Cross-Modal Supervision for Learning Active Speaker Detection in Video

Punarjay Chakravarty(✉) and Tinne Tuytelaars

ESAT-PSI-iMinds, KU Leuven, Leuven, Belgium
{Punarjay.Chakravarty,Tinne.Tuytelaars}@esat.kuleuven.be

Abstract. In this paper, we show how to use audio to supervise the learning of active speaker detection in video. Voice Activity Detection (VAD) guides the learning of the vision-based classifier in a weakly supervised manner. The classifier uses spatio-temporal features to encode upper body motion - facial expressions and gesticulations associated with speaking. We further improve a generic model for active speaker detection by learning person specific models. Finally, we demonstrate the online adaptation of generic models learnt on one dataset, to previously unseen people in a new dataset, again using audio (VAD) for weak supervision. The use of temporal continuity overcomes the lack of clean training data. We are the first to present an active speaker detection system that learns on one audio-visual dataset and automatically adapts to speakers in a new dataset. This work can be seen as an example of how the availability of multi-modal data allows us to learn a model without the need for supervision, by transferring knowledge from one modality to another.

Keywords: Active speaker detection · Cross-modal supervision · Weakly supervised learning · Online learning

1 Introduction

The problem of detecting active speakers in video is a central one to several applications. In video conferencing, knowing the active speaker allows the application to focus on and transmit the video of one amongst several people at a table. In a Human-Computer-Interaction (HCI) setting, a robot/computer can use active speaker information to address the correct interlocuter. Active speaker detection is also a part of the pipeline in video diarization, the automatic annotation of speakers, their speech and actions in video. Video diarization is useful for movie sub-titling, multimedia retrieval and for video understanding in general.

Traditionally, visual active speaker detection has been done using lip motion detection [1–4]. However, facial expressions and gestures from the upper body,

This work was supported by the KU Leuven GOA project *CAMETRON* and iMinds.

B. Leibe et al. (Eds.): ECCV 2016, Part V, LNCS 9909, pp. 285–301, 2016.
DOI: 10.1007/978-3-319-46454-1_18

movement of the hands, etc., are all cues that can be utilized to assist with this task, as shown in [5], where better detection results are achieved using spatio-temporal features extracted from the entire upper body, compared with just lip motion detection.

Another powerful idea we borrow from [5], is to use audio to supervise the training of a video based active speaker detection system. In that work, a microphone array is used to get directional sound information (assumed to be speech sounds), and based on this input, upper body tracks are associated with speak/non-speak labels. These labels are then used to train an active speaker classifier using video only.

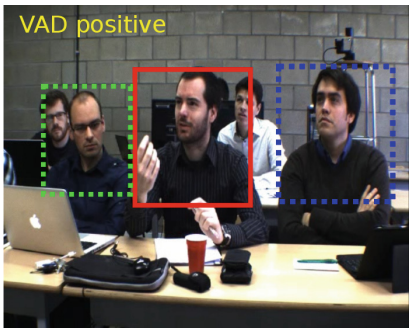

Fig. 1. Audio-based Voice Activity Detection (VAD) is used to weakly supervise the training of a video-based active speaker classifier. VAD tells us that someone in the frame is speaking, but not who. The problem is one of associating the voice activity with one of the people (solid red upper body bounding box) in the frame, and training the classifier at the same time. We use structured output learning to train a latent SVM classifier in the presence of partially observed (latent) inputs. (Color figure online)

However, the presence of reverberation and background noise prevents perfect active speaker identification using directional audio alone, which subsequently affects the training of the video-based classifier. Additionally, in the vast majority of videos, such as the millions of Youtube videos available online, in videos from films and TV series, only a single channel of sound is available, with no directional information. Even in those cases where 2 channels of audio are available, the relative position of the camera and the microphones varies, and no calibration information is available, making it impossible to apply the method of [5].

In the absence of directional information, we propose to use Voice Activity Detection (VAD) [6] to tell us when there is someone speaking in a frame. If there is only one person in the frame, then this can be used to train the video-based classifier directly. However, when this is not the case, the problem becomes one of simultaneously associating the voice activity with one of the people in the frame, and learning the classifier (Fig. 1). That's the challenge we address in this work.

Moreover, there's an additional challenge. Investigating our trained classifier, we find that it has some bias: it works better for some speakers, compared to others. We identify two reasons for this. First, the way people gesticulate while speaking varies a lot from person to person. Indeed, a person-specific model typically outperforms the generic model. Second, there is the domain shift problem: the change of data distribution between training and test data. We address both by extending our previous scheme to an online learning setting that, starting from a generic classifier, gradually adapts to a specific person. To this end, we retrain the model with an incrementally increasing number of training samples coming from a new video of a previously unseen person at each iteration. The online training is also weakly supervised by VAD from audio. The generic classifier is used to label and pick the training samples for each speaker and temporal continuity constraints allow the classification performance to improve in spite of imperfectly labelled training data from the generic classifier.

