# Group-Level Focus of Visual Attention for Improved Active Speaker Detection

Chris Birmingham
cbirming@usc.edu
University of Southern California
Los Angeles, California, USA

Maja J Matarić
mataric@usc.edu
University of Southern California
Los Angeles, California, USA

Kalin Stefanov
kalin.stefanov@monash.edu
Monash University
Melbourne, Victoria, Australia

## ABSTRACT

This work addresses the problem of active speaker detection in physically situated multiparty interactions. This challenge requires a robust solution that can perform effectively across a wide range of speakers and physical contexts. Current state-of-the-art active speaker detection approaches rely on machine learning methods that do not generalize well to new physical settings. We find that these methods do not transfer well even between similar datasets. We propose the use of group-level focus of visual attention in combination with a general audio-video synchronizer method for improved active speaker detection across speakers and physical contexts. Our dataset-independent experiments demonstrate that the proposed approach outperforms state-of-the-art methods trained specifically for the task of active speaker detection.

## CCS CONCEPTS

• **Computing methodologies** → Neural networks; Supervised learning; Unsupervised learning; • **Human-centered computing** → Collaborative interaction; Laboratory experiments.

## KEYWORDS

active speaker detection; focus of visual attention; neural networks

**ACM Reference Format:**
Chris Birmingham, Maja J Matarić, and Kalin Stefanov. 2021. Group-Level Focus of Visual Attention for Improved Active Speaker Detection. In *Companion Publication of the 2021 International Conference on Multimodal Interaction (ICMI '21 Companion), October 18–22, 2021, Montréal, QC, Canada.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3461615.3485430

## 1 INTRODUCTION

Understanding the roles on the conversational floor (*i.e.*, speaker, addressee, bystander), known as *footing* [12, 13], is a prerequisite for natural and effective verbal interaction. Therefore, to successfully and fluently participate in a situated, multiparty human-machine conversation, a system must understand those roles. *Active speaker detection* is the task of identifying the current speaker (if any) from a set of candidate speakers. It is necessary for recognizing who is talking and for attributing any thoughts, ideas, and opinions to the speaker.


*ICMI '21 Companion, October 18–22, 2021, Montréal, QC, Canada*

ACM ISBN 978-1-4503-8471-1/21/10.
https://doi.org/10.1145/3461615.3485430

Although most humans can perform active speaker detection with relative ease, machines struggle to do so accurately. This task is inherently multimodal, requiring an accurate synthesis of and reasoning about visual, auditory, and linguistic information. In physically situated interactions, this challenge is amplified by sensor limitations, *e.g.*, monocular cameras and far-field microphones. Additionally, natural conversations can be noisy, with overlaps, cut-ins, and backchannels that can blur the distinction between the active speaker and other group members.

Due to these challenges, we hypothesise that it is helpful to incorporate information beyond the candidate speaker's visual and auditory data alone. Such information can include objects of interest in the environment or, in the case of multiparty interactions, it can include information from other group members, such as their focus of visual attention. In this work we propose to utilize group members' focus of visual attention to improve the performance of state-of-the-art active speaker detection methods. The main contributions of this work are as follows.

- We evaluate state-of-the-art methods for the task of active speaker detection in situated multiparty interactions.
- We analyse some of the conditions under which these methods fail when used in situated multiparty interactions.
- We propose and evaluate methods that utilize group members' focus of visual attention to address some of these shortcomings.

## 2 BACKGROUND

Active speaker detection (ASD) is the task of determining if a certain speaker is active at any point in time. In clean acoustic conditions, and with single speaker, the acoustic information is fundamental for the ASD task, and methods for **audio-only** ASD have been extensively studied. Anguera et al. [2] and Tranter and Reynolds [28] offer comprehensive reviews of the research in this field. Audio-only ASD systems usually suffer from noisy environments, far-field microphones, and speakers that overlap in time. Additionally, audio-only approaches are limited in multiparty interactions, where it is important to assign the detection to speakers that might be physically close. **Video-only** methods attempt to directly model the face, *e.g.*, [1, 27], or some aspects of the face (*e.g.*, lip movements [20]). The drawbacks of these methods are related to motions, *e.g.*, facial expressions, that can be misinterpreted as speaking.

