Active Speaker Detection for Humanoid Robots in Human-Robot Interactions

Adrian Auer[1], Arunima Chaurasia[2] and Priya George[3]

I. Contribution

We contribute to the field of Active Speaker Detection (ASD) for Human-Robot Interaction (HRI) by providing a comprehensive evaluation of existing ASD pipelines on the UniTalk dataset [1], with a specific focus on real-time performance on resource-constrained hardware. Our work includes the following key contributions:

- We implemented a lipshape-based ASD approach as a lightweight alternative to the more complex state-of-the-art (SOTA) models,
- We benchmark several SOTA ASD pipelines, including TalkNet [2], LoCoNet [3], LR-ASD [4], and models provided by PAZ [5] against our lipshape-based approach on the UniTalk dataset [1] to assess their performance under real-time constraints.
- We developed a benchmark to evaluate in which scenarios models perform best to make informed decisions about model selection for real-world applications.
- We open-sourced our development and evaluation code on GitHub to facilitate further research and development in the field of ASD for HRI.

II. Motivation

Recent advances in ASD have significantly improved benchmark performance, yet most high-performing systems depend on large deep neural networks and Graphics Processing Unit (GPU)-accelerated inference to sustain real-time operation. This dependency is a major limitation for humanoid HRI: a robot must simultaneously run multiple perception and interaction modules, including speech processing, natural language understanding, and vision-based scene analysis. In this setting, a GPU-bound ASD solution is often not an acceptable design choice, because the available compute, memory, and power budget is shared across the entire robot stack. As a result, many current ASD systems are strong in isolation but difficult to deploy on embodied platforms where reliable, responsive behavior must be maintained under real-world constraints. This creates a clear need for lightweight ASD models that can run in real time on embedded hardware such as a Raspberry Pi, even if this comes at a modest cost in accuracy. Such systems are not merely a pragmatic compromise; they are essential for enabling humanoid robots to integrate conversational grounding, social attention, and multimodal interaction in a scalable and deployable manner. Our work addresses this gap by targeting real-time ASD on resource-constrained hardware, thereby improving the practical utility of humanoid robots in HRI scenarios where overall system robustness matters more than maximizing ASD accuracy alone.

[1] Robotics Research Group, University of Bremen, 28359 Bremen, Germany a.auer@uni-bremen.de

[2] German Research Center for Artificial Intelligence, 28359 Bremen, Germany arunima.chaurasia@dfki.de

[3] German Research Center for Artificial Intelligence, 28359 Bremen, Germany priya.george@dfki.de

III. Related Work

ASD has been studied extensively in recent years, but most of the strongest datasets and pipelines still originate from television-style data. The AVA-ActiveSpeaker dataset [6] became a de facto training and evaluation resource for many modern audio-visual (AV) ASD models, yet its public benchmark does not provide a test set that can be used to verify generalization to HRI scenarios. In contrast, UniTalk [1] was explicitly designed as a more challenging, real-world dataset for ASD and is therefore a better reference point for evaluating robustness outside broadcast video. The VVAD-LRS3 dataset [7] is another important resource that focuses on vision-only (VO) ASD. On the model side, the current literature includes TalkNet [2], which introduced long-term temporal modeling for AV ASD, LoCoNet [3], which improves context handling in multi-speaker scenes, and LR-ASD [4], which targets lightweight deployment on constrained hardware. PAZ [5] offers a variety of pipelines for different vision-based tasks for autonomous systems including several models in the ASD pipeline. Together, these datasets and models define the current ASD landscape: AVA-ActiveSpeaker as a dominant but TV-oriented training source, UniTalk as a more realistic evaluation target, and VVAD-LRS3 as a VO benchmark that complements the broader AV literature.

IV. Methodology

We evaluate the selected ASD pipelines on the validation split of Unitalk. The videos are annotated with face tracks and speech activity labels, which allow us to compute frame-level metrics for the Visual Voice Activity Detection (VVAD) models. The evaluation is designed to reflect the full pipeline performance, including preprocesing e.g. face detection and face shape prediction.

We compare the PAZ family pipelines (VVAD-LRS3-LSTM, CNN2Plus1D, CNN2Plus1D_Filters, CNN2Plus1D_Layers, CNN2Plus1D_Light), our lipshape-based implementation, TalkNet, LR-ASD, and LoCoNet. Face tracks are generated from OpenCV face detections using a

heuristic ordering strategy without any tracker. This means that the evaluation captures both the quality of the VVAD model and detection backend.

For model quality, we report frame-level accuracy, precision, recall, and F1 score. These metrics are computed per video and then aggregated into an overall score for the validation split. In addition to the model metrics, we report a detection fail rate that captures False Positive (FP) and False Negative (FN) detections.

To understand the deployment trade-off between accuracy and speed, we measure prediction time as the full pipeline latency, including pre-processing and post-processing, and report speed in frames per second.

The runtime evaluation is performed on the Raspberry Pi 4 B and on a desktop system with a GPU, using the same pipeline configurations on both platforms.

To analyze performance in a setting that is closer to HRI than to broadcast television, we additionally developed a benchmark derived from the annotations. The benchmark provides automatically extracted rating for the following categories: Diversity (number of unique entities), Interactivity (number of entities per frame), Audibility (rate of non-audible speech), Occlusion (pairwise IoU over all face pairs), Dynamic (motion of the detected entities over time), and Engagement (face-track length). For each category, we log the average, maximum, and variance where applicable, and we use these attributes as stratification criteria when analyzing the results. This benchmark is intended to expose where the pipelines fail or succeed under HRI-like conditions, where we expect lower interactivity, lower occlusion, lower dynamics, and longer engagement compared to television material.

