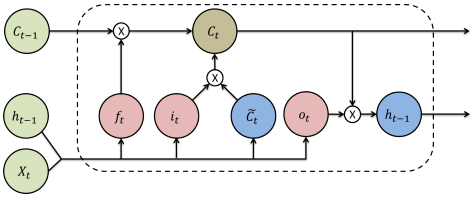

Figure 3: Temporal models. LSTM illustration. Red circles denote sigmoid activation of the gates while blue circles denote tanh activation of the states. $\times$ circles denote point-wise multiplication.

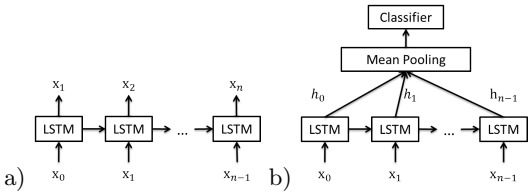

Figure 4: LSTM model. a) At each step i the LSTM learns to predict the feature vector x_{i+1} from the next time step. b) The LSTM is applied to the input sequence, and the sequence of hidden states h_i are averaged and used as input for classification.

the cross-correlation between the projected samples is maximized. For instance, let $X_A \in \mathbb{R}^{N_\text{A} \times N}$ and $X_V \in \mathbb{R}^{N_\text{V} \times N}$ be N audio and visual samples, respectively. Looking at a one dimensional subspace, CCA looks for the projection weights $w_\text{A} \in \mathbb{R}^{N_\text{A}}$ and $w_\text{V} \in \mathbb{R}^{N_\text{V}}$ such that:

$$\max_{w_\text{A}, w_\text{V}} corr \left[w_\text{A}^T X_A, w_\text{V}^T X_V \right] \qquad s.t. \; ||w_\text{A}|| = 1, ||w_\text{V}|| = 1.$$

Such optimization is conducted by finding w_A and w_V using the eigenvalue decomposition method on the correlation matrix, and then generalized to find the common subspace in which the audio stream and visual stream are most correlated [13] Thus, features from this subspace can represent how two modalities harmonize with each other, which will be important to detect dubbing events.

3.3 Temporal modeling and classification

In this Section, we introduce the LSTM architecture and then describe how it is used in our dubbing detection task.

Long Short Term Memory. In sequence modelling, the typical challenge is to learn a model mapping an input sequence $\{x_0, x_1, ..., x_n\}$ to an output sequence $\{y_0, y_1, ..., y_n\}$ where predictions at step t should use the knowledge from x_0 to x_t. To tackle this challenge, RNNs were introduced and shown to learn both high level representation of input signals and temporal dependencies. However, due to gradients multiplications during back propagation through time, they suffered from *exploding* or *vanishing* gradients, making it hard to learn long range dependencies [2].

LSTMs were introduced to overcome these issues [12]. The key ideas were to add a memory cells C_t to store useful information to model long term dependencies, as well as explicit gating mechanisms to regulate the memory updates, as illustrated in Fig. 3 and indicated by the formulae below:

$$\begin{aligned}
Gates: \quad f_t &= sigm(W_{xf} x_t + W_{hf} h_{t-1} + b_f), && (1)\\
i_t &= sigm(W_{xi} x_t + W_{hi} h_{t-1} + b_i), && (2)\\
o_t &= sigm(W_{xo} x_t + W_{ho} h_{t-1} + b_o), && (3)\\
States: \quad \tilde{C}_t &= tanh(W_{xc} x_t + W_{hc} h_{t-1} + b_c), && (4)\\
C_t &= f_t \times C_{t-1} + i_t \times \tilde{C}_t, h_t = o_t \times c_t && (5)\\
Output: \quad y_t &= W_y h_t + b_y, && (6)
\end{aligned}$$

where $W_.$ and $b_.$ denote weight matrices and biases. The mechanism works as follows. First, new information are processed from current states x_t and h_{t-1} to yield $\tilde{C}_t$. Then, to update C_t, the LSTM can selectively decide how much information from the past needs to be "remembered" or forgotten by passing C_{t-1} through the forget gate f_t, and replaced (reset) by new information $\tilde{C}_t$ through the input gate i_t. Finally, through the output gate o_t, the LSTM selects which C_t components to use to generate the hidden states h_t, from which the LSTM output y_t is produced. Importantly, the strategy to open or close gates is data driven and automatically learned from the data through the trainable $W_.$ and $b_.$. Also, the weighted addition of $\tilde{C}_t$ and C_{t-1} is crucial for LSTMs to avoid the vanishing gradient issue and to propagate gradient through long intervals.