Our method is completely unsupervised, in the sense that there is no *human* supervision/labelling. We use audio to supervise the learning. This supervision comes "for free" with the video, but is only partial - VAD tells us that one of the persons in the frame is speaking, but not who. As opposed to [5], who use full supervision from directional audio, we use weak supervision from VAD. This work can be seen as an example of how the availability of multi-modal data allows us to learn a model without the need for supervision, by transferring knowledge from one modality to another.

The remainder of the paper is organized as follows. We discuss prior work in this area in Sect. 2. We discuss the use of audio for active speaker detection in Sect. 3, with Subsects. 3.1, 3.2 and 3.3 discussing the weakly supervised learning with Latent SVMs, speaker specific classification and online learning, respectively. Experimental results are discussed in Sect. 4 and concluding remarks and potential for future work in Sect. 5.

2 Related Work

Weakly supervised and multimodal learning. The learning of a classifier in the presence of weak supervision, or partially labelled data, has been studied mostly in the context of object recognition, where labels are available for images, but localization information - bounding boxes around the objects to be classified, are missing [7–11]. Best results in this context are obtained with Structured Output Learning [12], i.e. by learning a classifier that outputs not only the class labels, but also the bounding box coordinates or index. We use the same approach for training a classifier for active speaker detection with only VAD-based supervision, which gives us labels for the images, but not for individual bounding boxes. Audio weakly supervises the training of video. The work of Bojanowski et al. [13] is another example of one mode of information weakly supervising another. They use scripts to weakly supervise the learning of actors and actions in movies. However, scripts are not always available for video data, while audio is.

Dealing with domain shift. In our work, we find that an active speaker classifier trained on a first set of speakers performs less well on previously unseen speakers, while best results are obtained with person-specific classifiers. This is because of the mismatch between the distributions of different speakers. On the one hand, training a generic classifier means that it has seen a larger number of training samples, is less prone to overfitting compared to a person specific classifier, and should generalize well for unseen speakers. However, the generic classifier still suffers from person-specific biases, and gives better classification results for some people over others. The same problem exists for object recognition - a classifier trained on one dataset typically has lower performance when applied to images from another dataset. This is known as the dataset bias problem, and there have been some efforts at reducing this for object recognition [14,15]. One way to deal with person or dataset specific biases is to adapt the source classifier to the target classifier, and this is called Domain Adaptation [16,17]. Transfer Learning [18–20], a related problem, is about using the information available from the source data to aid the learning of the target classifier utilizing only a small number of target training samples. For instance, Aytar et al. [18] use an Adaptive SVM (ASVM) that incrementally adapts an SVM learnt on source data (e.g. motorbike class) to target data (e.g. bicycle class) in the context of object recognition. The source classifier acts as a regularizer for the target classifier in the adaptive SVM framework, and they demonstrate successful adaptation based on a relatively small number of training samples of the target class. This work lies at the basis of our online adaptation to previously unseen persons.

Person-specific models. There has been some work on person specific facial expression recognition and transferring generic to specific models for improving classification performance [21–23]. Chen et al. [21] show that facial expression recognition results improve when using person specific classifiers. They use an Inductive Transfer Learning (ITL) approach, where they learn a source classifier, which is a collection of weak learners in a boosting framework. Subsequently a subset of these are used for training the target classifier with a small number of labeled target samples.

Chu et al. [22] propose a Selective Transfer Machine (STM) approach to re-weight the source samples so that they are closer to the target samples. The algorithm simultaneously learns the parameters of the classifier and the source sample weights that minimize the error between the source and target distributions. They thus personalize a generic classifier to individual, with the resulting personalized classifier improving on the generic classifier on facial action unit detection tasks. However, STM requires the storage of all source samples, with a higher memory requirement than storing just the source classifier, which could be the weights of an SVM.

Zen et al. [23] demonstrate unsupervised adaptation of a generic classifier to a target classifier on single frame expression datasets. They learn a regression function between the "shape" or sample distribution of each user in the labelled source dataset and his/her classifier (source weight vector w_i in the SVM). Applying this function on the unlabelled sample distribution of the target

user then gives them the target classifier (target weight vector w_t). They do not require to keep in memory all the samples from the source dataset and outperform the STM method of [22]. However, their approach requires that the relative distribution of positive and negative samples in every user's data is relatively constant and can be learnt using the source users. However, this is not the case in our data. Additionally, we learn the generic source classifier using unlabelled data as well - so our process requires no human supervision from beginning to end.

Online learning is the incremental learning of a classifier with an increasing number of training samples as and when they become available. In our context, we adapt the generic source classifier to the person-specific target classifier with an increasing number of samples from the speaker. This is somewhat similar to the problem of Active Learning, where a new classifier is to be learnt with the minimum budget in terms of time spent in labelling training samples, and the task is one of selecting the most relevant samples to be used for training. Gavves et al. [24] demonstrate Active Transfer Learning, in that the selection of relevant training samples is done with the help of previously learnt classifiers on other datasets. Both [23,24] use the source classifiers as zero-shot priors, giving a baseline performance using only the target classifier, with classification performance gradually increasing with an increasing number of samples from the target dataset. We use this as our inspiration for our online learning problem, except again, our learning is without any manual supervision.

