Proposal: *Improving Human-Robot Conversation with Active Speaker Detection*

Adrian Auer

September 10, 2026

Contents

1 Introduction

Human-Robot Interaction (HRI) becomes more and more important in a world where robots integrate fast in all aspects of our lives. The industry is currently changing from static robotic environments to dynamic environments where humans collaborate with robots instead of operating robots. In private homes robotic applications like robot vacuums and digital assistants for home automation are becoming regular household items. Although this movements is drastically changing our society, interaction and dialogues between humans and robots are very command driven and unnatural in contrast to Human-Human Interaction (HHI) where more modalities like gesture, posture, facial expressions and gaze are involved. We identified Active Speaker Detection (ASD) as one key aspect in HHI which is currently not appropriately implemented for modern Social Robots in Human-Robot Convesation (HRC). Social Robots cannot see who is speaking and therefore constantly fail in noisy environments. Furthermore ASD enables a robot to generate meaningful backchannels in multi-speaker scenarios, e.g. head-tilt, eye-gaze towards the speaker in a group. While ASD has very active research and application in television and conferencing, the developed models do not fit the real-time and resource constraints of social robotics. In this work we aim to develop a low resource ASD based on lip features which is suitable for HRI applications. We aim to evaluate the developed models, pre- and postprocessing methods within HRI User Studies in noisy environments and for backchanneling.

2 Related Work

In the time from 2020 to 2025 3 new datasets for ASD have been published, all bringing their pros and cons. The AVA-ActiveSpeaker dataset [9] was published in 2020 and is the defacto-standard for ASD and is widely used in the literature. It contains about 38.5 hours of human labeled face tracks with frame level annotations, and the corresponding audio. The VVAD-LRS3 dataset [7] was published in 2023 and contains over 180 hours of automatically labeled face tracks with sample level annotations for each sample lasting 1.5 seconds. It comes without audio but provides different levels of preprocessed features for the face tracks. The UniTalk dataset [8] was published in 2025 and contains over 44.5 hours of manually labeled face tracks with frame level annotations, and the corresponding audio. It is designed overcome the limitations of the AVA-ActiveSpeaker dataset by providing a more diverse set of speakers, languages, and speaking styles.

The state-of-the-art (SOTA) models for ASD are based on deep learning which provides high accuracy but requires a lot of computational resources. TalkNet-ASD [10] was published in 2021 and is a deep learning model based on a combination of convolution based audio and visual encoding combined with cross and self attention for the fusion of the audio and visual features. It is trained on a combination of LRS3-TED [1] and VoxCeleb2 [5] called TalkSet, which was never published. It is evaluated on the AVA-ActiveSpeaker dataset and achieves a Mean Average Precision (mAP) of 92.3%. LoCoNet [11] was published in 2024 and introduces Short-term Inter-speaker Modeling to improve the mAP to 95.2% on the AVA-ActiveSpeaker dataset. It is trained on AVA-ActiveSpeaker and fine-tuned on the Talkies dataset [2]. LR-ASD

[6] was published in 2025 and is a lightweight model combining a simple channel attention module for multi-modal feature fusion with Gated Recurrent Units (GRUs) for temporal modeling. It achieves comparable results with a mAP of 94.5% on the AVA-ActiveSpeaker dataset while using 41 times less model parameters and 10 times less Floating Point Operations (FLOPs). While all of the above models are performing well the do not provide a full pipeline to generate the face tracks in real time from a video frame and therefore are not suitable for HRI applications, the PAZ framework [3] provides a pipeline to generate the face tracks in real time but only provides model with sample based predictions which pose another problem for HRI applications.

To evaluate models for ASD the most common metrics are mAP, Intersection over Union (IoU), and the F1-Score. For HRI applications additional metrics are required to evaluate the models in real world scenarios. These include Latency, Memory Footprint, Computational Efficiency, Task completion rate, False Activation Rate, and the Godspeed Questionnaire [4]

3 Research Gap

We define the research gap in the field of ASD for HRI as follows: While ASD has been extensively studied and applied in television broadcasting and conferencing, the existing models do not fully address the unique challenges posed in HRC. In television broadcasting, real-time processing is generally not a critical requirement, as automatic subtitling systems typically rely on asynchronous processing pipelines. Similarly, in conferencing scenarios, the absence of multi-speaker interactions simplifies the computational demands, allowing to skip parallel processing of multiple face tracks or rely solely on audio input. In contrast, social robotics presents significantly stricter constraints: systems must operate under limited computational resources (e.g., embedded platforms with constrained CPU/GPU capabilities) and often require concurrent execution of multiple models (e.g., speech recognition, natural language understanding, and perception), necessitating efficient resource sharing. Besides limited resources real-time responsiveness is a critical requirement. Consequently, the deployment of ASD in such environments demands careful optimization for latency, memory footprint, and computational efficiency. Additionally to the computational constraints for social robots the need to accurately evaluate the developed models in real HRI scenarios is currently not addressed in the literature.