Multimodal LSTM. Let $X = \{x_0, x_1, ..., x_n\}$ be a sequence of CCA projections for one segment. Because our task is binary, we have only one supervised signal denoting the class (Authentic or Dubbing). Straightforwardly, one could thus define the output sequence as a series of only 0s or of only 1s when appropriate and learn the LSTM classifier. However, such an approach does not constrain enough the network parameters, thus quickly leads to overfitting. Furthermore, in one dubbing segment, not all frames look asynchronous, thus forcing the label to 0 at every step can be misleading for the network to learn.

To overcome this challenge, similarly to [34, 33], we propose to train the LSTM in an unsupervised fashion with a bottleneck hidden layer h of size N_h: at each step t, the LSTM needs to predict the feature x_{t+1} of the next step, as shown in Fig. 4a. This architecture can learn good features for two related reasons. First, the hidden layer must be able to extract and compress the essential information from the input vector to make predictions. Since an input vector x_i is composed of two feature vectors of equal size coming from each modality, several hidden units will be able to capture the existing correlation between modalities, whereas others will perform intra-modality predictions (see Fig. 6). Second, to make better predictions and learn retention and anticipation temporal phenomena across modalities, the LSTM must also rely on features observed several steps in the past.

Finally, on a test sequence, the extracted hidden representations are mean pooled over the whole segment to form a single vector used for classification, as shown in Fig. 4b.

4. EXPERIMENTS

We describe below our experimental protocol and analysis of the talking face and dubbing detections results.

4.1 Experimental protocol

DW-Dubbing dataset[1]**.** We collect face tracks with their corresponding audio from Deutsche-Welle broadcast programs including debates and documentaries. Each track was divided into $2s$ segments. Segments with multiple arguing voices, inaudible speeches, or profile faces were discarded. The statistics of the dataset is shown in Tab. 1. Data from different videos were split into subsets used for unsupervised training, training and test data.The language of au-

[1] http://www.idiap.ch/scientific-research/resources

	Training set	Testing set	Unsupervised set
Authentic	617	444	1598
Dubbing	440	209	0
Silence	406	237	771

Table 1: Number of segments belonging to different splits and classes in the DW-dubbing dataset

	CV Acc.	Testing Acc.
MSD [4]	80.67	77.79
Mv [1]	78.92	82.16
HOF + SVM	78.39	79.06
HOF + LSTM	81.59	83.08

Table 2: Talking face detection results.

thentic speech/speaker segments was English, and dubbing segments taken from DW international documentaries had English voice dubbing a wide range of languages including Spanish, German, or other minority languages.

Protocol. For all models below, the authentic segments of the unsupervised set and of the training set were used to learn the PCA, CCA, and LSTM representations. Linear SVM classifiers were trained from the authentic and dubbing segments from the training set, using cross validation (CV) to determine hyperparameters. Evaluation was done on the test set, using accuracy as performance measures, along with recall and precision of authentic segments.

Models. To evaluate the contributions of the different elements, we tested several models. (i) **Audio.** This uses only MFCC features as the input for a SVM classifier. (ii) **PCA.** It consists of applying another PCA on the concatenation of the PCA representation of each modality. Keeping 95% of the representation, we obtained a 75 dimension vector for each block of frames, which were averaged and used as input to a SVM classifier. (iii) **CCA.** For each block, as shown in Fig. 1, the CCA projections (32 dimensions per modality) were computed, averaged and fed to a SVM. (iv) **PCA+LSTM.** It consists of a LSTM with $N_h = 16$ applied to the multimodal PCA representation of the PCA baseline. (v) **CCA+LSTM.** A LSTM with $N_h = 16$ is trained with the CCA projection vectors of the CCA baseline.

4.2 Experimental Results

Talking faces. In a preliminary experiment, we trained a LSTM model to detect talking faces from optical flow histograms computed at every frame. As in dubbing, the LSTM was trained to predict the next frame observations, and the average hidden state was used as input to a silent-vs-speaking classifier. Results in Tab. 2 demonstrate the benefit of the temporal information over other baselines (see Sec. 2 for details of [4] and [1]).