3 Audio Supervised Training

In the original experiment of [5], a 2-mic array was used to associate upper bodies detected in the video, with sound directions. They used a technique proposed by [25] for estimating the number and direction of sound sources. A non-linear function of the Generalized Cross Correlation Phase Transform (GCC-PHAT) between the audio signals is calculated over all the angles of arrival with respect to the microphone array baseline. This is done over short time intervals corresponding to the Time Frequency cells of a Short Term Fourier Transform. This gives an angle of arrival spectrum at each point in time that can be associated with the people detected in the image. In each frame, the sound direction is associated with a speaker's upper body bounding box, and features within that bounding box are used to train the classifier. We use the same data as [5], available on request from the authors, and consider the case when directional information is absent. We simulate the output of VAD by removing the speak/non-speak bounding box labels. We assign a label of speak to the frame if any of the bounding boxes in it are tagged as speaking and non-speak if none of the bounding boxes is speaking. Our problem is one of associating one of the bounding boxes in the image with the sound and training a classifier at the same time. We treat this as a structured output prediction problem [26].

3.1 Classifier Training Under Weak Audio Supervision Using Structured Output Learning

In the absence of information about which upper body bounding box is associated with the active speaker in each frame, the problem can be posed as a structured training problem [7–9,12], in the presence of partially observed training data. In the context of object recognition, there are databases with images labelled with the presence of one or more objects in the scene, but no localization (bounding box) information for the object in the image. [7–9] deal with this by using a Latent SVM formulation, which alternates between the guessing of object bounding boxes, and training a classifier for the object inside the bounding box. They use object proposals [27] to narrow down the search for objects in the image.

Here, we adapt [7–9] to our setting. Our object proposals are the upper body bounding boxes. We know that one of the bounding boxes is an active speaker, but not which one - the speak/non-speak label for the individual bounding boxes are our latent variables. Using structured output prediction, we jointly learn which of the bounding boxes in the image is associated with the active speaker, together with learning the active speaker classifier. Given an image x and upper body bounding box h, let $\phi(x,h)$ denote an image description computed over bounding box h. Given all upper body bounding boxes $h_1, ...h_n$, the algorithm then needs to select the bounding box that contains the active speaker. The labels of the images, speaking/non-speaking, $y = \pm 1$, are obtained from the sound using VAD or, in our experiment, by removing the directional information from the training data. Once the classifier is trained, the best bounding box h is found by

$$h^* = \underset{h}{\operatorname{argmax}} \langle w, \phi(x,h) \rangle \tag{1}$$

where w is the weights vector of the SVM. We define $\Phi(x,y,h) = \phi(x,h)$ if $y = 1$, and 0 otherwise. The learning task is to optimize the following:

$$\hat{w} = \underset{w}{\operatorname{argmin}} \sum_{i=1}^{N} l(w, x^i, y^i) + \frac{C}{2}\|w\|^2 \tag{2}$$

where $l(w, x^i, y^i)$ is the per example loss, $\frac{C}{2}\|w\|^2$ is the regularizer and N is the total number of training data. The max-margin loss function is defined as

$$l_{mm}(w, x^i, y^i) = \max_{y,h}(\langle w, \Phi(x^i, y, h) \rangle + \Delta(y^i, y)) - \max_{h}(\langle w, \Phi(x^i, y^i, h) \rangle) \tag{3}$$

where $\Delta(y^i, y)$ is the zero-one error, which is 0 if $y^i = y$ and 1 otherwise.

This loss function tries to maximize the margin between the score of the selected active speaker's bounding box and the non-speaking bounding boxes. Following the work of [8,9], we replace the max-margin loss with a soft-max loss function:

$$l_{sm}(w, x^i, y^i) = \frac{1}{\beta} log \sum_{y,h} exp(\beta \langle w, \Phi(x^i, y, h) \rangle + \beta \Delta(y^i, y)) - \frac{1}{\beta} log \sum_{h} exp(\beta \langle w, \Phi(x^i, y^i, h) \rangle) \quad (4)$$

where β controls the sharpness of the distribution. It can be shown that as $\beta \to \infty$, the loss function limits to the standard structured SVM formulation. The softmax loss function allows for multiple active speakers in the same frame. It also makes the optimization function smoother and less prone to local minima. We use the LBFGS solver from minFunc[1] to optimize our cost function and train our classifier.

3.2 Speaker Specific Models

Using the motion of the face and upper body over time assists with active speaker detection. At the same time, it maybe has the disadvantage of making the detector more speaker specific, as different people are likely to have different mannerisms while speaking. We explore this hypothesis by training several person specific Active Speaker classifiers. We do this in two settings: one using the directional audio (i.e., supervised), as a baseline, and subsequently, in the VAD setting, where the learning is weakly supervised by audio, as detailed in the previous section.

In the first case, the learning is straightforward: we have a separate track for each person in the video, and knowledge of the frames in which that track is speaking (from the directional audio).

