Self-Supervised Vision-Based Detection of the Active Speaker as Support for Socially Aware Language Acquisition

Kalin Stefanov, Jonas Beskow, and Giampiero Salvi

***Abstract*—This paper presents a self-supervised method for visual detection of the active speaker in a multiperson spoken interaction scenario. Active speaker detection is a fundamental prerequisite for any artificial cognitive system attempting to acquire language in social settings. The proposed method is intended to complement the acoustic detection of the active speaker, thus improving the system robustness in noisy conditions. The method can detect an arbitrary number of possibly overlapping active speakers based exclusively on visual information about their face. Furthermore, the method does not rely on external annotations, thus complying with cognitive development. Instead, the method uses information from the auditory modality to support learning in the visual domain. This paper reports an extensive evaluation of the proposed method using a large multiperson face-to-face interaction data set. The results show good performance in a speaker dependent setting. However, in a speaker independent setting the proposed method yields a significantly lower performance. We believe that the proposed method represents an essential component of any artificial cognitive system or robotic platform engaging in social interactions.**

***Index Terms*—Active speaker detection and localization, cognitive systems and development, language acquisition through development, transfer learning.**

I. Introduction

THE ABILITY to acquire and use language in a similar manner as humans may provide artificial cognitive systems with a unique communication capability and the means for referencing to objects, events, and relationships. In turn, an artificial cognitive system with this capability will be able to engage in natural and effective interactions with humans. Furthermore, developing such systems can help us further understand the underlying processes in language acquisition during the initial stages of the human life. As mentioned in [1], modeling language acquisition is very complex and should integrate different aspects of signal processing, statistical learning, visual processing, pattern discovery, and memory access and organization.

According to many studies (e.g., [2]) there are two alternatives to human language acquisition—individualistic learning and social learning. In the case of individualistic learning, the infant exploits the statistical regularities in the multimodal sensory inputs to discover linguistic units, such as phonemes and words and word-referent mappings. In the case of social learning, the infant can determine the intentions of others by exploiting different social cues. Therefore, in social learning, the participants in the interaction with the infant play a crucial role by constraining the interaction and providing feedback.

From a social learning perspective, the main prerequisite for language acquisition is the ability to engage in social interactions. For an artificial cognitive system to address this challenge, it must at least: 1) be aware of the people in the environment; 2) detect their state: *speaking* or *not speaking*; and 3) infer possible objects the active speaker is focusing attention on.

In this paper, we address the problem of detecting the active speaker in a multiperson language learning scenario. The auditory modality is fundamental for this task and much research has been devoted to audio-based active speaker detection (Section II-B). In this paper, however, we propose to take advantage of the temporal synchronization of the visual and auditory modalities in order to improve the robustness of audio-based active speaker detection. This paper proposes and evaluates three *self-supervised* methods that use the auditory input as reference in order to learn an active speaker detector based on the visual input alone. The goal is not to replace the auditory modality, but to complement it with visual information whenever the auditory input is unreliable.

In order to impose as little constraints as possible on the social interaction, we have two requirements for the proposed methods. The first is that any particular method must operate in real-time (possibly with a short lag), which in practice means that the method should not require any future information. The second requirement is that the methods should make as few assumptions as possible about the environment in which the artificial cognitive system will engage in social interactions. Therefore, the methods should not assume noise-free

Manuscript received November 12, 2017; revised July 1, 2018, October 30, 2018, and April 15, 2019; accepted July 2, 2019. Date of publication July 10, 2019; date of current version June 10, 2020. This work was supported by the CHIST-ERA European Project Interactive Grounded Language Understanding (IGLU). *(Corresponding author: Giampiero Salvi.)*

K. Stefanov is with the Institute for Creative Technologies, University of Southern California, Los Angeles, CA 90089 USA (e-mail: kalins@kth.se).

J. Beskow is with the Department of Speech, Music and Hearing, KTH Royal Institute of Technology, 10044 Stockholm, Sweden (e-mail: beskow@kth.se).

G. Salvi is with the Department of Electronic Systems, NTNU Norwegian University of Science and Technology, 7491 Trondheim, Norway, and also with the Department of Speech, Music and Hearing, KTH Royal Institute of Technology, 10044 Stockholm, Sweden (e-mail: giampi@kth.se).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCDS.2019.2927941

environment, known number of participants in the interaction, or known spatial configuration. The proposed methods address the requirements for engagement in social interactions outlined above, by detecting the people in the environment and detecting their state—speaking or not speaking. In turn, this information is a prerequisite to hypothesizing the possible objects a speaking person is focusing his/her attention on, which has been shown to play an important role in language acquisition (Section II-A).

The rest of this paper is organized as follows. First, we examine previous research that forms the context for the current study in Section II, and then we describe the proposed methods in Section III. The experiments we conducted are described in Section IV, and the results of these experiments are presented in Section V. Discussion on the used evaluation metric, together with the assumptions made, can be found in Section VI. We conclude this paper in Section VII.

II. Related Work

This section is divided in two parts. First, we introduce research on language acquisition which supports our motivation to build an active speaker detector for a language learning artificial cognitive system. Second, we turn our focus on research related to the problem of identifying the active speaker through visual and auditory perceptual inputs.

A. Language Acquisition

The literature on language acquisition offers several theories of how infants learn their first words. One of the main problems which researchers face in this field is the *referential ambiguity* as discussed, for example, in [3]–[5]. Referential ambiguity stems from the idea that infants must acquire language by linking heard words with perceived visual scenes, in order to form word-referent mappings. In everyday life however, these visual scenes are highly cluttered which results in many possible referents for any heard word, within any learning event [6], [7]. Similarly, many computational models of language acquisition are rooted in finding statistical associations between verbal descriptions and the visual scene [3], [8]–[10], or in more interactive robotic manipulation experiments [11]. However, nearly all of them assume a clutter-free visual scene, where objects are observed in isolation on a simplified background (often white table).

Different theories offer alternative mechanisms through which infants reduce the uncertainty present in the learning environment. One such mechanism is statistical aggregation of word-referent co-occurrences across learning events. The problem of referential ambiguity within a single learning event has been addressed by Smith *et al.* [12], [13], suggesting that infants can keep track of co-occurring words and potential referents across learning events and use this aggregated information to statistically determine the most likely word-referent mapping. However, the authors argued that this type of statistical learning may be beyond the abilities of infants when considering highly cluttered visual scenes. In order to study the visual scene clutter from the infants' perspective, Pereira *et al.* [4] and Yurovsky *et al.* [5] performed experiments in which the infants were equipped with a head-mounted eye-tracker. The conclusion was that some learning events are not ambiguous because there was only one dominant object when considering the infants' point of view. As a consequence, the researchers argued that the input to language learning must be understood from the infants' perspective, and only regularities that make contact with the infants' sensory system can affect their language learning. Although not related to language acquisition, an attempt at modeling the saliency of multimodal stimuli from the learner's (robot's) perspective was proposed in [14]. This bottom up approach is based exclusively on the statistical properties of the sensory inputs.

