# Active Speaker Detection with Audio-Visual Co-Training

## ABSTRACT

In this work, we show how to co-train a classifier for active speaker detection using audio-visual data. First, audio Voice Activity Detection (VAD) is used to train a personalized video-based active speaker classifier in a weakly supervised fashion. The video classifier is in turn used to train a voice model for each person. The individual voice models are then used to detect active speakers. There is no manual supervision - audio weakly supervises video classification, and the co-training loop is completed by using the trained video classifier to supervise the training of a personalized audio voice classifier.

## 1. INTRODUCTION

Detecting the presence of active speakers is an important task for several applications. Human-robot or human-computer interactions require the machine to know when a human, possibly one amongst several interlocuters, is speaking. Video conferencing systems could benefit from the identification of individual speakers, so that the system can zoom in on the active speaker, and broadcast her image. Video-diarization, the process of automatically annotating video with scene descriptions and the dialogues and actions of actors also requires the detection of speakers in the scene.

The detection of active speakers can be done using video, or audio, or a combination of the two. Early work in using audio-visual features for speaker recognition was done by [4, 18]. Cutler et al. [4] used a combination of audio-visual features for determining whether a single person in front of the computer was speaking or not. Temporally differenced frames from video and Mel frequency correlation coefficients (MFCC) from audio were used as separate features in a neural network classifier. Neti et al. [18] used Gabor-wavelet video features, combined with MFCC audio features for recognizing newsreaders in broadcast TV news.

Video and audio features can also be correlated using Canonical Correlation Analysis (CCA). Izadinia et al. [14] combined audio and video features for detecting parts of the scene responsible for sound. Spatio-temporal video features were clustered for identifying moving regions in the scene. Simultaneously, audio MFCC features were extracted from these frames and CCA was used to find canonical audio and video sub-spaces which maximize the correlation of the two features. Video regions highly correlated with audio were used to locate the dominant sound source in the image. Li et al. [17] also used CCA to maximize the correlation between video (eigen-faces) and audio MFCC features to detect talking heads in video.

Ren et al. [19] considered the problem of determining the identity of active speakers in TV series. CNN video features (used for recognizing actors), combined with MFCC audio features were used to train a Long-Short-Term-Memory (LSTM) network, and they demonstrated superior identification of the speaking actor compared to using either modality alone.

Directional sound information from microphone arrays have been used by [2, 8] for detecting active speakers. Gebru et al. [8] used a multi-target video tracker and directional sound information to assign speak/non-speak labels to people in a scene.

Audio can also be used to supervise the learning of a video-based active speaker classifier, as was shown by Chakravarty et al. [2]. Upper bodies of speakers were detected and tracked in video. Spatio-temporal features extracted from inside upper body tracks and speak/non-speak labels obtained from directional sound information (using a mic array) were used to train a video-only classifier. The use of spatio-temporal features was found to outperform lip-motion to detect speaking [4, 7, 17], especially in videos with multiple speakers in the scene that lack adequate resolution to discern the motion of individual lips. In further work [3], they demonstrated that a generic visual active speaker classifier can be modified, online, again using weak supervision from audio VAD, for individual speakers. Modifying a generic classifier to adapt to the quirks in expressions and gestures associated with individual speakers resulted in improved classification performance.

In this paper, we use the above video-based person-specific active speaker detection models to train personalized audio voice models. This further improves the performance of the detection of active speakers in the dataset used by [3], to almost 100%.

A common audio-only approach to speaker identification is clustering i-vectors [5]. However, they are mainly used in forensic applications with a large number of known speakers, where the identity of an audio segment with a single speaker

© 2016 ACM. ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

has to be found [11]. It has been shown that Non-negative Matrix Factorization (NMF) achieves comparable results to i-vectors [6], is able to give a time dependent speaker identity estimation for multiple active speakers [13] and can be extended to cope with overlapping speech [21]. So, this is the technique that we use for voice modelling in this paper.

Our key contribution is that *there is no manual supervision in the training of our active speaker detection system.* We use co-training between audio and video. The term co-training was first introduced by Blem et al. [1], where complementary sources of information are used to train classifiers and the most confident labelled samples from one classifier are used to train the other classifier, and classifier 2 can subsequently be used to generate more training samples to re-train classifier 1.