**Audio-visual** methods combine information from both the audio and visual modalities; complementing the audio approach with its video counterpart generally produces better performance due to increased robustness [6, 7, 11, 16, 18]. Recently, researchers have employed Artificial Neural Networks for ASD from audio-visual input. A multimodal Long Short-Term Memory (LSTM) model that

learns shared weights between modalities was proposed in [18]. A combination of a pre-trained Convolutional Neural Network (CNN) model used as image encoder and an LSTM model used as classifier was presented in [24]. Stefanov et al. [25] proposed a self-supervised method in the context of language acquisition. Hu et al. [14] proposed a CNN model that learns the fusion of face and audio information.

**Other** approaches to ASD include a general pattern recognition framework used by Besson and Kunt [4]. Visual activity (the amount of movement) and the focus of visual attention were used as inputs by Hung and Ba [15]. Stefanov et al. [27] used facial action units as inputs to Hidden Markov Models and Vajaria et al. [29] demonstrated that information from body movements can improve detection performance.

# 3 METHODOLOGY

Given a number of candidate speakers, active speaker detection (ASD) consists of the task of determining, at any point in time, what speakers are active using information from that point in time. This is a binary classification problem (per candidate speaker) where the input is the part of the image capturing the candidate speaker and the associated audio. The backbone model used in this work is the state-of-the-art audio-video *synchronizer* [9, 10, 21]. Following the state-of-the-art in ASD [8], the synchronizer is turned into *detectors* by training a Temporal Convolution Network (TCN) and a Bidirectional Long Short-Term Memory (BLSTM). We propose novel methods to augment the synchronizer with information for the group members' focus of visual attention for improved cross-dataset ASD.

## 3.1 Datasets

The recent introduction of the AVA-ActiveSpeaker dataset [19] consisting of nearly 40 hours of video data from movies has allowed for benchmarking different methods for ASD. However, the use of data from movies does not support the development of ASD methods for physically situated interactions. To address this issue, we used two private multiparty interaction datasets described next.

*3.1.1 Robot-Facilitated Support Group Dataset (RFSG).* The robot-facilitated support group dataset [5] consists of 27 multiparty interactions between three students and a robot. The robot lead the support group by asking questions and making disclosures to encourage the human members of the group to share and receive support, using the set up shown in Figure 1a. The robot's utterances were selected from a predefined set of questions and statements by a human "wizard". The total number of participants in the dataset is 81. The average duration of the interactions is 20 minutes, resulting in a total of 10 hours of data per recording device. The active speaker labels were obtained by manual annotations. In this work we consider the color video stream generated by the camera pointed at each participant and the audio stream generated by a single microphone in the middle of the table.

*3.1.2 Focus of Visual Attention Dataset (FOVA).* We also analyzed the multimodal multiparty dataset described in Stefanov and Beskow [23]. Each interaction consisted of three participants: one moderator and two interactants, using the spatial configuration shown in Figure 1b. A total of 15 sessions were recorded, each lasting approximately 30 minutes, resulting in 7.5 hours of data per recording device. The moderator was the same in all interactions, while the other participants varied, totalling 24 unique participants. The active speaker labels were obtained by manual annotations. In this work we consider the color video stream generated by the Kinect RGB-D camera pointed at each participant and the audio stream generated by the participants' close-talking microphone.