In the second case, the audio does not tell us which track/person is speaking at any given time, just that one among the multiple tracks in the frame is speaking. For this, we do the training in two steps. We first learn a generic classifier in the weakly supervised case, as detailed previously. Subsequently, we use the generic (source) classifier to guide the selection of the positive samples for the person specific (target) classifier. We run the generic classifier on each "speaking" frame's bounding boxes to get an idea of which track/bounding box is speaking. However, the generic classifier does not always give the highest score to the active speaker in the frame. This is because of the dataset bias and domain shift problem discussed earlier - the generic classifier performs better for some speakers compared to others. So we bring in another cue: temporal continuity.

So far, we have discussed active speaker detection on each frame in isolation. However, people's speech tends to be for periods longer than a single frame. If a person is speaking in one frame, it is more likely than not, that they will be speaking in the next frame as well. We use temporal continuity to reduce the effect of mis-classifications of the generic classifier and guide the sample selection for the speaker specific classifier. The highest scoring sample at each VAD-positive frame is taken to be the positive sample for the associated speaker, and all other samples are selected as negative samples for the other speakers. Both positive

[1] http://people.cs.ubc.ca/~schmidtm/Software/minFunc.html.

and negative samples are weighted according to temporal continuity, measured as the number of contiguous neighbouring frames with consistent labels. We use a weighted logistic loss function l_{wll}

$$l_{wll}(w, \Phi(x, y, h^*), \alpha) = \alpha \cdot log\{1 + exp(-\langle w, \Phi(x, y, h^*)\rangle)\} \tag{5}$$

where $\Phi(x, y, h^*)$ is the feature vector from the best scoring bounding box, w is the weights vector of the speaker-specific SVM and α is the temporal continuity weight of the sample.

Note how this integration of temporal continuity directly in the weakly supervised learning framework (as opposed to keeping it as a postprocessing step, as is usually done) reflects again one of the core ideas behind our work, that combining multiple, independent sources of information - be it multiple modalities, or temporal vs. spatial information - allows learning models with less supervision.

3.3 Online Learning

In this section, we deal with the problem of learning the specific model in an online fashion for a speaker who has not been seen earlier during training. This can be the case during a live setting, where we don't have the entire data available to us at any given time, just what we have seen so far. To this end, we use a model inspired by the Tabula Rasa Transfer Learning model of Aytar et al. [18].

The idea is that the generic model is used as a zero-shot prior, and already gives a baseline performance, that can be improved as a new speaker specific model is trained incrementally with every additional batch of samples that trickle in from the new speaker. This allows us to have a model that performs better than the prior, generic model in an iterative fashion, without needing to see all the target samples. The process of online learning of speaker-specific classifiers is again weakly supervised by audio: it assumes that VAD is available for the target speaker data as well.

As in the offline case for training speaker specific models (Subsect. 3.2), we use VAD to detect the frames in which human speech is present. Subsequently, the generic (source) classifier is used to guide the selection of the positive samples for each new speaker (target). We select the highest scoring bounding box in each VAD-positive frame as the positive sample for the speaker associated with it, and the remaining bounding boxes are selected as negative samples for the other speakers. Temporal continuity is used to weigh both the positive and negative samples (Eq. 5). Motivated by [24], we use the prior (source) model, not just for the selection of the target speaker's positive training samples, but also for target prediction. During prediction, the generic model scores are added to the target model scores so that the prediction score from online learning, at each iteration is given as:

$$f^t(\phi(x, h)) = \langle w^{gen}, \phi(x, h)\rangle + \langle w^t, \phi(x, h)\rangle \tag{6}$$

Each time step t has an increasing number of training samples to train the classifier w^t at that iteration. w^{gen} remains constant during online learning. This

results in the person-specific target classifier being at least as good as the generic source classifier, and getting progressively better with an increasing number of training samples.

4 Experiments

We use the audio-visual dataset made available by the authors of [5]. It consists of 7 recordings of masters student thesis presentations to a jury of examiners. Each student presents for 25 min, followed by 5 min of questioning by the jury. The microphone array, with its directional sound information in a cone of 180 degrees in front of it is associated with upper body tracks of the jury. We will call this the Masters student dataset in the rest of the paper. An example frame of this data is shown in Fig. 1. [5] used the directional sound information from the microphone array, associated with the bounding boxes of persons in the frame to train their video-based active speaker classifier. We simulate VAD by removing this directional information from the data, leaving only a label of speak/non-speak per frame. Like [5], we only use the 3 people from the jury in the front row of the audience, as others behind them are obscured. The people in all the experiments are the same, and do not change positions. We train the active speaker detection classifier in a Leave-One-Out-Cross-Validation (LOOCV) fashion, where the data from 6 presentations are used for training, and tested on the 7th presentation. This is repeated 7 times.