Another mechanism to cope with the uncertainty in the learning environment might be related to social cues to the caregivers' intent, as mentioned in the above studies. Although a word is heard in the context of many objects, infants may not treat the objects as equally likely referents. Instead, infants can use social cues to rule out contenders to the named object. Yu and Smith [15] used eye-tracking to record gaze data from both caregivers and infants and found that when the caregiver visually attended to the object to which infants' attention was directed, infants extended the duration of their visual attention to that object, thus increasing the probability for successful word-referent mapping.

Infants do not learn only from interactions they are directly involved in, but also observe and attend to interactions between their caregivers. Handl *et al.* [16] and Meng *et al.* [17] performed studies to examine how the body orientation can influence the infants' gaze shifts. These studies were inspired by large body of research on gaze following which suggests that infants' use others' gaze to guide their own attention, that infants pay attention to conversations, and that joint attention has an effect on early learning. The main conclusion was that static body orientation alone can function as a cue for infants' observations and guides their attention. Barton and Tomasello [18] also reasoned that multiperson context is important in language acquisition. In their triadic experiments, joint attention was an important factor facilitating infants' participation in conversations; infants were more likely to take a turn when they shared a joint attentional focus with the speaker. Yu and Ballard [9] also proposed that speakers' eye movements and head movements among others, can reveal their referential intentions in verbal utterances, which could play a significant role in an automatic language acquisition system.

The above studies do not consider how infants might know which caregiver is actively speaking and therefore requires attention. We believe that this is an important prerequisite to modeling automatic language acquisition. The focus of the study described in this paper is, therefore, to investigate different methods for inferring the active speaker. We are interested in methods that are plausible from a developmental cognitive system perspective. One of the main implications is that the methods should not require manual annotations.

B. Active Speaker Detection

Identifying the active speaker is important for many applications. In each area, different constraints are imposed to the methods. Generally, there are three different approaches: 1) audio-only; 2) audio-visual; and 3) approaches that use other forms of inputs for detection.

Audio-only active speaker detection is the process of finding segments in the input audio signal associated with different speakers. This type of detection is known as *speaker diarization*. Speaker diarization has been studied extensively. Anguera *et al.* [19] offered a comprehensive review of recent research in this field. In realistic situations, with far-field microphones, or microphone arrays, the task of active speaker detection from audio is far from trivial. Most methods (e.g., [20] and [21]), use some form of model-based supervised training. This is one of the motivation for this paper: First, we believe that complementing the auditory modality with visual information can be useful if not necessary for this task, especially in the more challenging acoustic conditions. Second, we want to comply with a developmental approach, where the learning system only uses the information available through its senses in the interaction with humans. We therefore want to avoid the need for careful annotations that are required by the aforementioned supervised methods.

Audio-visual speaker detection combines information from both the audio and the video signals. The application of audio-visual synchronization to speaker detection in broadcast videos was explored by Nock *et al.* [22]. Unsupervised audio-visual detection of the speaker in meetings was proposed in [23]. Zhang *et al.* [24] presented a boosting-based multimodal speaker detection algorithm applied to distributed meetings, to give three examples. Mutual correlations to associate an audio source with regions in the video signal was demonstrated by Fisher *et al.* [25], and Slaney and Covell [26] showed that audio-visual correlation can be used to find the temporal synchronization between audio signal and a speaking face. An elegant solution was proposed in [27] where the mutual information between the acoustic and visual signals is computed by means of a joint multivariate Gaussian process, with the assumption that only one audio and one video streams were present and that locating the source corresponds to finding the pixels in the image that correlate with acoustic activity. In more recent studies, researchers have employed artificial neural network architectures to build active speaker detectors from audio-visual input. A multimodal long short-term memory (LSTM) model that learns shared weights between modalities was proposed in [28]. The model was applied to speaker naming in TV shows. Hu *et al.* [29] proposed a convolutional neural network (CNN) model that learns the fusion function of face and audio information.

Other approaches for speaker detection include a general pattern recognition framework used by Besson and Kunt [30] applied to detection of the speaker in audio-visual sequences. Visual activity (the amount of movement) and focus of visual attention were used as inputs by Hung and Ba [31] to determine the current speaker on real meetings. Stefanov *et al.* [32] used action units as inputs to hidden Markov Models to determine the active speaker in multi-party interactions and Vajaria *et al.* [33] demonstrated that information for body movements can improve the detection performance.

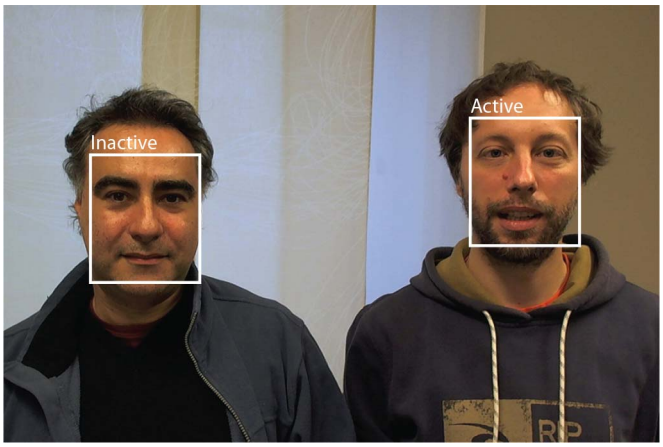

Fig. 1. Example of an output of a visual active speaker detector.

Most of the approaches cited in this section are either evaluated on small amounts of data, or have not been proved to be usable in real-time settings. Furthermore, they usually require manual annotations and the spatial configuration of the interaction and the relative position of the input sensors is known. The goal is usually an offline video/audio analysis task, such as semantic indexing and retrieval of TV broadcasts or meetings, or video/audio summarization. We believe that the challenge of real-time detection of the active speaker in dynamic and cluttered environments remains. In the context of automatic language acquisition, we want to infer the possible objects the active speaker is focusing attention on. In this context, assumptions, such as known sensor arrangement or participants' position and number in the environment are unrealistic, and should be avoided. Therefore, in this paper we present methods which have several desirable characteristics for such types of scenarios: 1) they work in real-time; 2) they do not assume specific spatial configuration (sensors or participants); 3) the number of possible (simultaneously) speaking participants is free to change during the interaction; and 4) no externally produced labels are required, but rather the acoustic inputs are used as reference to the visually based learning.