## 3.2 Experimental Setup

We evaluated the performance of the ASD methods with three experiments: within dataset speaker-dependent (10-fold cross-validation), within dataset speaker-independent (leave-6-out cross-validation), and cross-dataset speaker-independent (train on one dataset and test on the other). The within-dataset speaker-dependent experiment trained models with data for all participants and evaluated them on independent data from all participants in the same dataset. The within-dataset speaker-independent experiment trained models with data for a subset of participants and evaluated them on the left-out participants in the same dataset. This experiment tested the transferability (generalization capabilities) of the models to unseen participants from the same physical context. The cross-dataset speaker-independent experiment used all data from one dataset to train the models, and all data from the other dataset to evaluate them. This experiment tested the transferability (generalization capabilities) of the models to both unseen participants and physical contexts. The experiments directly demonstrated the contribution of the work in terms of error analysis and proposed strategies to address the identified shortcomings of the ASD methods.

*3.2.1 Features.* We present experiments with two sets of features termed PerfectMatch and VisualAttention.

**PerfectMatch** – in Chung and Zisserman [9], the authors trained a Convolutional Neural Network termed SyncNet for the purpose of synchronizing audio and video tracks of individuals talking. The network consisted of 6 convolutional layers followed by 2 fully connected layers for both audio and video, separately. The model takes as input 5 video frames and the corresponding audio samples. This model was shown to be effective for ASD by comparing the magnitude of the difference between the final audio and video features. In Chung et al. [10], the authors employed a different strategy for training the original SyncNet model. The new model termed PerfectMatch was shown to outperform the original SyncNet model. In our experiments, we used the features of the final convolutional layer from the PerfectMatch model that has been trained on the VoxCeleb dataset [17].

**VisualAttention** – we used visual attention features inspired by Stefanov et al. [26]. We implemented both binary and continuous representations of whether any of the other group members are looking at the candidate speaker. To measure the direction of the visual attention of each group member, we used the $3D$ position and orientation of the head. For the FOVA dataset, an RGB-D Kinect camera with calibrated position and orientation was used to acquire those measures. For the RFSG dataset, the measures were approximated through OpenFace [3] from cameras with calibrated positions and orientations. For both datasets, the position and orientation of the participants' head was recreated in the same $3D$

(a) Robot-Facilitated Support Group.

(b) Focus of Visual Attention.

**Figure 1: Spatial configuration of the sensors and participants in the datasets.**

space for each video frame where the head position and orientation was used to create a vector of visual attention. The binary representation creates a cylinder around the candidate speaker's head and judges a group member to be looking at that person if the head pose vector of the group member intersects with the cylinder. We used 5× the average male head size for the dimensions of the cylinder. The continuous representation measures the angle between the group member's head pose and the vector from the group member's head to the candidate speaker's head. The measured angle is mapped to a range between 0 and 1, in which an angle of 0 degrees (the group member is looking directly at the candidate speaker) is 1 and the angle of 75 degrees (the group member is looking away from the candidate speaker) is 0.

**Augmentation** – we combined the VisualAttention features for a candidate speaker by averaging the features produced by each of the other group members. This produced a single value for both the binary and continuous features. In both cases, the closer this value was to 1, the more likely the candidate speaker was the focus of attention; conversely, the closer the value was to 0, the less likely it was that they are the focus of attention. To augment the detection of a given model, which is also between 0 and 1, we implemented three preliminary methods: 1) we combined the VisualAttention feature with the model detection in an unweighted average, 2) we combined them in a weighted average skewed towards the VisualAttention feature, and 3) we multiplied the model detection value with a VisualAttention feature shifted from the range $[0, 1]$ to the range $[0.5, 1.5]$ in order to allow the feature to modify the model detection.

*3.2.2 Models.* For each experiment, we independently evaluated the synchronizer, detectors, and augmented synchronizer in order to compare the models and provide evidence for some of the situations in which they fail.

The **synchronizer** was implemented as described in [10]. This included an input of 5 video frames and the corresponding audio, resulting in $1024D$ audio and video features. These features were cross-correlated to find the minimum difference (the matching synchronization), then the difference between the features was median-filtered and normalized for each video. This produced a single value in the range $[0, 1]$ that was used as the detection of whether the candidate speaker was speaking. For the synchronizer we used the state-of-the-art pre-trained model.