Dubbing. Tab. 3 displays the obtained results. Because one can possibly distinguish dubbing cases based on languages or quality of voices in the audio, Audio only can give some positive results. However, using both streams in

		Testing		
	CV Acc.	Acc.	Prec.	Recall
Audio	67.50	76.92	96.31	72.08
PCA	91.01	79.91	97.03	73.65
CCA	74.58	81.80	89.64	83.78
PCA + LSTM	85.44	83.76	94.69	81.53
CCA + LSTM	86.36	88.03	95.78	86.79

Table 3: Dubbing classification results on DW data.

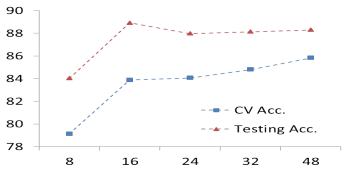

Figure 5: Training and testing accuracies for different values of N_h for the CCA+LSTM model.

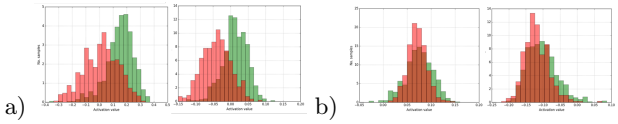

Figure 6: Hidden neurons activation distributions. Green distributions are from the authentic samples, red ones from dubbing samples. a) discriminative neurons. b) non-discriminative neurons.

PCA slightly improves the accuracy, this signifies the importance of multimodal analysis in this task. Nevertheless, the joint PCA subspace computed by maximizing variance is not expressive enough and results in confusing class observations, the classifier cannot be well generalized for the test set. CCA learns a better space where high or low correlation are expected depending on the class, leading to more stable results. By modeling the temporal dynamics within segments rather than averaging, the hidden state representation extracted from LSTM better discriminates the two classes and boosts the performance of both types of input. In this view, CCA offers a more suitable space for LSTM predictions of normal speech, and LSTM trained on CCA inputs outperforms LSTM trained from PCA.

This is confirmed by visualizing the activation distribution of the hidden neurons (i.e. each dimension of the hidden state). Typical examples are illustrated in Fig. 6 (CCA+ LSTM with $N_h = 16$). The two left neurons fire stronger when the two streams are correlated (in green), and are inhibited otherwise (in red). Neurons on the right fire similarly regardless of the classes, suggesting that they are probably specialized to process single modality inputs, whereas left ones incorporate cross-modality information, thus contributing significantly to detecting asynchrony.

Finally, to explore the LSTM parameter space, we vary the hidden size N_h from 8 to 48. Results are shown in Fig. 5. As N_h increases, the cross validation training accuracy increases, but not the testing results. This shows that large hidden size can lead to overfitting on the training set.

5. CONCLUSION

We have addressed dubbing detection in broadcast data, which involves detecting asynchrony between a visible speaker and the actual audio. In this context we proposed a multimodal algorithm comprising a CCA step, to capture the correlation between the 2 modalities, and a LSTM to capture the joint evolution of audio and mouth features in the common CCA space. For further research, we have made our DW-dubbing dataset available. Future improvements may include features learned with deep networks and deep CCA, resulting in an end-to-end trainable network. In addition, to detect more challenging dubbing situations, semantic understanding of the asynchrony origin will be needed.

Acknowledgement This research was supported by the European Union project EUMSSI (FP7-611057).

6. REFERENCES

[1] M. Bendris, D. Charlet, and G. Chollet. Lip activity detection for talking faces classification in TV-Content. In *ICMV*, 2010.

[2] Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE Trans. on Neural Networks*, 1994.

[3] G. Chetty and M. Wagner. Audio-visual multimodal fusion for biometric person authentication and liveness verification. In *NICTA-HCSNet Multimodal User Interaction Workshop*, 2006.

[4] M. Everingham, J. Sivic, and A. Zisserman. Hello! my name is... Buffy–automatic naming of characters in TV video. In *BMVC*, 2006.

[5] M. Everingham, J. Sivic, and A. Zisserman. Taking the bite out of automated naming of characters in TV video. *Image and Vision Computing*, 2009.

[6] G. Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Image analysis*. Springer, 2003.

[7] J. W. Fisher and T. Darrell. Speaker association with signal-level audiovisual fusion. *IEEE Trans. on Multimedia*, 6(3):406–413, 2004.

[8] P. Gay, E. Khoury, S. Meignier, J.-M. Odobez, and P. Deleglise. A Conditional Random Field approach for Audio-Visual people diarization. In *ICASSP*, 2014.

[9] P. Gay, S. Meignier, p. Deleglise, and J.-M. Odobez. Crf based context modeling for person identification in broadcast videos. *Frontiers in ICT*, 3, 2016.