III. Methods

The goal of the methods described in this section is to detect in real-time the state (speaking or not speaking) of all visible faces in a multiperson language learning scenario, using only visual information (the RGB color data). An illustration of the desired output of an active speaker detector can be seen in Fig. 1.

We use a self-supervised learning approach to construct an active speaker detector: the machine learning methods are supervised, but the labels are obtained automatically from the auditory modality to learn models in the visual modality. An

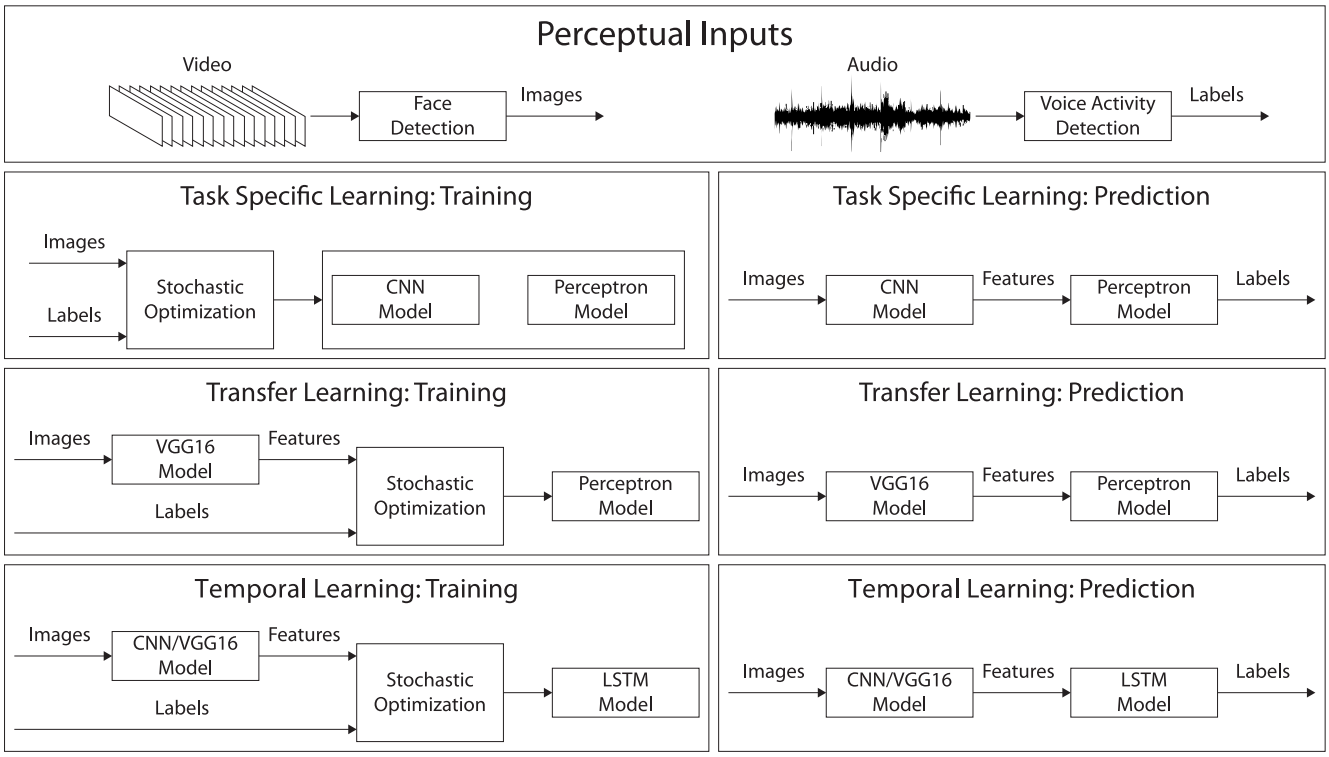

Fig. 2. Approaches to visual active speaker detection considered in the study. In the first row are the perceptual inputs automatically extracted from the video and audio streams. These inputs are passed to the task specific learning (second row), transfer learning (third row) and temporal learning (forth row) methods.

overview of the approaches considered in the study is given in Fig. 2. The first row in the figure illustrates the perceptual inputs that are automatically extracted from the raw audio and video streams. The visual input consists of RGB images of each face extracted from the video stream with the Viola and Jones's face detector [34]. The auditory input consists of labels extracted from the audio stream which correspond to the voice activity. The used audio-only voice activity detector (VAD) [35] is based on two thresholds on the energy of the signal, one to start a speech segment and one to end it. These thresholds are adaptive and based on a histogram method. The ability to extract face images and VAD labels is given as a starting point to the system and is motivated in Section VI.

The methods use a feature extractor based on a CNN, followed by a classifier. Two types of classifiers are tested: 1) nontemporal (Perceptron) and 2) temporal (LSTM network). Additionally, two techniques for training the models are considered: 1) transfer learning that employs a pretrained feature extractor and only trains a classifier specifically for the task and 2) task specific learning that trains a feature extractor and a classifier simultaneously for the task.

Each method outputs *a posterior* probability distribution over the two possible outcomes (speaking or not speaking). Since the goal is a binary classification, the detection of the active speaker happens when the corresponding probability exceeds 0.5. The evaluation of each method is performed by computing the accuracy of the predictions on frame-by-frame basis (Section IV).

A. Task Specific Learning

An illustration of the task specific learning method is shown in the second row of Fig. 2. This method trains a CNN feature extractor in combination with a Perceptron classifier with the goal of classifying each input image either as speaking or not speaking. During the training phase both images and labels are used by a gradient-based optimization procedure [36] to adjust the weights of the CNN and Perceptron models. During the prediction phase, only images are used by the trained models to generate labels. The CNN and Perceptron models work on a frame-by-frame basis and have no memory of past frames.

B. Transfer Learning

An illustration of this method can be seen in the third row of Fig. 2. Similarly to the previous method, the transfer learning method uses a CNN and a Perceptron model. In this method, however, the CNN model is pretrained on an object recognition task (i.e., VGG16 [37]). To adapt the VGG16 model to the active speaker detection task, the object classification layer is removed and the truncated VGG16 model is used as a feature extractor. Then the method consists of training only a Perceptron model to map the features generated by the VGG16 model to the speaker activity information. As for the task specific learning method, this method has no memory of past frames.