The **detectors** were implemented as described in [8]. A window of 5 of the $512D$ audio and video features was used as input to a 2-layer time series model before being combined in a fully connected layer. The fully connected layer was then connected to a softmax layer to produce the final output probabilities. As in [8], we tested both a Temporal Convolution Network and a Bidirectional Long Short-Term Memory as the time series model. The detectors were trained as described in [8]. The input features from the synchronizer were held constant and the models were trained on the FOVA and RFSG datasets individually for each experiment.

The **augmented** models were implemented by combining the detection of the synchronizer with the VisualAttention features using the three different augmentation methods.

*3.2.3 Evaluation.* Each of the models was evaluated on an unseen test set. In the case of the synchronizer, the model was trained on the VoxCeleb dataset and was not fine-tuned. We report the result of using the model as-is on the RFSG and FOVA datasets. In the case of the detectors, we report the result of within dataset speaker-dependent, within dataset speaker-independent, and cross-dataset speaker-independent models separately. We follow the reporting requirements for the AVA-ActiveSpeaker dataset in reporting the mean average precision (mAP) scores for each model. Average precision (AP) is the average of precision scores calculated for each recall threshold, $\text{AP} = \sum_n (R_n - R_{n-1}) P_n$, where R is the recall and P is the precision for each threshold n. The results are reported in terms of frame-by-frame mAP, mAP = (AP0 + AP1)/2, where AP0 and AP1 are the AP score of the negative and positive class, respectively.

| Model | Speaker-dependent | | Speaker-independent | | Cross-dataset | |
|---|---|---|---|---|---|---|
| | RFSG | FOVA | RFSG | FOVA | RFSG | FOVA |
| BLSTM | **0.973** (0.006) | **0.987** (0.002) | **0.894** (0.028) | 0.975 (0.004) | 0.663 (0.022) | 0.685 (0.012) |
| TCN | 0.966 (0.004) | 0.986 (0.002) | 0.891 (0.038) | **0.976** (0.002) | 0.638 (0.024) | **0.702** (0.011) |
| PerfectMatch | - | - | - | - | **0.807** | 0.677 |

**Table 1: Performance of the synchronizer and detectors. Mean mAP and standard deviation in parenthesis.**

## 4 RESULTS

The results from all experiments are reported in Table 1, with the mean and standard deviation of the models' mAP. The results from the **speaker-dependent** experiment are based on 10-fold cross-validation. The best performing detector was the BLSTM. This detector achieved a score of 0.973 on the RFSG dataset and 0.987 on the FOVA dataset. Both detectors performed similarly well on both datasets. The results from the **within-dataset speaker-independent** are based on leave-6-out cross-validation. The BLSTM detector performed better on the RFSG dataset, reaching the score of 0.894 while the TCN detector performed better on the FOVA dataset, with score of 0.976. Both detectors performed similarly within dataset, however there was a large difference between datasets, where the performance on the FOVA dataset was 0.08 better. The results from the **cross-dataset speaker-independent** experiment are based on 10-fold cross-validation for the detectors and include the performance of the synchronizer on each dataset, reported as a single value. The PerfectMatch synchronizer performed best on the RFSG dataset, with a score of 0.807. The detectors that had been fine-tuned on FOVA performed significantly worse, with the mean score of 0.651. On the FOVA dataset, the TCN detector performed best, with the score of 0.702.

### 4.1 Error Analysis

In order to investigate the decrease in performance on the challenging task of cross-dataset speaker-independent detection, we evaluated the cross-dataset models for different head poses of the candidate speaker. In Figures 2a and 2c we report the performance of the models across different head poses for both datasets. The candidate speaker's head pose was gathered on a per-frame basis into 6 buckets spanning 20 degrees each. The performance of each model was then calculated for all frames in each bucket. Across all models the performance formed an inverted relationship with the distance from the center, which is to say that the performance tended to get worse as the candidate speaker looked further away from the camera (*i.e.*, profile faces). Although not always the case, this relationship can be seen from the general concave shapes in the figures. This observation further motivates the introduction of information that is independent from the candidate speaker for the task of ASD. In the next section we present the relative improvement when information for the group-level focus of visual attention is incorporated into the detection result.