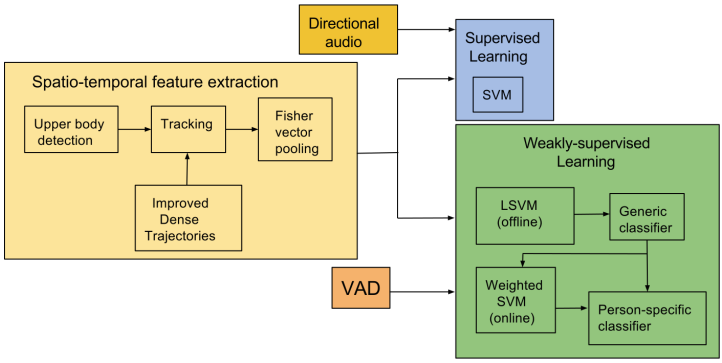

Fig. 2. Experimental setup

Finally, we test the model learnt on the Masters dataset on an entirely new dataset that we present - the Columbia dataset. It is an 87-minute-long video of a panel discussion at Columbia university, available from YouTube[2]. There are 7 speakers on the panel, and the camera focusses on smaller groups of speakers at a time. We only focus on the parts of the video where there is more than

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

one person in the frame, and ignore people on the margins of the video who are not detected by the upper body detector. This gives us sections of video for 5 speakers, with 2–3 speakers visible at any one time. We have annotated the upper body bounding boxes of each speaker with speak/non-speak labels, about 35 min of video in all, which are available at http://www.jaychakravarty.com/?page_id=432. We update the generic classifier learnt on the Masters dataset online, in a completely unsupervised fashion, with the generic classifier adapting to each new speaker in the Columbia dataset, with subsequent improvement in performance.

4.1 Implementation Details

We use the same improved trajectory features (ITF) [28] recommended by [5], for training our active speaker detection classifier. ITF are spatio-temporal features used for state of the art action recognition, and comprise of a concatenation of Histogram of Oriented Gradients (HOG), Histogram of Flow (HoF) and Motion Boundary Histogram (MBH) features. HoG, HoF and MBH features are calculated in the immediate neighbourhood of each point on the grid. We use 15 consecutive frames for calculating the ITF - this corresponds to about 7 s of video in the Masters dataset. The HoG, HoF and MBH features are independently reduced to half their original dimensions using PCA, and feature vectors from within an upper body track are pooled using Fisher vectors (FV) [29]. We apply intra-class L2 normalization, power and a final L2 normalization of the whole FV before classification using a linear SVM. We use a codebook size of 256 for the FV encoding. The FV encoding is done independently for HOG, HoF and MBH, before they are concatenated to a single, 101,376 dimensional vector. Intra-class L2 normalization - normalization within each block of the FV related to a single codeword, is used to balance weights of the different codewords in the FV, and reduces the "burstiness" in the FVs (often resulting from features belonging to the background). Training a linear SVM with a non-linear feature map (obtained using the power normalization) has the advantage of approximating a non-linear SVM at lower computational complexity [30]. These techniques, recommended as best practice in [31], have been to shown to considerably boost performance of FVs.

Table 1. Average AUC (with standard deviations) for active speaker detection fully supervised by directional audio [5], and weakly supervised by VAD, over all experimental folds (Masters dataset).

	Directional audio	VAD
Avg. AUC	$0.69_{\pm 0.07}$	$0.71_{\pm 0.05}$

For upper body detection, we use a detector trained using the Deformable Parts Model from [32]. The tracking is relatively straight-forward, because people don't change positions and there are no crossing tracks. ITF are grouped

by their start frame (calculated from the following 15 frames), and a FV is calculated for all the improved trajectories within a bounding box (person) track starting from that frame. A training sample is thus one FV pooling all ITFs from within an upper body track starting in a given frame, with each ITF covering 15 consecutive frames (about 7 s of video at 2 fps). The active speaker classifier is sensitive to the frame-rate of the dataset on which it is trained. To have the classifier transfer between datasets, we subsample the Columbia dataset so that its frame-rate matches the frame-rate of the Masters dataset (2 fps). A pipeline of the system is shown in Fig. 2.

4.2 Weak Supervision Using Audio

VAD results in frames with speak/non-speak labels. There are no speak/non-speak labels for individual bounding boxes and the FVs extracted from them. Section 3.1 details the Structured Output SVM classifier that is used for training the active speaker detection classifier in the absence of training labels for individual bounding boxes. Table 1 displays the results of our experiments with the active speaker detection classifier trained using VAD. The results with weak supervision (structured output learning) are comparable with the results from fully supervised learning from directional sound. This shows that the structured output formulation and the soft-max loss function for optimization transfers well from the object localization application of [8,9], to our task of active speaker localization in the absence of bounding box labels for training.

4.3 Speaker Specific Models

Section 3.2 makes the hypothesis that training person specific active speaker detection models will give better results than training a generic model for all speakers. To validate this hypothesis, we perform three experiments:

1. Full directional audio (giving speak/non-speak labels for all bounding boxes in the frame) for training the person specific classifier.
2. VAD audio (speak/non-speak label for the frame, but without information about individual bounding boxes) for training the person-specific classifier. This highest scoring sample using the generic classifier is used to get positive training samples for each person in a VAD-positive frame.
3. Experiment 2, with samples weighted according to temporal continuity (see Eq. 5).

When full directional audio supervision is available (expt. 1), the speaker specific models show better results, a 10 % improvement over the generic classifier of Table 1.