Because the VGG16 model was originally trained in a supervised manner to classify objects, this raises the question on how suitable this model is in the context of developmental

language acquisition. Support to the use of this model comes from the literature on visual perception that demonstrates the ability of infants to recognize objects very early in their development [38], [39].

C. Temporal Learning

The temporal learning method is illustrated by the forth row of Fig. 2. This method is based on the previously described feature extractors, but introduces a model of the time evolution of the perceptual inputs. During the training phase a custom (CNN) or pretrained (VGG16) feature extractor constructs a feature vector for each input image. Then the features and labels are used by a gradient-based optimization procedure [36] to adjust the weights of a LSTM model [40]. During the prediction phase, images are converted into features with a custom CNN or VGG16 model, which features are then used by the trained detector (LSTM) to generate labels.

D. Acoustic Noise

In order to test the effect of noise on the audio-only VAD, stationary noise is added to the audio signal. The noise is sampled from a Gaussian distribution with zero mean and variance σ^2. For every recording, the active segments are first located by means of the audio-only VAD. These are then used to estimate the energy E_x of the signal as the mean squares of the samples. Then σ^2 is computed as the ratio between the energy of the signal and the desired signal-to-noise ratio (SNR)

$$\sigma^2 = \frac{E_x}{10^{\frac{\text{SNR}}{10}}}. \tag{1}$$

Finally, the noise is added to the signal, and the samples are renormalized to fit in the 16 bit linear representation. The audio-only VAD is used again on the noisy signal and its accuracy is computed on the result.

IV. Experiments

This section is divided in two parts. The first part describes the data set used to build and evaluate the active speaker detectors. The second part describes the general setup of the conducted experiments.

A. Data Set

The methods presented in Section III are implemented and evaluated using a multimodal multiparty interaction data set described in [41]. The main purpose of the data set is to explore patterns in the focus of visual attention of humans under the following three different conditions: two humans involved in task-based interaction with a robot; the same two humans involved in task-based interaction where the robot is replaced by a third human, and a free three-party human interaction. The data set contains two parts: 1) six sessions with duration of approximately 30 min each and 2) nine sessions, each of which is with duration of approximately 40 min. The data set is rich in modalities and recorded data streams. It includes the streams generated from three Kinect v2 devices (color, depth, infrared, body, and face data), three high-quality audio streams generated from close-talking microphones, three high-resolution video streams generated from GoPro cameras, touch-events stream for the task-based interactions generated from an interactive surface, and the system state stream generated by the robot involved in the first condition. The second part of the data set also includes the data streams generated from 3 Tobii Pro Glasses 2 eye trackers. The interactions are in English and all data streams are spatially and temporally synchronized and aligned. The interactions occur around a round interactive surface and all 24 unique participants are seated. Fig. 3 illustrates the spatial configuration of the setup in the data set.

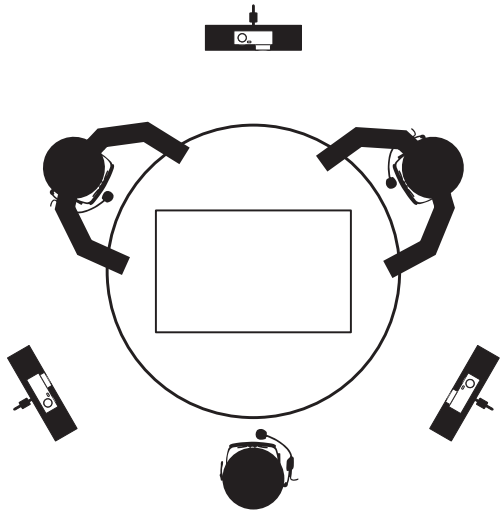

Fig. 3. Spatial configuration of the sensors and participants in the data set.

Fig. 4. Example of a difficult visual input from the first and second condition in the data set.

As described previously, each interaction in the data set is divided into three conditions, with the first and second condition being related to a collaborative task-based interaction in which the participants play a game on a touch surface. During this two conditions the participants interact mainly with the touch surface and discuss with their partner how to solve the given task. Therefore, the participants' overall gaze direction (head orientation) is toward the touch surface. This raises some very challenging visual conditions for extracting speech activity information from the face. We show three examples in Fig. 4. This observation motivated experiments using only the data from the third condition of each interaction.

B. Experimental Setup

This section describes the general setup of the experiments. In all experiments the video stream is generated by

TABLE I
SPEAKER DEPENDENT RESULTS (TENFOLD CROSS-VALIDATION); MEAN ACCURACY AND STANDARD DEVIATION

Features	Perceptron	LSTM_15	LSTM_30	LSTM_150	LSTM_300
CNN	**73.13** (7.81)	72.92 (8.47)	**73.13** (8.67)	72.61 (9.54)	72.46 (9.56)
VGG16	72.61 (8.27)	72.90 (8.85)	**73.27** (9.14)	72.46 (9.97)	72.55 (10.22)

the Kinect v2 device directed at the participant under consideration and the audio stream is generated by the participant's close-talking microphone. The total amount of frames used in the experiments is 690 000 (∼6.5 h).

The CNN models comprise three convolutional layers of width 32, 32, and 64 with receptive fields of 3×3 and rectifier activations, interleaved by max pooling layers with window size of 2×2. The output of the last max pooling layer is used by a densely connected layer of size 64 with rectifier activation functions and finally by a perceptron layer with logistic sigmoid activations. The LSTM models include one LSTM layer of size 128 with hyperbolic tangent activations, followed by a densely connected and a perceptron layer similarly to the CNN models.

During the training phase the models use Adam optimizer with default parameters ($\alpha = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$) and binary crossentropy loss function. Each nontemporal model (CNN and Perceptron) is trained for 50 epochs and each temporal model (LSTM) is trained for 100 epochs. The LSTM models are trained with 15, 30, 150, and 300 frame (500 ms, 1 s, 5 s, and 10 s) long segments without overlaps. The models corresponding to the best validation performance are selected for evaluation on the test set. The models are implemented in Keras [42] with TensorFlow [43] backend. During the prediction phase only the RGB color images extracted with the face detector are used as input. As described previously, each of the considered methods outputs *a posterior* probability distribution over the two possible outcomes—speaking or not speaking. Therefore, when evaluating the models' performance, 0.5 is used as a threshold for assigning a class to each frame-level prediction. The results are reported in terms of frame-by-frame weighted accuracy which is calculated with

$$\text{wacc} = 100 \times \frac{\frac{\text{tp}}{\text{tp+fn}} + \frac{\text{tn}}{\text{fp+tn}}}{2} \tag{2}$$

where tp, fp, tn, and fn are the number of true positives, false positives, true negatives, and false negatives, respectively. As a consequence, regardless of the actual class distribution in the test set (which is in general different for each participant), the baseline chance performance using this metric is always 50%. Although this metric allows an easy comparison of results between different participants and methods, it is a very conservative measure of performance (Section VI-A).