### 4.2 Augmentation Improvement

On the RFSG dataset, the best performing augmented model resulted from combining the synchronizer detection with the continuous focus of visual attention feature through multiplication. This yielded a mAP score of 0.850. For each of the detectors on this dataset, the multiplication with the continuous focus of visual attention feature yielded an improvement. For the FOVA dataset, the best performing augmented model involved combining the synchronizer detection with the continuous focus of visual attention feature through weighted averaging. This yielded a mAP score of 0.861. There was no improvement when augmentation was applied to the detectors.

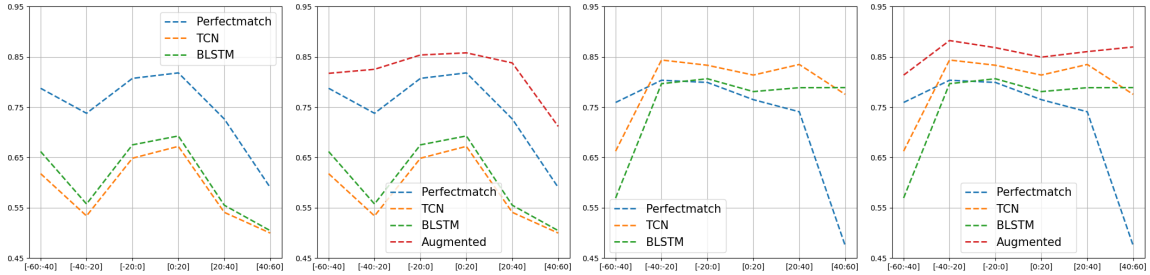
Figures 2b and 2d illustrate how each of the models performed across the different head poses. We included the original synchronizer, TCN and BLSTM detectors, and the augmented synchronizer. The results show that the augmented model performed better across the full range of head poses, and partially managed to correct for the typical deterioration of performance that happened when the head pose is at an extreme angle with respect to the camera plane.

## 5 DISCUSSION

As can be seen in Table 1, the state-of-the-art detectors performed well in speaker-dependent 10-fold and speaker-independent leave-6-out cross-validation experiments on both datasets. Both datasets provide well-posed problems, with cameras pointed at participants' faces and clear audio during a normal conversation. However, when these fine-tuned detectors are applied to a new setting (cross-dataset), we observe a significant decrease in performance. Even though both datasets consist of seated conversations around a table, the detectors appear to learn the distribution specific to the context present in the dataset used for fine-tuning. Although this is a well understood limitation of machine learning, it presents a significant challenge for creating accurate ASD methods that can generalize across environments and physical contexts.

Given prior work on spatial bias [22] in vision-based voice activity detection, we investigated how spatial bias hampered the performance of these detectors when transferred to new physical contexts. As expected, we found that the models often performed worse when the faces were pointed away from the camera. This occurs because detection is not as good and fewer training examples exist for faces seen in profile.

To improve the performance of the models, we utilized a feature that is independent from the candidate speaker's head pose. We found that the focus of visual attention of the other group members could be used to augment the output of the synchronizer and detectors, improving the cross-dataset performance in almost all cases. Furthermore, we found a significant improvement of the models' performance when combined with the focus of visual attention of the other group members across the entire spectrum of head poses. However, when the head is turned away from the camera,




(a) Models on RFSG (b) Augmented model on RFSG (c) Models on FOVA (d) Augmented model on FOVA

**Figure 2: Performance of all cross-dataset speaker-independent models. The $Y$-axis is the models' mAP for different head poses in the range $[-60, 60]$ degrees. (a) and (c) show the performance the original detectors and synchronizer on each dataset. (b) and (d) show the improved performance of the synchronizer augmented with VisualAttention features.**

the increase is larger than when the candidate speaker is facing the camera.