When using VAD for weak supervision with a hard-max posterior (expt. 2), the person-specific classifier performs worse (16 % worse mean average AUC) than the person-specific classifier with full audio supervision (expt. 1), and worse

Table 2. Mean Avg AUC (with standard deviations) for person-specific active speaker detection using (1) directional audio, (2) VAD - no temporal weighting & (3) VAD with temporally weighted samples (All expts. on Masters dataset).

Expt.	Speaker 1	Speaker 2	Speaker 3	Mean Avg. AUC
1	0.79 ± 0.08	0.76 ± 0.03	0.88 ± 0.05	0.81 ± 0.07
2	0.60 ± 0.10	0.59 ± 0.07	0.75 ± 0.03	0.65 ± 0.10
3	0.79 ± 0.10	0.80 ± 0.03	0.88 ± 0.04	0.82 ± 0.07

even than the generic classifier. This confirms the dataset bias problem we discussed in Sect. 2. The generic classifier might be more biased towards one speaker compared to the others and occasionally score the true positive speaker lower than another non-speaker in the same VAD-positive frame. This leads to the use of mis-classified samples for the training of the person-specific classifiers in the weakly supervised case, and their subsequent poor performance.

In experiment 3, a temporal weight is added to each sample - the number of contiguous neighbouring frames in which it has been consistently labelled (see Eq. 5). We use a temporal window of 3 s. This results in a mean average AUC of 0.82, comparable to the fully supervised case (expt. 1). This shows that the temporal weighting of samples correctly guides the sample selection. Thus, it acts as another weak supervisor (apart from the VAD) for the training of the speaker specific classifer. Table 2 presents results for all 3 speaker-specific experiments in the Masters dataset. It should be noted here that for all experiments in this sub-section, the evaluations are performed on individual frames and temporal continuity is exploited as an extra cue during training, not as a postprocessing step to correct results afterwards.

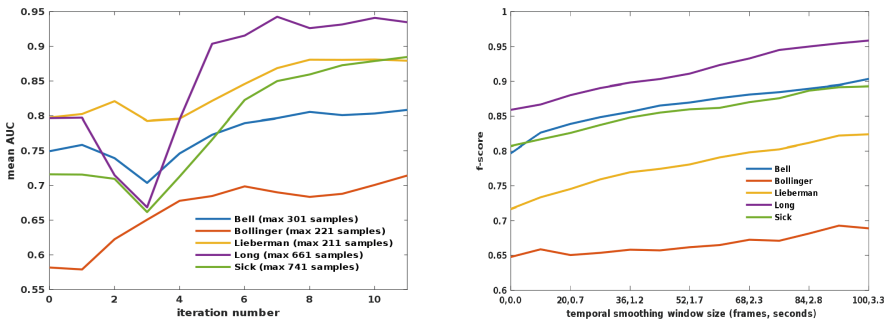

Fig. 3. (a) Online Learning: Mean AUC over all speakers in the new Columbia dataset with an increasing number of training samples in each iteration. (b) Temporal smoothing: F-scores for all speakers at the end of online learning, after thresholding and temporal smoothing, with increasing size of temporal window in the Columbia datset.

4.4 Online Learning

Here, we report results of experiments that demonstrate how a generic classifier trained on speakers in the Masters dataset, can be modified online, to specific speakers in the Columbia dataset. We only select sections of video in which there are 2 or more people in the frame at the same time. This is to demonstrate the unsupervised selection of training samples from one among many speakers. The selection of training samples when only 1 speaker is present in the frame is trivial (VAD can be used to detect positive and negative samples for the speaker), and is not considered in this experiment.

The prior classifier is run on each VAD-positive frame in the new dataset and the highest scoring bounding box is taken to be the positive sample for that speaker in the frame, and the remaining bounding boxes are taken to be the negative samples for the other speakers. This assumes that there is only one person speaking at a time in the video, which is actually the case in most target applications. The samples are weighted according to their temporal continuity - a positive sample with a higher number of contiguous positive samples around it gets a higher weight, as was done in number 3 of the speaker specific experiments (Subsect. 4.3). The experiment begins by using the prior classifier to detect active speakers in the new data. Then, with each iteration of online learning, a balanced selection of positive and negative samples are selected from each speaker, and used for training the person-specific classifier. The number of training samples increases with each iteration. Figure 3a displays the mean average AUC results for experiments conducted per speaker over the training iterations. We see that the performance of the iteratively trained person specific target classifier starts out at the performance of the generic source classifier, and gradually improves with increasing number of target training samples. There is an initial dip in the performance of the classifier learnt online for 3 of the 5 people, when there is a small number of training samples. If some of these samples are wrongly selected by the prior classifier, then the classifier's performance will decrease to a level below the generic classifier performance. But, as the speaker speaks for longer, and more correct samples weighted by their temporal continuity are picked, the online learning adapts to the target distribution. We use a maximum of 10 s of video per person for the online learning in the Columbia dataset in our experiments, and see an improvement of about 5–15% over the performance of the prior classifier. Thus, our method of selecting samples weighted by their temporal continuity is resilient against the selection of some wrong samples, and very quickly - within a few seconds - adapts to each new speaker. We use temporal continuity to further improve the performance of the online-learnt classifier during inference as well. The scores from the classifier learnt during the last iteration of online learning are thresholded (at the intersection of the ROC curve with the diagonal) and smoothed over increasing lengths of time (from 0 to about 3 s). Figure 3b shows that the f-scores for all the speakers improve with increasing amounts of temporal smoothing, with plateauing of results at around 3 s. A potential downside of too much smoothing is that if a person speaks for short durations (single, yes/no utterances for example),