This paper presents three experiments with the proposed methods: 1) speaker dependent; 2) multispeaker dependent; and 3) speaker independent. The speaker dependent experiment builds a model for each participant and tests it on independent data from the same participant. This process is repeated ten times per participant with splits generated through a tenfold cross-validation procedure. The multispeaker dependent experiment uses the splits generated in speaker dependent experiment. This experiment, however, builds a model with the data for all participants and tests it on the independent data from all participants. This experiment tests the scalability of the proposed methods to more than one participant. The speaker independent experiment uses a leave-one-out cross-validation procedure to build and evaluate the models. This experiment tests the transferability of the proposed methods to unseen participants.

Finally, as described in Section III-D, the effect of noise is tested on the audio-only VAD. The proposed video-only active speaker detectors are compared with audio-only VAD where the SNR varies from 0 to 30 in increments of 5.

V. RESULTS

This section presents the numerical results obtained from the experiments.

A. Speaker Dependent

The mean accuracy and standard deviation per method obtained in the speaker dependent experiment are provided in Table I. The highest mean result in this experiment is 73.13% for the LSTM_30 models when using custom CNN feature extractors and 73.27% for the LSTM_30 models when using pretrained VGG16 feature extractors. The complete results are illustrated in Fig. 5. The figure shows that the accuracy varies significantly between participants. Also the variability between participants is higher than the difference obtained with different methods per participant. A comparison between the best performing video-only method and an audio-only VAD is illustrated in the left plot of Fig. 6. The two methods give similar results for a range of SNRs around 12. The video-only method outperforms the audio-only VAD for more noisy conditions, whereas the opposite is true if the SNR is greater than 20.

B. Multispeaker Dependent

The summarized results of the multispeaker dependent experiment are provided in Table II. The highest mean result in this experiment is 75.76% for the LSTM_150 models when using custom CNN feature extractors. A comparison between the best performing video-only method and an audio-only VAD is illustrated in the center plot of Fig. 6. Similarly to the speaker dependent case, the two methods give similar results for a range of SNRs around 12. However, in this case the spread around the mean is much reduced because every fold includes a large collection of samples from all participants.

C. Speaker Independent

The summarized results of the speaker independent experiment are provided in Table III. The highest mean result in this experiment is 57.11% for the LSTM_30 models when using

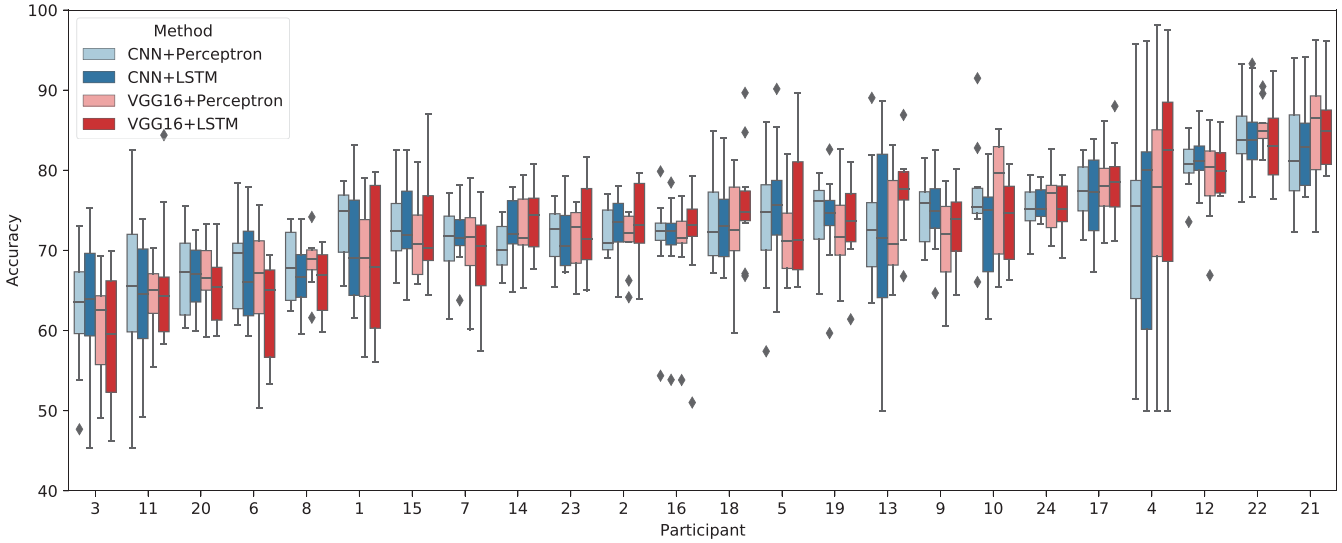

Fig. 5. Accuracy versus participant and method. The participants are sorted by overall accuracy. The segment length for the LSTMs is 15 frames (500 ms). The boxplots show the results over all tenfolds.