V. Preliminary Results

TABLE I
Model Performace across UniTalk Dataset (val split)

Model	Accuracy	FPS	F1	Precision	Recall
CNN2Plus1D	0.6561	17.40	0.3901	0.2765	0.6624
CNN2Plus1D_Filters	0.7592	17.29	0.4447	0.3603	0.5809
CNN2Plus1D_Layers	0.8185	11.31	0.4296	0.4505	0.4105
CNN2Plus1D_Light	0.8044	20.97	0.4533	0.4237	0.4874
LipShape	0.8180	27.38	0.3094	0.1994	0.6899
LipsShape (PI)	0.7718	8.81	0.2905	0.1891	0.6266
CNN2Plus1D_Light (PI)	0.6719	2.95	0.5112	0.4309	0.6282

While this work is still in progress our preliminary results show that the lipshape-based model performs comparably to the CNN2Plus1D-based models in terms of accuracy, while being significantly faster. On the Raspberry Pi 4 B, the lipshape-based model achieves 8.81 FPS, while the fastest CNN-based model (CNN2Plus1D_Light) only achieves 2.95 FPS. Although we could not reach real-time on the Raspberry Pi 4 B, this performance difference provides a strong hint that a lipshape-based approach may be a good compromise for real-time applications on constrained hardware.

We noticed that beside the performance of the models on the predictions they make there is a large number of frames that were missed by the detection and tracking backend and therefore do not have any predictions at all. On average 63.16% of the frames in the validation set do not have any predictions. We noticed though that this is tightly linked to the length of the video and the scene cuts, longer videos miss around 20% of the frames, while very short videos miss up to 97% of the frames. At this point we could not draw any conclusions about the performance of the models on the different categories of our benchmark, which suggests that there is no single model that performs best across all conditions.

VI. Discussion

[8] highlights that ASD in HRI improves the quality of the interaction in noisy environments, but application on a mobile robot platform is still an open challenge. Additionally we believe that backchanneling gaze through ASD can improve the quality of the interaction. We are aiming to deepen our evaluation of available models compared to our lipshape-based implementation to highlight that compromising accuracy for speed can be a good choice for real-time applications on constrained hardware. We also want to show that there is still work to achieve in the detection and tracking backend, which is an important part of the pipeline that is often overlooked in the literature. Additionally we want to draw a better conclusion from the fine-grained analysis provided with our benchmark.

References

[1] L. T. P. Nguyen, Z. Yu, K. Q. N. Cao, Y. Guo, T. H. M. Pham, T. T. Nguyen, T. N. D. Vo, L. Poon, S. Lee, and Y. J. Lee, "UniTalk: Towards Universal Active Speaker Detection in Real World Scenarios," May 2025, arXiv:2505.21954 [cs.CV]. [Online]. Available: http://arxiv.org/abs/2505.21954

[2] R. Tao, Z. Pan, R. K. Das, X. Qian, M. Z. Shou, and H. 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*, Oct. 2021, pp. 3927–3935, arXiv:2107.06592 [eess]. [Online]. Available: http://arxiv.org/abs/2107.06592

[3] X. Wang, F. Cheng, and G. Bertasius, "LoCoNet: Long-Short Context Network for Active Speaker Detection," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, WA, USA: IEEE, Jun. 2024, pp. 18 462–18 472. [Online]. Available: https://ieeexplore.ieee.org/document/10656597/

[4] J. Liao, H. Duan, K. Feng, W. Zhao, Y. Yang, L. Chen, and Y. Chen, "LR-ASD: Lightweight and Robust Network for Active Speaker Detection," *Int J Comput Vis*, vol. 133, no. 7, pp. 4749–4769, Jul. 2025. [Online]. Available: https://link.springer.com/10.1007/s11263-025-02399-2

[5] O. Arriaga, M. Valdenegro-Toro, M. Muthuraja, S. Devaramani, and F. Kirchner, "Perception for Autonomous Systems (PAZ)," *arXiv:2010.14541 [cs]*, Oct. 2020, arXiv: 2010.14541. [Online]. Available: http://arxiv.org/abs/2010.14541

[6] J. Roth, S. Chaudhuri, O. Klejch, R. Marvin, A. Gallagher, L. Kaver, S. Ramaswamy, A. Stopczynski, C. Schmid, Z. Xi, and C. Pantofaru, "AVA-ActiveSpeaker: An Audio-Visual Dataset for Active Speaker Detection," May 2019, arXiv:1901.01342 [cs]. [Online]. Available: http://arxiv.org/abs/1901.01342

[7] A. Lubitz, M. Valdenegro-Toro, and F. Kirchner, "The VVAD-LRS3 Dataset for Visual Voice Activity Detection," Mar. 2023, pp. 39–46. [Online]. Available: https://www.scitepress.org/PublicationsDetail.aspx?ID=unPCBBhJ+SU=&t=1

[8] A. Gopikrishnan, A. Auer, and L. Gutzeit, "Improving Human-Robot Communication in Noisy Environments with Visual Voice Activity Detection," in *Computer-Human Interaction Research and Applications*, J. F. Krems, H. P. da Silva, and P. Cipresso, Eds. Cham: Springer Nature Switzerland, 2026, pp. 71–91.