[10] A. Giraudel, M. Carré, V. Mapelli, J. Kahn, O. Galibert, and L. Quintard. The {REPERE} Corpus: a multimodal corpus for person recognition. In *LREC*, 2012.

[11] J. Hershey and J. Movellan. Audio-vision: Using audio-visual synchrony to locate sounds. In *NIPS*, 2000.

[12] S. Hochreiter, S. Hochreiter, J. Schmidhuber, and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–80, 1997.

[13] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

[14] Y. Hu, J. S. Ren, J. Dai, C. Yuan, L. Xu, and W. Wang. Deep multimodal speaker naming. In *ACM Multimedia*, 2015.

[15] G. Iyengar, H. J. Nock, and C. Neti. Audio-visual synchrony for detection of monologues in video archives. In *ICME*. IEEE, 2003.

[16] B. Jou, H. Li, J. G. Ellis, D. Morozoff-Abegauz, and S.-F. Chang. Structured exploration of who, what, when, and where in heterogeneous multimedia news sources. *ACM MM*, 2013.

[17] V. Jousse, S. Petit-Renaud, S. Meignier, Y. Esteve, and C. Jacquin. Automatic named identification of speakers using diarization and {ASR} systems. In *ICASSP*, 2009.

[18] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *CVPR*, 2014.

[19] S. Kumagai, K. Doman, T. Takahashi, D. Deguchi, I. Ide, and H. Murase. Detection of inconsistency between subject and speaker based on the co-occurrence of lip motion and voice towards speech scene extraction from news videos. In *Multimedia (ISM), 2011 IEEE International Symposium on*, 2011.

[20] N. Le, D. Wu, S. Meignier, and J.-M. Odobez. Eumssi team at the mediaeval person discovery challenge. In *MediaEval 2015 Workshop*, 2015.

[21] C. Ma, P. Nguyen, and M. Mahajan. Finding speaker identities with a conditional maximum entropy model. In *ICASSP*, 2007.

[22] A. Morris, A. Hagen, H. Glotin, and H. Bourlard. Multi-stream adaptive evidence combination for noise robust asr. *Speech Communication*, 2001.

[23] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *ICML*, 2011.

[24] H. J. Nock, G. Iyengar, and C. Neti. Assessing face and speech consistency for monologue detection in video. In *ACM Multimedia*, 2002.

[25] A. Noulas, G. Englebienne, and B. J. A. Krose. Multimodal speaker diarization. *TPAMI*, 2012.

[26] F. Patrona, A. Iosifidis, A. Tefas, N. Nikolaidis, and I. Pitas. Visual voice activity detection in the wild. *Transactions on Multimedia*, 2016.

[27] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy. Cuave: A new audio-visual database for multimodal human-computer interface research. In *ICASSP*. IEEE, 2002.

[28] L. Pigou, A. Van Den Oord, S. Dieleman, M. V. Herreweghe, and J. Dambre. Beyond Temporal Pooling: Recurrence and Temporal Convolutions for Gesture Recognition in Video. *Arxiv*, pages 1–9, 2015.

[29] J. Poignant, L. Besacier, and G. Quénot. Unsupervised Speaker Identification in {TV} Broadcast Based on Written Names. *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, 2014.

[30] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior. Recent advances in the automatic recognition of audiovisual speech. In *Proceedings of the IEEE*, pages 1306–1325, 2003.

[31] J. S. Ren, Y. Hu, Y.-W. Tai, C. Wang, L. Xu, W. Sun, and Q. Yan. Look, Listen and Learn - A Multimodal LSTM for Speaker Identification. In *AAAI Conference on Artificial Intelligence*, 2016.

[32] E. A. Rúa, H. Bredin, C. G. Mateo, G. Chollet, and D. G. Jiménez. Audio-visual speech asynchrony detection using co-inertia analysis and coupled hidden markov models. *Pattern Analysis and Applications*, 2009.

[33] N. Srivastava, E. Mansimov, and R. Salakhutdinov. Unsupervised Learning of Video Representations using LSTMs. *Int. Conf. Machine Learning (ICML)*, 2015.

[34] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *NIPS*, 2014.

[35] M. Tapaswi, O. M. Parkhi, E. Rahtu, E. Sommerlade, R. Stiefelhagen, and A. Zisserman. Total Cluster: A person agnostic clustering method for broadcast videos. In *ICVGIP*, 2014.