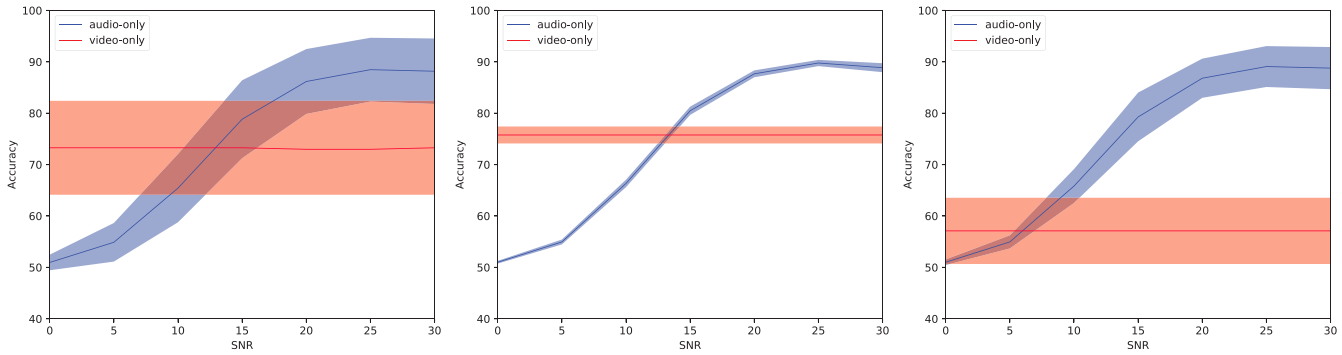

Fig. 6. Comparison between audio-only and video-only method in noise (the solid lines are accuracies and the shaded areas are standard deviations). The accuracy in the speaker dependent experiment (left) is averaged over 24 participants and tenfolds. The accuracy in the multispeaker dependent experiment (center) is averaged over tenfolds each containing data from 24 participants. The accuracy in the speaker independent experiment (right) is averaged over 24 folds each containing data from the participant that was left out during training. In all cases the performance of the audio-only method degrades with the reduction in SNR, whereas the video-only method is not affected by acoustic noise.

TABLE II
MULTISPEAKER DEPENDENT RESULTS (TENFOLD CROSS-VALIDATION); MEAN ACCURACY AND STANDARD DEVIATION

Features	Perceptron	LSTM_15	LSTM_30	LSTM_150	LSTM_300
CNN	74.80 (1.63)	74.91 (1.54)	75.11 (1.57)	**75.76** (1.65)	75.26 (1.46)

TABLE III
SPEAKER INDEPENDENT RESULTS (LEAVE-ONE-OUT CROSS-VALIDATION); MEAN ACCURACY AND STANDARD DEVIATION

Features	Perceptron	LSTM_15	LSTM_30	LSTM_150	LSTM_300
CNN	55.39 (5.74)	56.33 (6.56)	**57.11** (6.44)	56.96 (6.50)	57.55 (7.02)

custom CNN feature extractors. A comparison between the best performing video-only method and an audio-only VAD is illustrated in the right plot of Fig. 6. As can be observed, the results from the video-only method are only slightly above chance level, hence falling far behind the audio-based VAD.

VI. DISCUSSION

In order to interpret the results presented in Section V we need to make a number of considerations about the evaluation method. We will also consider the advantages and limitations of the metric used and detail the assumptions made in the methods and the main contributions of this paper.

The proposed methods estimate the probability of speaking independently for each face. This has the advantage of being able to detect several speakers that are active at the same time, but for many applications it might be sufficient to select the active speaker among the detected faces. Doing this would allow us to combine the single predictions into a joint probability, thus increasing the performance.

It is important to note that the conditions in the experiment that compared audio-only and video-only methods were

favorable to the audio-only method due to the use of stationary noise. The VAD employed for the audio-based detection uses adaptive thresholds that are specifically suitable for stationary noise. Therefore, we would expect a larger advantage for the video-based speaker detection in low to medium SNRs in the presence of nonstationary noises often present in natural communication environments.

A. Metric

Evaluating the proposed methods on a frame-by-frame basis gives a detailed measure of performance. However, one might argue that frame-level (33 ms) accuracy is not necessary for artificial cognitive systems employing the proposed methods in the context of automatic language acquisition. Evaluating the methods on a fixed-length sliding time window (e.g., 200 ms) might be sufficient for this application.

Furthermore, the definition of the weighted accuracy amplifies short mistakes. For example, if in 100 frames, 98 belong to the active class and 2 to the inactive class, a method that classifies all frames as active will have $\text{wacc} = (100/2) \times [98/(98+0) + 0/(2+0)] = 50\%$. If we consider a case of continuous talking, where the speaker takes short pauses to recollect a memory or structure the argument, then a perfect audio-only method will detect silence of certain length (at least 200 ms) in the acoustic signal and label the corresponding video frames as not speaking. However, from the interaction point of view the speaker might be still active, resulting also in visual activity. A video-only method that misses these short pauses would be strongly penalized by the used metric, achieving as low as 50% accuracy when all other frames are classified correctly. Similar situation occurs when a person is listening and gives short acoustic feedbacks which are missed by the video-only methods.

The advantage of the weighted accuracy metric, however, is that it enables us to seamlessly compare the performance between participants and methods. This is because, the different underlying class distributions due to each particular data set, are accounted for by the metric and the resulting baseline is 50% for all considered experimental configurations.

B. Assumptions

The proposed methods make the following assumptions.

1) The system is able to detect faces.
2) The system is able to detect speech for a single speaker.
3) There are situations in which the system only interacts with one speaker, and can therefore use the audio-only VAD to train the video-only active speaker detector.

In order to motivate the plausibility of these assumptions in the context of a computational method for language acquisition, we consider research in developmental psychology. According to studies reported in [44] and [45] infants can discriminate between several facial expressions which suggests that they are capable of detecting human faces. The assumption that the system can detect speech seems to be supported by research on recognition of mother's voice in infants (e.g., [46]). However, whereas infants can detect the voice at a certain distance from the speaker, here we make the simplifying assumption that we can record and detect speech activity from close-talking microphones for each speaker. It remains to be verified if we can obtain similar performance from the audio-only VAD in case we use far-field microphones or microphone arrays, or in noisy acoustic conditions. The final assumption is reasonable considering that infants interact with small number of speakers in their first months, and in many cases only one parent is available as caregiver at any specific time.

C. Contributions

This paper extends our previous work [47] on vision-based methods for detection of the active speaker in multiparty human–robot interactions. We will summarize the main differences between this paper and [47] in this section. The first difference is the use of a better performing pretrained CNN model for feature extraction (i.e., VGG16 [37]) compared to the previously used AlexNet [48]. We also significantly extended the set of experiments to evaluate and compare the proposed methods. In this paper, we evaluated the effect of using temporal models by comparing the performance of LSTM models similar to the ones evaluated in [47], to non-temporal Perceptron models. Furthermore, we compared the performance of transfer learning models, with models that are built specifically for the current application and trained exclusively on the task specific data. Finally, we reported results for multispeaker and speaker independent experiments.

One of our findings is that, given that we optimize the classifier to the task (Perceptron or LSTM), it is not necessary to optimize the feature extractor (the custom CNNs perform similarly to the pretrained VGG16). This suggests that a pretrained feature extractor, such as VGG16 works well independently of the speaker and can be used to extend the results beyond the participants in the present data set. Also, the result of the multispeaker dependent experiment shows that the proposed methods can scale beyond a single speaker without decrease in performance. Combining this observation with the observation for the applicability of transfer learning suggests that a mixture of the proposed methods can be indeed an useful component of a real life artificial cognitive system.