We use a preliminary model for generic active speaker detection from [2]. VAD was subsequently used to modify and adapt this generic visual model to individual speakers [3]. These individual visual active speaker detection models are used in this work, to learn person-specific audio voice models, to achieve almost perfect classification results. With this, we close the loop in active speaker detection. Earlier work [2, 3] demonstrated the ability of using audio to supervise video. In this paper, we show the reverse, that the learnt video models are capable of successfully supervising the training of audio voice models, thus demonstrating co-training. The experimental pipeline is shown in Figure 1. The rest of the paper is organized as follows: a description of the video and audio components of the system is given in sections 2.1 and 2.2 respectively, followed by experimental results in section 3 and conclusions in section 4.

# 2. SYSTEM DESCRIPTION

## 2.1 Video-based Active Speaker Detection

We use Improved Trajectory (IT) features, spatio-temporal features originally used for action recognition [20], and adapted by [2, 3] for active speaker detection. These features are a concatenation of Histogram of Oriented Gradients (HoG), Histogram of Optical Flow (HoF) and Motion Boundary Histogram (MBH) features calculated around feature points tracked over a sequence of 15 frames. Upper body detections using Deformable Part Models [10] are separately tracked, one track for each person in the scene, and IT features are collected from within these tracked upper body bounding boxes.

The concatenated [HoG+HoF+MBH] features are dimensionality reduced (using PCA), and subsequently pooled using a Fisher vector representation, as in [2, 3].

### *Prior Generic Model.*

We use a prior generic classifier model learnt on speakers in the Masters dataset collected by the authors of [2]. This model was trained using directional audio - a directional microphone supplied sound directional information that was fused with the video to obtain speak/non-speak training samples (comprising of spatio-temporal features pooled with Fisher vectors).

### *Person-specific Model with Online Learning.*

Chakravarty et al. [3] found that the generic classifier, trained on one dataset (Masters) could be modified online, to specific speakers in a new dataset, and this adaptation, weakly supervised by audio, gives better results on the new dataset, compared to using the prior generic classifier.

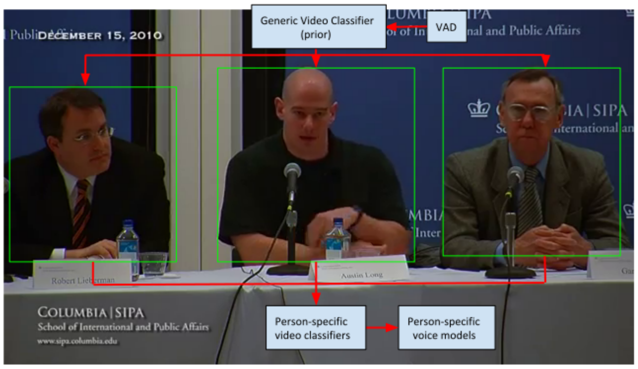
DECEMBER 15, 2010
Generic Video Classifier (prior)
VAD
Robert Lieberman
Austin Long
COLUMBIA | SIPA
School of International and Public Affairs
www.sipa.columbia.edu
Person-specific video classifiers
Person-specific voice models

**Figure 1: Experimental setup**

The prior model is used to train speaker-specific models on the Columbia dataset (used in all experiments in this paper), made available by [3]. The Columbia dataset is made from a YouTube recording[1] of a panel discussion at Columbia university, with 5 speakers, and the camera focusing on 2-3 people at any one time. A frame from the dataset is shown in Figure 1.

Following the setup of [3], Voice Activity Detection (VAD) [9] is first used as a weak supervisor, to separate frames where there is speech, from frames where there is none. The prior classifier is used to label speak/non-speak upper bodies for each person tracked in the frame. This is then used to learn a speaker-specific classifier for the Columbia dataset. Because the prior classifier has imperfect classification results on the new dataset, temporal continuity is used as an additional prior to weigh the samples. These are used to train an SVM with a weighted logistic loss function. With only a small amount of training (about 10 seconds per speaker), the baseline results from the prior classifier improve by 10-15 %. More details about online learning of the person-specific classifier are available in [3].