4 Research Questions

For this work we formulate two Research Questions (RQs):

- RQ 1: How does ASD affect the accuracy and perceived naturalness of HRC in noisy and multi-speaker settings?
- RQ 2: Can ASD be implemented on constrained hardware without compromising performance?

5 Objectives

We list the following technical objectives for this work:

- Make the developed models, pre- and postprocessing methods publicly available to the research community via PyPI and GitHub.
- Document the developed models, pre- and postprocessing methods in a way that they can be easily used by other researchers.
- Release the developed models, pre- and postprocessing methods as a ROS2 package to enable easy integration into existing robotic systems.
- Release the ROS2 package as a pre-build Docker image to enable easy deployment on different platforms.

6 Timeline

To answer the RQs we define the following timeline for this work:

First Year

- Conduct a literature review on ASD and its application in HRI.
- Evaluate existing ASD models in the context of HRC in noisy environments and publish a paper on the findings.

Second Year

- Develop a low resource ASD model based on lip features suitable for HRI applications.
- Evaluate the developed model in comparison to SOTA models and publish a paper on the findings.

Third Year

- Conduct HRI User Studies to evaluate the developed models, pre- and postprocessing methods for backchanneling and publish a paper on the findings.
- Fulfil the technical objectives of this work.

References

[1] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. LRS3-TED: A large-scale dataset for visual speech recognition, 2018.

[2] Juan Leon Alcazar, Fabian Caba Heilbron, Ali K. Thabet, and Bernard Ghanem. MAAS: Multi-modal Assignation for Active Speaker Detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 265–274, Montreal, QC, Canada, October 2021. IEEE.

[3] Octavio Arriaga, Matias Valdenegro-Toro, Mohandass Muthuraja, Sushma Devaramani, and Frank Kirchner. Perception for Autonomous Systems (PAZ). *arXiv:2010.14541 [cs]*, October 2020.

[4] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. *International Journal of Social Robotics*, 1(1):71–81, January 2009.

[5] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. VoxCeleb2: Deep Speaker Recognition. In *Interspeech 2018*, pages 1086–1090. ISCA, September 2018.

[6] Junhua Liao, Haihan Duan, Kanghui Feng, Wanbing Zhao, Yanbing Yang, Liangyin Chen, and Yanru Chen. LR-ASD: Lightweight and Robust Network for Active Speaker Detection. *International Journal of Computer Vision*, 133(7):4749–4769, July 2025.

[7] Adrian Lubitz, Matias Valdenegro-Toro, and Frank Kirchner. The VVAD-LRS3 Dataset for Visual Voice Activity Detection. In *7th International Conference on Human Computer Interaction Theory and Applications*, pages 39–46, March 2023.

[8] Le Thien Phuc Nguyen, Zhuoran Yu, Khoa Quang Nhat Cao, Yuwei Guo, Tu Ho Manh Pham, Tuan Tai Nguyen, Toan Ngo Duc Vo, Lucas Poon, Soochahn Lee, and Yong Jae Lee. UniTalk: Towards Universal Active Speaker Detection in Real World Scenarios, May 2025.

[9] Joseph Roth, Sourish Chaudhuri, Ondrej Klejch, Radhika Marvin, Andrew Gallagher, Liat Kaver, Sharadh Ramaswamy, Arkadiusz Stopczynski, Cordelia Schmid, Zhonghua Xi, and Caroline Pantofaru. Ava Active Speaker: An Audio-Visual Dataset for Active Speaker Detection. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4492–4496, Barcelona, Spain, May 2020. IEEE.

[10] Ruijie Tao, Zexu Pan, Rohan Kumar Das, Xinyuan Qian, Mike Zheng Shou, and Haizhou Li. Is Someone Speaking? Exploring Long-term Temporal Features for Audio-visual Active Speaker Detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3927–3935, October 2021.

[11] Xizi Wang, Feng Cheng, and Gedas Bertasius. LoCoNet: Long-Short Context Network for Active Speaker Detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18462–18472, Seattle, WA, USA, June 2024. IEEE.