Finally, the speaker independent experiment yields significantly lower performance compared to the other two experiments. We should mention, however, that, from a cognitive system's perspective, this might be an unnecessarily challenging condition. We can in fact expect infants to be familiar with a number of caregivers, thus justifying a condition more similar to the settings in the multispeaker dependent experiment.

VII. Conclusion

In this paper, we proposed and evaluated three methods for automatic detection of the active speaker-based solely on visual input. The proposed methods are intended to complement acoustic methods, especially in noisy conditions, and could assist an artificial cognitive system to engage in social interactions which has been shown to be beneficial for language acquisition.

We tried to reduce the assumptions about the language learning environment to a minimum. Therefore, the proposed methods allow different speakers to speak simultaneously as well as to be all silent; the methods do not assume a specific number of speakers, and the probability of speaking is estimated independently for each speakers, thus allowing the number of speakers to change during the social interaction.

We evaluated the proposed methods on a large multiperson data set. The methods perform well on a speaker dependent and multispeaker dependent fashion, reaching accuracy of over 75% (baseline 50%) on a weighted frame-based evaluation metric. The combined results obtained from the transfer learning and multispeaker learning experiments are promising and suggest that the proposed methods can generalize to unseen perceptual inputs by incorporating a model adaptation step for each new face.

We should acknowledge the general difficulty of the problem addressed in this paper. Humans generally produce many facial configurations when they are not speaking that might be highly overlapping to the configurations associated with when they are speaking.

The methods proposed in this paper are in support to socially aware language acquisition and they can be seen as mechanisms for constraining the visual input thus providing higher quality and more appropriate data for a statistical learning of word-referent mappings. Therefore, the main purpose of the methods is to help bringing an artificial cognitive system one step closer to resolving the referential ambiguity in cluttered, dynamic, and noisy environments.

ACKNOWLEDGMENT

The authors would like to thank the NVIDIA Corporation for donating the GeForce GTX TITAN cards used for this paper, and the Swedish National Infrastructure for Computing at the Parallel Data Center at KTH for computational time allocation. They also would like to thank the anonymous reviewers for their insightful comments.

REFERENCES

[1] L. T. Bosch, L. Boves, H. Van Hamme, and R. K. Moore, "A computational model of language acquisition: The emergence of words," *Fundamenta Informaticae*, vol. 90, no. 3, pp. 229–249, 2009.

[2] L. Steels and F. Kaplan, "Aibo's first words: The social learning of language and meaning," *Evol. Commun.*, vol. 4, no. 1, pp. 3–32, 2000.

[3] E. M. Clerkin, E. Hart, J. M. Rehg, C. Yu, and L. B. Smith, "Real-world visual statistics and infants' first-learned object names," *Philos. Trans. Roy. Soc. London B Biol. Sci.*, vol. 372, no. 1711, 2016, Art. no. 20160055.

[4] A. F. Pereira, L. B. Smith, and C. Yu, "A bottom-up view of toddler word learning," *Psychonomic Bull. Rev.*, vol. 21, no. 1, pp. 178–185, 2014.

[5] D. Yurovsky, L. B. Smith, and C. Yu, "Statistical word learning at scale: The baby's view is better," *Develop. Sci.*, vol. 16, no. 6, pp. 959–966, 2013.

[6] W. V. O. Quine, *Word and Object*. Cambridge, MA, USA: MIT Press, 2013.

[7] P. Bloom, *How Children Learn the Meanings of Words*. Cambridge, MA, USA: MIT Press, 2000.

[8] D. K. Roy and A. P. Pentland, "Learning words from sights and sounds: A computational model," *Cogn. Sci.*, vol. 26, no. 1, pp. 113–146, 2002.

[9] C. Yu and D. H. Ballard, "A multimodal learning interface for grounding spoken language in sensory perceptions," *ACM Trans. Appl. Percept.*, vol. 1, no. 1, pp. 57–80, 2004.

[10] O. Räsänen and H. Rasilo, "A joint model of word segmentation and meaning acquisition through cross-situational learning," *Psychol. Rev.*, vol. 122, no. 4, pp. 792–829, 2015.

[11] G. Salvi, L. Montesano, A. Bernardino, and J. Santos-Victor, "Language bootstrapping: Learning word meanings from perception–action association," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 3, pp. 660–671, Jun. 2012.

[12] L. B. Smith and C. Yu, "Infants rapidly learn word-referent mappings via cross-situational statistics," *Cognition*, vol. 106, no. 3, pp. 1558–1568, 2008.

[13] L. B. Smith, S. H. Sumarga, and C. Yu, "The unrealized promise of infant statistical word-referent learning," *Trends Cogn. Sci.*, vol. 18, no. 5, pp. 251–258, 2014.

[14] J. Ruesch, M. Lopes, A. Bernardino, J. Hornstein, J. Santos-Victor, and R. Pfeifer, "Multimodal saliency-based bottom-up attention a framework for the humanoid robot iCub," in *Proc. IEEE Int. Conf. Robot. Autom.*, Pasadena, CA, USA, 2008, pp. 962–967.

[15] C. Yu and L. B. Smith, "The social origins of sustained attention in one-year-old human infants," *Current Biol.*, vol. 26, no. 9, pp. 1235–1240, 2016.

[16] A. Handl, T. Mahlberg, S. Norling, and G. Gredebäck, "Facing still faces: What visual cues affect infants' observations of others?" *Infant Behav. Develop.*, vol. 36, no. 4, pp. 583–586, 2013.

[17] X. Meng, Y. Uto, and K. Hashiya, "Observing third-party attentional relationships affects infants' gaze following: An eye-tracking study," *Front. Psychol.*, vol. 7, p. 2065, Jan. 2017.

[18] M. E. Barton and M. Tomasello, "Joint attention and conversation in mother-infant-sibling triads," *Child Develop.*, vol. 62, no. 3, pp. 517–529, 1991.

[19] X. Anguera, N. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 2, pp. 356–370, Feb. 2012.

[20] X. Anguera, C. Wooters, B. Peskin, and M. Aguiló, "Robust speaker segmentation for meetings: The ICSI-SRI spring 2005 diarization system," in *Machine Learning for Multimodal Interaction*. Berlin, Germany: Springer, 2006, pp. 402–414.

[21] C. Fredouille and G. Senay, "Technical improvements of the E-HMM based speaker diarization system for meeting records," in *Machine Learning for Multimodal Interaction*. Berlin, Germany: Springer, 2006, pp. 359–370.