The positive samples with high certainty are then used to train the person-specific voice classifier. High certainty positive samples for each speaker are identified by using the samples that are temporally consistent for a threshold amount of time - a threshold of about 3 seconds of speaking was found to work well.

The complete pipeline, starting from the prior, generic classifiers to training the person-specific video and audio classifiers is shown in Figure 1, along with a frame from the Columbia dataset. A demonstration of the system is shown in the video submitted as supplementary material.

## 2.2 Audio-based Voice Classifier

In order to characterize speakers or voices, the magnitude spectrogram obtained from a Short Term Fourier Transform (SFTF) is represented as a weigted sum of dictionary elements. The dictionaries learned with Non-negative Matrix Factorization (NMF), are speaker-specific [6, 13, 21] and can therefore be used in the speaker identification process.

### *Nonnegative Matrix Factorization.*

NMF is a method that factorises a data matrix $\mathbf{X} \in \mathbb{R}_+^{F \times N}$ in to a (nonnegative) linear combination of (nonnegative)

[1] https://youtu.be/6GzxbrO0DHM

atoms in a dictionary $\mathbf{T} \in \mathbb{R}_+^{F \times K}$ and corresponding nonnegative activations (or linear coefficients) in an activation matrix $\mathbf{V} \in \mathbb{R}_+^{K \times N}$, such that $\mathbf{X} \approx \hat{\mathbf{X}} \triangleq \mathbf{TV}$ where $\mathbf{X}$ is the magnitude spectrogram and $K$ is the amount of atoms or basic building blocks in the dictionary [15]. To find such a dictionary and activation matrix, one first has to define a discrepancy measure between $\mathbf{X}$ and $\hat{\mathbf{X}}$. For this the generalized Kullback-Leibler (KL) divergence is used. To minimize this divergence, multiplicative update formulas for $\mathbf{T}$ and $\mathbf{V}$ have been found with convergence guarantees [16].

$$t_{fk} \leftarrow t_{fk} \frac{\sum_n \frac{x_{fn}}{\hat{x}_{fn}} v_{kn}}{\sum_n v_{kn}}, v_{kn} \leftarrow v_{kn} \frac{\sum_f \frac{x_{fn}}{\hat{x}_{fn}} t_{fk}}{\sum_f t_{fk}} \tag{1}$$

### *Audio-based Voice Classifier Supervised by Video.*

Section 2.1 details the video-based person-specific active speaker detection, and the use of temporal continuity to get segments of the video belonging to each active speaker with high confidence. The audio feature vectors in the corresponding time segments are grouped in a training data matrix $\mathbf{X}^j$ for the $j^{th}$ estimated speaker identity. Using NMF, a dictionary $\mathbf{T}^j$ is created for each speaker $j$. Note that the data matrices are constructed on estimated speaker identities and not on actual (ground truth) speaker identities. This means that $\mathbf{X}^j$ might contain data not belonging to the $j^{th}$ speaker. However, by keeping the amount of atoms in a dictionary low, these error frames should not fit in the dictionary. The dictionaries are collected in a library $\mathbf{T}_{tot} = [\mathbf{T}^1, \mathbf{T}^2, \ldots, \mathbf{T}^J]$. In the test phase this library is used to reconstruct the voice samples of all the speakers $X^{test}$. Using only the second part of equation 1 and keeping the library fixed, the global activations $\mathbf{V}_{tot}^{test}$ are found. The sum of the atom activations of a dictionary is a measure of the total dictionary activation and thus also speaker activation.

$$y_n^j = \sum_{k \in \kappa^j} v_{tot,kn}^{test} \tag{2}$$

where $\kappa^j$ are the indices of the atoms belonging to the dictionary of the $j^{th}$ target speaker. $y_n^j$ is smoothed over time with a moving average filter of window size 5 and is then normalized per frame. When $y_n^j$ is larger than some threshold, the $j^{th}$ speaker is assumed active in the $n^{th}$ timeframe.