Despite their similarities, the RFSG and FOVA datasets have significant differences due to their set ups and physical arrangement. The distribution of head poses for the FOVA dataset is bi-modal, caused by the seating of 3 people evenly around a circular table. The RFSG distribution is a steep unimodal curve, likely caused by the shape of the table and the location of the robot on that table. The fine-tuned detectors are well suited to the distributions of the respective datasets, leading to a poor fit when faced with the challenge of transferring to a new dataset. However, this type of distribution shift is expected every time the ASD models are employed in new interactions.

## 6 CONCLUSION & FUTURE WORK

In this work we show how a simple group-level focus of visual attention feature can improve the performance of a general purpose synchronizer to above the level of fine-tuned detectors on the task of context- and person-independent active speaker detection. We show that, even when employed on similar datasets, fine-tuned detectors struggle to generalize well. We demonstrate that spatial bias can contribute to performance degradation and offer an investigation of the possible causes and remedies for this important problem. The proposed methods utilized simple but effective approaches such as averaging and multiplying for augmentation. Our future work will explore improved ways for combining the detection and group-level focus of visual attention of the models. We will also benchmark our approach to general active speaker detection on public datasets.

## ACKNOWLEDGMENTS

This work was supported by the National Science Foundation Expeditions in Computing IIS-1925083.

## REFERENCES

[1] R. Ahmad, S. P. Raza, and H. Malik. 2013. Visual Speech Detection Using an Unsupervised Learning Framework. In *Proceedings of the International Conference on Machine Learning and Applications*, Vol. 2. 525–528.

[2] X. Anguera, N. Bozonnet, S.and Evans, C. Fredouille, G. Friedland, and O. Vinyals. 2012. Speaker Diarization: A Review of Recent Research. *IEEE Transactions on Audio, Speech, and Language Processing* 20, 2 (2012), 356–370.

[3] T. Baltrusaitis, P. Robinson, and L. P. Morency. 2016. OpenFace: An Open Source Facial Behavior Analysis Toolkit. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*. 1–10.

[4] P. Besson and M. Kunt. 2008. Hypothesis Testing for Evaluating a Multimodal Pattern Recognition Framework Applied to Speaker Detection. *Journal of Neuro-Engineering and Rehabilitation* 5, 1 (2008), 11.

[5] C. Birmingham, Z. Hu, K. Mahajan, E. Reber, and M. J. Matarić. 2020. Can I Trust You? A User Study of Robot Mediation of a Support Group. In *Proceedings of the IEEE International Conference on Robotics and Automation*. 8019–8026.

[6] P. Chakravarty, S. Mirzaei, T. Tuytelaars, and H. Van Hamme. 2015. Who's Speaking? Audio-Supervised Classification of Active Speakers in Video. In *Proceedings of the ACM on International Conference on Multimodal Interaction*. 87–90.

[7] P. Chakravarty and T. Tuytelaars. 2016. Cross-Modal Supervision for Learning Active Speaker Detection in Video. In *Proceedings of the European Conference on Computer Vision*. 285–301.

[8] J. S. Chung. 2019. Naver at ActivityNet Challenge 2019 – Task B Active Speaker Detection (AVA). *arXiv preprint arXiv:1906.10555* (2019).

[9] J. S. Chung and A. Zisserman. 2016. Out of Time: Automated Lip Sync in the Wild. In *Proceedings of the Workshop on Multi-view Lip-reading*.

[10] S.-W. Chung, J. S. Chung, and H.-G. Kang. 2019. PerfectMatch: Improved Cross-modal Embeddings for Audio-visual Synchronisation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. 3965–3969.

[11] R. Cutler and L. Davis. 2000. Look Who's Talking: Speaker Detection Using Video and Audio Correlation. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, Vol. 3. 1589–1592.

[12] E. Goffman. 1974. *Frame Analysis: An Essay on the Organization of Experience*. Harvard University Press.

[13] E. Goffman. 1981. *Forms of Talk*. University of Pennsylvania Press.