[22] H. J. Nock, G. Iyengar, and C. Neti, "Speaker localisation using audio-visual synchrony: An empirical study," in *Proc. 2nd Int. Conf. Image Video Retrieval*, 2003, pp. 488–499.

[23] G. Friedland, C. Yeo, and H. Hung, "Visual speaker localization aided by acoustic models," in *Proc. 17th ACM Int. Conf. Multimedia*, Beijing, China, 2009, pp. 195–202.

[24] C. Zhang *et al.*, "Boosting-based multimodal speaker detection for distributed meeting videos," *IEEE Trans. Multimedia*, vol. 10, no. 8, pp. 1541–1552, Dec. 2008.

[25] J. W. Fisher, T. Darrell, W. T. Freeman, and P. Viola, "Learning joint statistical models for audio-visual fusion and segregation," in *Proc. 13th Adv. Neural Inf. Process. Syst.*, Denver, CO, USA, 2001, pp. 772–778.

[26] M. Slaney and M. Covell, "FaceSync: A linear operator for measuring synchronization of video facial images and audio tracks," in *Proc. 13th Adv. Neural Inf. Process. Syst.*, Denver, CO, USA, 2001, pp. 814–820.

[27] J. R. Hershey and J. R. Movellan, "Audio-vision: Using audio-visual synchrony to locate sounds," in *Proc. Adv. Neural Inf. Process. Syst.*, Denver, CO, USA, 2000, pp. 813–819.

[28] J. Ren *et al.*, "Look, listen and learn—A multimodal LSTM for speaker identification," in *Proc. 30th AAAI Conf. Artif. Intell.*, Phoenix, AZ, USA, 2016, pp. 3581–3587.

[29] Y. Hu, J. Ren, J. Dai, C. Yuan, L. Xu, and W. Wang, *Deep Multimodal Speaker Naming*. New York, NY, USA: ACM, 2015, pp. 1107–1110.

[30] P. Besson and M. Kunt, "Hypothesis testing for evaluating a multimodal pattern recognition framework applied to speaker detection," *J. Neuroeng. Rehabil.*, vol. 5, no. 1, p. 11, 2008.

[31] H. Hung and S. O. Ba, "Speech/non-speech detection in meetings from automatically extracted low resolution visual features," Idiap, Martigny, Switzerland, Rep. Idiap-RR-20-2009, 2009.

[32] K. Stefanov, A. Sugimoto, and J. Beskow, "Look who's talking: Visual identification of the active speaker in multi-party human–robot interaction," in *Proc. 2nd Workshop Adv. Soc. Signal Process. Multimodal Interact.*, Tokyo, Japan, 2016, pp. 22–27.

[33] H. Vajaria, S. Sarkar, and R. Kasturi, "Exploring co-occurence between speech and body movement for audio-guided video localization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 11, pp. 1608–1617, Nov. 2008.
[34] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1, 2001, pp. I511–I518.
[35] G. Skantze and S. Al Moubayed, "IrisTK: A statechart-based toolkit for multi-party face-to-face interaction," in *Proc. 14th ACM Int. Conf. Multimodal Interact.*, Santa Monica, CA, USA, 2012, pp. 69–76.
[36] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent.*, 2015.
[37] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Represent.*, 2015.
[38] E. S. Spelke, "Principles of object perception," *Cogn. Sci.*, vol. 14, no. 2, pp. 29–56, 1990.
[39] N. Z. Kirkham, J. A. Slemmer, and S. P. Johnson, "Visual statistical learning in infancy: Evidence for a domain general learning mechanism," *Cognition*, vol. 83, no. 2, pp. B35–B42, 2002.
[40] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[41] K. Stefanov and J. Beskow, "A multi-party multi-modal dataset for focus of visual attention in human-human and human–robot interaction," in *Proc. 10th Int. Conf. Lang. Resources Eval. (LREC)*, 2016, pp. 4440–4444.
[42] F. Chollet *et al.* (2015). *Keras*. [Online]. Available: https://github.com/fchollet/keras
[43] M. Abadi *et al.* (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. [Online]. Available: http://tensorflow.org/
[44] J. D. LaBarbera, C. E. Izard, P. Vietze, and S. A. Parisi, "Four- and six-month-old infants' visual responses to joy, anger, and neutral expressions," *Child Develop.*, vol. 47, no. 2, pp. 535–538, 1976.
[45] G. Young-Browne, H. M. Rosenfeld, and F. D. Horowitz, "Infant discrimination of facial expressions," *Child Develop.*, vol. 48, no. 2, pp. 555–562, 1977.
[46] M. Mills and E. Melhuish, "Recognition of mother's voice in early infancy," *Nature*, vol. 252, pp. 123–124, Nov. 1974.
[47] K. Stefanov, J. Beskow, and G. Salvi, "Vision-based active speaker detection in multiparty interaction," in *Proc. Int. Workshop Grounding Lang. Understand. (GLU)*, 2017, pp. 47–51.
[48] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. 25th Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

Kalin Stefanov received the M.Sc. degree in artificial intelligence from the University of Amsterdam, Amsterdam, The Netherlands, and the Ph.D. degree in computer science from the KTH Royal Institute of Technology, Stockholm, Sweden.

He is currently a Postdoctoral Fellow with the Institute for Creative Technologies, University of Southern California, Los Angeles, CA, USA. His current research interests include machine learning, computer vision, and speech technology.

Jonas Beskow received the M.Sc. degree in electrical engineering and the Ph.D. degree in speech communication from the KTH Royal Institute of Technology, Stockholm, Sweden, in 1995 and 2003, respectively.

He is a Professor of speech communication with the KTH Royal Institute of Technology, Stockholm, Sweden. His current research interests include multimodal speech technology, modeling and generating verbal and nonverbal communicative behavior as well as embodied conversational agents or social robots that use speech, gesture and/or other modalities in order to accomplish human-like interaction. He is also a Co-Founder of Furhat Robotics, Stockholm, a startup developing an innovative social robotics platform based on KTH research.

Giampiero Salvi received the M.Sc. degree in electrical engineering from Università la Sapienza, Rome, Italy, and the Ph.D. degree in computer science from the KTH Royal Institute of Technology, Stockholm, Sweden.

He was a Postdoctoral Fellow with the Institute of Systems and Robotics, Lisbon, Portugal. He is currently a Professor of machine learning with the NTNU Norwegian University of Science and Technology, Trondheim, Norway. His current research interests include machine learning, speech technology, and cognitive systems.