Group sparsity is enforced on the activations [12] to express that atoms should preferably be selected from the same speaker. In addition to the KL divergence, an extra cost term is used, which increases when atoms from different dictionaries are activated. Even if dictionaries contain some similar atoms, a solution will be sought where a minimal amount of dictionaries are active.

## 3. EXPERIMENTS

We use the Columbia dataset, made available by [3], for our experiments. The data is from a panel discussion of 5 speakers, with the camera focusing on 2-3 people at a time. At any one time, only one person is speaking, except during margins of speaker change, where both speakers are briefly speaking at the same time.

Each of the panel members has a microphone approximately 30 cm in front of them. These audio signals are sent to the electronic speakers in the auditorium which in turn are recorded by the microphone of the camera. The final received microphone signal is a combination of the direct path from the electrical speakers to the microphone of the camera, as well as some multipath reflection. Some reverberation is thus expected. The audio is downsampled from 44.1 kHz to 16 kHz for our experiments and the STFT is calculated using a window size of 64 ms and a stride-length of 32 ms. The size of the NMF dictionaries $K$ is empirically chosen at 40.

### *Baselines.*

We test the performance of the audio voice classifiers, under training with different amounts of noise in the supervision labels. This simulates mistakes in the labelling of the video-based personalized active speaker detectors, that will later be used to train the audio voice models. We conduct the following supervision baseline experiments:

- *Full supervision*: Speaker dictionaries are trained with ground-truth voice samples, and these trained dictionaries are used to test the identities of the same samples. This should give nearly 100% performance.
- *Noisy supervision*: Speaker dictionaries are trained with increasing amounts of noise added to the labels, and tested on the ground truth samples.
- *Incremental supervision*: In addition to the noise added to the labels, the amount of training data, $N_{tr}$, is varied, from 4 to 80 seconds per speaker. This simulates online learning. Small amounts of labelled data can be used to train a voice classifier online that can be tested on the remainder of the data.

Area under Curve (AUC) values for the above experiments are shown in Table 1. It is evident (from Table 1) that large amounts of incorrectly assigned training data are acceptable - with an AUC close to 1 even with 30-40% of noise in the labels at 8 and 12 seconds of training data per speaker . This makes the system robust to errors in supervision - relevant to the experiment where the outputs of the personalized video-based active speaker detectors are used to train the audio voice models.

We conduct 2 unsupervised experiments with only audio:

- *Clustering (Aud Clus)*: No annotation data is used, apart from the total number of active speakers $J$. The STFT features are used to determine MFCC features with differentials (MFCC$\Delta$) for each time frame. A moving window of 1.5s is shifted over these MFCC($\Delta$) features and each time a 15-dimensional i-vector is determined (using a Universal Background Model (UBM) with 512 components and a total variability space (T)) [5]. A Gaussian Mixture Model (GMM) is fitted to this data, where the number of clusters is equal to $J$. It is expected that each cluster represents a different speaker and thus time segments are clustered per speaker. Results per speaker are presented as Aud Clus in Table 2.
- *Clustering supervises training of audio dictionaries (Aud Sup)*: Using the clustered samples as supervision for NMF further improves the performance. Per speaker AUC results for the NMF voice models trained with supervision from the unsupervised clustering (Aud Clus) are presented as Aud Sup in Table 2.

| $N_{tr}$ | Noise labels in % | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 0 | 20 | 30 | 40 | 50 | 60 | 70 | 80 |
| 4 | 0.98 | 0.99 | 1.00 | 0.98 | 0.93 | 0.84 | 0.59 | 0.56 |
| 8 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.94 | 0.78 | 0.79 |
| 12 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.97 | 0.83 | 0.49 |
| 20 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.97 | 0.92 | 0.40 |
| 40 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.99 | 0.98 | 0.53 |
| 80 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.97 | 0.50 |