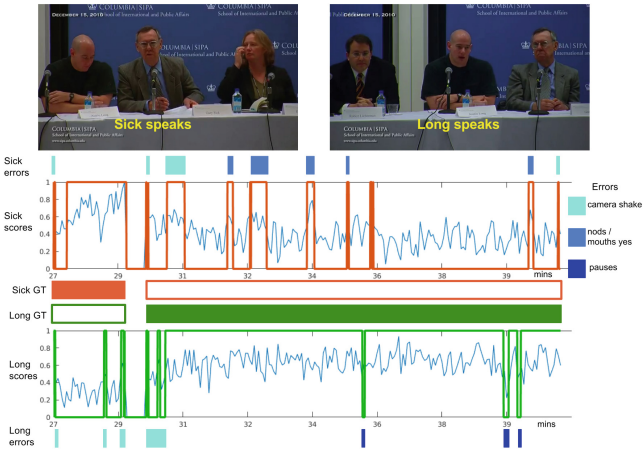

Fig. 4. Normalized raw scores (blue) with the online-learnt classifier and thresholded and temporally smoothed speak/non-speak values for speakers Sick (red) and Long (green), along with Ground Truth (GT, solid colour = speak), in minutes 27:00 to 40:00 in the Columbia dataset. (Color figure online)

then these are not going to be registered. The amount of temporal smoothing applied would depend on the application. For video conferencing, it might not be appropriate to switch focus between speakers for such short utterances, and a smoothing of 3 s (the maximum smoothing applied in our experiments during inference), would probably be adequate.

Figure 4 shows a timeline for Active Speaker Detection in the Columbia dataset, for speakers Sick and Long, during minutes 27:00 to 40:00 in the video. The classifiers for these speakers are learnt online earlier in the video, and the raw scores for these speakers over time are shown in blue. The scores are thresholded and temporally smoothed to obtain speak/non-speak values, shown in red and green for Sick and Long respectively. Ground truth speak/non-speak values for these speakers are also given. It can be seen that the parts of the video where the algorithm apparently makes a mistake can be explained by camera shake, or where a non-active speaker actually nods and mouths yes in response to another active speaker (ground truth does not mark this as speech), or when an active speaker pauses mid-sentence.

5 Conclusions

This paper demonstrates the use of audio for cross-modal supervision of the training of a video-based active speaker detector. The problem is posed in terms of a structured output prediction problem - given information about the presence of an active speaker in a frame from audio-based Voice Activity Detection, find out which particular person is speaking, among the people in the frame, and at the same time, learn the video-based classifier for active speaker detection.

Person-specific classifiers are shown to perform better than generic classifiers, and the learning of the specific classifiers is again weakly supervised by audio. The prior classifier adapts to the specific speaker using samples from just a few seconds of video, with additional improvement in results using temporal smoothing. This shows that the system has the potential to be used in a video conferencing application, and quickly learn the characteristics of new speakers.

In future work, we will close the loop between audio and video. In current work, audio supervises the learning of a video-based person-specific active speaker detector. The learnt video classifier will in turn supervise the learning of person-specific voice models and those voice models will be fed back into the video to further improve active speaker detection. This is expected to be particularly useful in the more challenging data encountered in video diarization: movies and TV series with non-frontal views of people, where the video-only classifier is expected to perform worse than in frontal-view video.

References

1. Khoury, E., Sénac, C., Joly, P.: Audiovisual diarization of people in video content. Multimedia Tools Appl. **68**(3), 747–775 (2014)
2. Everingham, M., Sivic, J., Zisserman, A.: Hello! my name is... buffy"-automatic naming of characters in tv video. In: BMVC, vol. 2, pp. 6 (2006)
3. Everingham, M., Sivic, J., Zisserman, A.: Taking the bite out of automatic naming of characters in TV video. Image Vis. Comput. **27**(5), 545–559 (2009)
4. Haider, F., Al Moubayed, S.: Towards speaker detection using lips movements for humanmachine multiparty dialogue. In: 2012 FONETIK (2012)
5. Chakravarty, P., Mirzaei, S., Tuytelaars, T., Vanhamme, H.: Who's speaking? audio-supervised classification of active speakers in video. In: ACM International Conference on Multimodal Interaction (ICMI) (2015)
6. Germain, F., Sun, D.L., Mysore, G.J.: Speaker and noise independent voice activity detection. In: INTERSPEECH, pp. 732–736 (2013)
7. Bilen, H., Namboodiri, V.P., Gool, L.J.: Object and action classification with latent window parameters. Int. J. Comput. Vis. **106**(3), 237–251 (2014)
8. Bilen, H., Pedersoli, M., Tuytelaars, T.: Weakly supervised object detection with posterior regularization. In: British Machine Vision Conference (2014)
9. Bilen, H., Pedersoli, M., Tuytelaars, T.: Weakly supervised object detection with convex clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1081–1089 (2015)
10. Deselaers, T., Alexe, B., Ferrari, V.: Weakly supervised localization and learning with generic knowledge. Int. J. Comput. Vis. **100**(3), 275–293 (2012)
11. Song, H.O., Girshick, R., Jegelka, S., Mairal, J., Harchaoui, Z., Darrell, T.: On learning to localize objects with minimal supervision. arXiv preprint arXiv:1403.1024 (2014)
12. Nguyen, M.H., Torresani, L., de la Torre, F., Rother, C.: Weakly supervised discriminative localization and classification: a joint learning process. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 1925–1932. IEEE (2009)
13. Bojanowski, P., Bach, F., Laptev, I., Ponce, J., Schmid, C., Sivic, J.: Finding actors and actions in movies. In: 2013 IEEE International Conference on Computer Vision (ICCV), pp. 2280–2287. IEEE (2013)