[14] Y. Hu, J. Ren, J. Dai, C. Yuan, L. Xu, and W. Wang. 2015. Deep Multimodal Speaker Naming. In *Proceedings of the ACM International Conference on Multimedia*. 1107–1110.

[15] H. Hung and S. O. Ba. 2009. *Speech/Non-Speech Detection in Meetings From Automatically Extracted Low Resolution Visual Features*. Technical Report. Idiap.

[16] V. P. Minotto, C. R. Jung, and B. Lee. 2014. Simultaneous-Speaker Voice Activity Detection and Localization Using Mid-Fusion of SVM and HMMs. *IEEE Transactions on Multimedia* 16, 4 (2014), 1032–1044.

[17] A. Nagrani, J. S. Chung, and A. Zisserman. 2017. VoxCeleb: A Large-Scale Speaker Identification Dataset. In *Proceedings of the Annual Conference of the International Speech Communication Association*.

[18] J. Ren, Y. Hu, Y.-W. Tai, C. Wang, L. Xu, W. Sun, and Q. Yan. 2016. Look, Listen and Learn - A Multimodal LSTM for Speaker Identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 3581–3587.

[19] J. Roth, S. Chaudhuri, O. Klejch, R. Marvin, A. Gallagher, L. Kaver, S. Ramaswamy, A. Stopczynski, C. Schmid, Z. Xi, and C. Pantofaru. 2019. AVA-ActiveSpeaker: An Audio-Visual Dataset for Active Speaker Detection. *arXiv preprint arXiv:1901.01342* (2019).

[20] S. Siatras, N. Nikolaidis, M. Krinidis, and I. Pitas. 2009. Visual Lip Activity Detection and Speaker Detection Using Mouth Region Intensities. *IEEE Transactions on Circuits and Systems for Video Technology* 19, 1 (2009), 133–137.

[21] J. S. Son and A. Zisserman. 2017. Lip Reading in Profile. In *Proceedings of the British Machine Vision Conference*. 1–11.

[22] K. Stefanov, M. Adiban, and G. Salvi. 2021. Spatial Bias in Vision-Based Voice Activity Detection. In *Proceedings of the International Conference on Pattern Recognition*. 10433–10440.

[23] K. Stefanov and J. Beskow. 2016. A Multi-Party Multi-Modal Dataset for Focus of Visual Attention in Human-Human and Human-Robot Interaction. In *Proceedings of the International Conference on Language Resources and Evaluation*.

[24] K. Stefanov, J. Beskow, and G. Salvi. 2017. Vision-Based Active Speaker Detection in Multiparty Interaction. In *Proceedings of the Grounding Language Understanding*. 47–51.

[25] K. Stefanov, J. Beskow, and G. Salvi. 2020. Self-Supervised Vision-Based Detection of the Active Speaker as Support for Socially-Aware Language Acquisition. *IEEE Transactions on Cognitive and Developmental Systems* 12, 2 (2020), 250–259.

[26] K. Stefanov, G. Salvi, D. Kontogiorgos, H. Kjellström, and J. Beskow. 2019. Modeling of Human Visual Attention in Multiparty Open-World Dialogues. *ACM Transactions on Human-Robot Interaction* 8, 2 (2019), 21.

[27] K. Stefanov, A. Sugimoto, and J. Beskow. 2016. Look Who's Talking: Visual Identification of the Active Speaker in Multi-Party Human-Robot Interaction. In *Proceedings of the Advancements in Social Signal Processing for Multimodal Interaction*. 22–27.

[28] S. E. Tranter and D. A. Reynolds. 2006. An Overview of Automatic Speaker Diarization Systems. *IEEE Transactions on Audio, Speech, and Language Processing* 14, 5 (2006), 1557–1565.

[29] H. Vajaria, S. Sarkar, and R. Kasturi. 2008. Exploring Co-Occurrence Between Speech and Body Movement for Audio-Guided Video Localization. *IEEE Transactions on Circuits and Systems for Video Technology* 18, 11 (2008), 1608–1617.