**Table 1: AUC-values for noisy supervision and variable training size (in seconds).**

| | S1 | S2 | S3 | S4 | S5 | avg. |
|---|---|---|---|---|---|---|
| Vid1 | 0.80 | 0.80 | 0.72 | 0.75 | 0.58 | 0.73 |
| Vid2 | 0.88 | 0.94 | 0.89 | 0.81 | 0.71 | 0.85 |
| Aud Clus | 1.00 | 1.00 | 1.00 | 0.98 | 0.95 | 0.98 |
| Aud Sup | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| Vid Sup | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table 2: The AUC values for speakers 1-5 (S1-S5) for the prior, generic video classifier (Vid1), the online learnt video classifier (Vid2), the unsupervised audio clustering mechanism (Aud Clus), the audio supervised audio classifier (Aud Sup) and the video supervised audio classifier (Vid Sup).**

| | -3 dB | 0 dB | 3 dB | No noise |
|---|---|---|---|---|
| Vid1 | 0.73 | 0.73 | 0.73 | 0.73 |
| Vid2 | 0.85 | 0.85 | 0.85 | 0.85 |
| Aud Clus | 0.78 | 0.82 | 0.87 | 0.98 |
| Aud Sup | 0.81 | 0.89 | 0.96 | 1.00 |
| Vid Sup | 0.96 | 0.99 | 1.00 | 1.00 |

**Table 3: The average AUC values for different SNR levels of added white Gaussian noise to the audio signal.**

### *Voice Classifiers Supervised by Video.*

Chakravarty et al. [3] demonstrated online learning of the video-based person-specific active speaker classifiers, using VAD for weak supervision, and these numbers (Vid2) are quoted in Table 2 for comparison. For online learning, a prior, generic classifier (learnt on another dataset) is used to pick out positive and negative samples (spatio-temporal IT features pooled using a Fisher vector representation) for each person. These samples are weighted by their temporal continuity, and used to train a Support Vector Machine (SVM) classifier, as described in Section 2.1. The training is done using a maximum of 10 seconds of data per speaker, and the first row, Vid1, in Table 2 shows the per-speaker AUC results for the prior, generic active speaker classifier. The second row, Vid2, shows the results with the online-learnt classifier, and it can be seen that the classifier, using weak audio supervision (VAD), already improves by an average of 16% with just 10 seconds of training.

The above person-specific video classifiers that are learnt online on this dataset, are subsequently used to train per-person audio voice-classifiers. As mentioned in section 2.1, the high-precision, low-recall area of the video-based classifier is used to train the audio voice classifier. These are parts of the data that have been continuously determined to be speak/non-speak for a fixed amount of time - a 3 second threshold is used in our experiments. The corresponding audio feature vectors of those classified video samples are used to train the person specific voice models. This way, errors in the training labels of the video classifier are minimized at the cost of the number of learning samples. We achieve perfect classification results using the voice models trained using video supervision - AUC of 1 for all 5 speakers (Vid Sup in Table 2). This compares well with the results from the unsupervised audio clustering mechanism (Aud Clus) and the audio classifier supervised by clustering (Aud Sup), and significantly improves the results from the two video classifiers Vid1 and Vid2.

To summarize, the generic video-based active speaker detector achieves an average AUC of 0.73 - this improves with online learning of personalized video-based active speaker detectors (under weak supervision from audio VAD), improving results by 16%. Using the outputs of these online active speaker detectors to train person-specific voice models further improves results by a further 15%, achieving perfect classification AUC scores of 1.0.

One might question the necessity of the video supervision (Vid Sup in Table 2) if the audio training of voice models using clustered samples (Aud Sup in Table 2) already give 100% classification results. The total number of speakers in the audio unsupervised clustering was assumed to be known in our experiment - this is information that can be gleaned from video. Additionally, video allows the linking of audio fragments with speaker images. Further, if there is noise in the audio, using video to supervise the training of the audio classifier becomes a necessity. We add artificial white Gaussian noise to the audio fragments, and repeat our experiments (Table 3). It can be seen that average AUC values for the video supervised voice classifier (Vid Sup) exceed the voice models trained in an unsupervised setting - Aud Clus and Aud Sup with the addition of noise. This is because noise (like coughing by a speaker) could be modeled as an additional cluster(s) and two or more speakers could be modeled by the same cluster. Thus, we demonstrate that using video supervision, the audio classifier is robust in noisy environments.