14. Khosla, A., Zhou, T., Malisiewicz, T., Efros, A.A., Torralba, A.: Undoing the damage of dataset bias. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7572, pp. 158–171. Springer, Heidelberg (2012). doi:10.1007/978-3-642-33718-5_12
15. Tommasi, T., Quadrianto, N., Caputo, B., Lampert, C.H.: Beyond dataset bias: multi-task unaligned shared knowledge transfer. In: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) ACCV 2012, Part I. LNCS, vol. 7724, pp. 1–15. Springer, Heidelberg (2013)
16. Aljundi, R., Emonet, R., Muselet, D., Sebban, M.: Landmarks-based kernelized subspace alignment for unsupervised domain adaptation. In: Computer Vision and Pattern Recognition (CVPR 2015) (2015)
17. Fernando, B., Habrard, A., Sebban, M., Tuytelaars, T.: Unsupervised visual domain adaptation using subspace alignment. In: 2013 IEEE International Conference on Computer Vision (ICCV), pp. 2960–2967. IEEE (2013)
18. Aytar, Y., Zisserman, A.: Tabula rasa: model transfer for object category detection. In: 2011 IEEE International Conference on Computer Vision (ICCV), pp. 2252–2259. IEEE (2011)
19. Tommasi, T., Caputo, B.: The more you know, the less you learn: from knowledge transfer to one-shot learning of object categories. In: BMVC, Number LIDIAP-CONF-2009-049 (2009)
20. Tommasi, T., Orabona, F., Caputo, B.: Safety in numbers: learning categories from few examples with multi model knowledge transfer. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3081–3088. IEEE (2010)
21. Chen, J., Liu, X., Tu, P., Aragones, A.: Person-specific expression recognition with transfer learning. In: 2012 19th IEEE International Conference on Image Processing (ICIP), pp. 2621–2624. IEEE (2012)
22. Chu, W.S., De la Torre, F., Cohn, J.F.: Selective transfer machine for personalized facial action unit detection. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3515–3522. IEEE (2013)
23. Zen, G., Sangineto, E., Ricci, E., Sebe, N.: Unsupervised domain adaptation for personalized facial emotion recognition. In: Proceedings of the 16th International Conference on Multimodal Interaction, pp. 128–135. ACM (2014)
24. Gavves, E., Mensink, T., Tommasi, T., Snoek, C.G., Tuytelaars, T.: Active transfer learning with zero-shot priors: reusing past datasets for future tasks. arXiv preprint arXiv:1510.01544 (2015)
25. Mirzaei, S., Van hamme, H., Norouzi, Y.: Blind audio source separation of stereo mixtures using bayesian non-negative matrix factorization. In: Signal Processing Conference (EUSIPCO), pp. 621–625, September 2014
26. Pletscher, P., Ong, C.S., Buhmann, J.M.: Entropy and margin maximization for structured output learning. In: Balcázar, J.L., Bonchi, F., Gionis, A., Sebag, M. (eds.) ECML PKDD 2010. LNCS (LNAI), vol. 6323, pp. 83–98. Springer, Heidelberg (2010). doi:10.1007/978-3-642-15939-8_6
27. Uijlings, J.R., Sande, K.E., Gevers, T., Smeulders, A.W.: Selective search for object recognition. Int. J. Comput. Vis. **104**(2), 154–171 (2013)
28. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: ICCV, Sydney, Australia, pp. 3551–3558, December 2013
29. Perronnin, F., Sánchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 143–156. Springer, Heidelberg (2010)
30. Vedaldi, A., Zisserman, A.: Efficient additive kernels via explicit feature maps. IEEE Trans. Pattern Anal. Mach. Intell. **34**(3), 480–492 (2012)

31. Peng, X., Wang, L., Wang, X., Qiao, Y.: Bag of visual words and fusion methods for action recognition: comprehensive study and good practice. CoRR abs/1405.4506 (2014)
32. Girshick, R.B., Felzenszwalb, P.F., McAllester, D.: Discriminatively trained deformable part models, release 5. http://people.cs.uchicago.edu/rbg/latent-release5/