## 4. CONCLUSIONS

In this paper, we have shown that person-specific voice models can be built using cross-modal supervision from video, thus completing the loop in audio-visual co-training. A generic, video-based active speaker classifier that was trained by directional audio [2] is used to train a video-based person-specific active speaker detection system on a new dataset, using just a few seconds of video per person. These online-learnt video classifiers are in-turn used to supervise the training of personalized voice models, leading to near-perfect active speaker detection accuracy. The whole process is (human) *unsupervised* from beginning to end.

In future work, we will adapt the co-training for audio-visual active speaker detection on more challenging datasets, from movies and TV series. The challenges for the visual classifier will be a moving camera and more varied views of people with non-frontal poses. The audio voice classifier training will have challenges from distracting background, music and noise.

# 5. REFERENCES

[1] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100. ACM, 1998.

[2] P. Chakravarty, S. Mirzaei, T. Tuytelaars, and H. Van hamme. Who's speaking? audio-supervised classification of active speakers in video. In *ACM International Conference on Multimodal Interaction (ICMI)*, 2015.

[3] P. Chakravarty and T. Tuytelaars. Cross-modal supervision for learning active speaker detection in video (http://arxiv.org/abs/1603.08907v1).

[4] R. Cutler and L. Davis. Look who's talking: Speaker detection using video and audio correlation. In *IEEE International Conference on Multimedia and Expo*, pages 1589–1592, 2000.

[5] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798, 2011.

[6] S. Drgas and T. Virtanen. Speaker verification using adaptive dictionaries in non-negative spectrogram deconvolution. In *Latent Variable Analysis and Signal Separation*, pages 462–469. Springer, 2015.

[7] M. Everingham, J. Sivic, and A. Zisserman. Taking the bite out of automated naming of characters in tv video. *Image and Vision Computing*, 27(5):545–559, 2009.

[8] I. Gebru, S. Ba, G. Evangelidis, and R. Horaud. Tracking the active speaker based on a joint audio-visual observation model. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 15–21, 2015.

[9] F. Germain, D. L. Sun, and G. J. Mysore. Speaker and noise independent voice activity detection. In *INTERSPEECH*, pages 732–736, 2013.

[10] R. B. Girshick, P. F. Felzenszwalb, and D. McAllester. Discriminatively trained deformable part models, release 5. http://people.cs.uchicago.edu/ rbg/latent-release5/.

[11] C. S. Greenberg, D. Bansé, G. R. Doddington, D. Garcia-Romero, J. J. Godfrey, T. Kinnunen, A. F. Martin, A. McCree, M. Przybocki, and D. A. Reynolds. The nist 2014 speaker recognition i-vector machine learning challenge. In *Odyssey: The Speaker and Language Recognition Workshop*, 2014.

[12] A. Hurmalainen, R. Saeidi, and T. Virtanen. Group sparsity for speaker identity discrimination in factorisation-based speech recognition. In *INTERSPEECH*, pages 2138–2141, 2012.

[13] A. Hurmalainen, R. Saeidi, and T. Virtanen. Noise robust speaker recognition with convolutive sparse coding. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[14] H. Izadinia, I. Saleemi, and M. Shah. Multimodal analysis for identification and segmentation of moving-sounding objects. *IEEE Transactions on Multimedia*, 15(2):378–390, 2013.

[15] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

[16] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems 13*, pages 556–562. MIT Press, 2001.

[17] D. Li, C. M. Taskiran, N. Dimitrova, W. Wang, M. Li, and I. K. Sethi. Cross-modal analysis of audio-visual programs for speaker detection. In *MMSP*, pages 1–4, 2005.

[18] B. Maison, C. Neti, and A. Senior. Audio-visual speaker recognition for video broadcast news. *Journal of VLSI signal processing systems for signal, image and video technology*, 29(1-2):71–79, 2001.

[19] J. 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.

[20] H. Wang and C. Schmid. Action recognition with improved trajectories. In *ICCV*, pages 3551–3558, Sydney, Australia, Dec. 2013.

[21] J. Zegers and H. Van hamme. Joint sound source separation and speaker recognition. In *Interspeech*, pages 2228–2232, 2016.