Audio-Visual Speech Recognition for Human-Robot Interaction: a Feasibility Study

Sander Goetzee, Konstantin Mihhailov, Roel van de Laar, Kim Baraka, and Koen V. Hindriks[1]

***Abstract*— Recent models for Visual Speech Recognition (VSR) have shown remarkable progress over the last few years. They have however been applied mainly to datasets such as Lip Reading Sentences 3 (LRS3), LRS2 or Lombard GRID, but not yet on social robots. As social robots struggle to recognize speech in more challenging acoustic and crowded environments, we believe such models are promising tools for real-time interaction with users. This paper presents a feasibility study focusing on integration of speech recognition (SR) using mixed modalities - audio, visual (lip-reading) and audio-visual - in social robots. To this end, this paper contributes a pipeline to detect an active speaker based on lip movement, post-processing of audio and video footage and inferencing it with the state-of-the-art Auto-AVSR model. In a user study ($N = 26$), we evaluated the feasibility of audio, visual and mixed modality speech recognition on a Pepper robot. We demonstrate the feasibility of using singular and mixed modalities with speech-to-text inference in natural interaction. The results show that it is feasible to deploy such models on social robots in a controlled, noiseless and non-interactive environment. Additionally, the results revealed that informing participants to emphasize their lip movements significantly improved text-to-speech inference results. Our work provides initial insights into the benefits and challenges of using VSR, ASR and AVSR for HRI.**

I. INTRODUCTION

Effective (social) human-robot interaction (HRI) hinges on accurately recognizing what a person is saying. Envisioning a cocktail party scenario [1], characterized by its cacophony of competing auditory and visual stimuli, this setting is difficult for both human and social robot in perceiving speech. While humans often use both visual and auditory signals to correctly decipher speech [2], robots currently only rely on auditory information. This work aims to investigate the use of a visual channel as a complement for the audio channel in speech recognition in HRI.

The analysis and processing of audio speech recognition (ASR), visual speech recognition (VSR), and combined modalities, audio-visual speech recognition (AVSR), in this context involves distinct challenges due to the unique properties and interference factors of each modality [3]. In the auditory domain, this includes acoustic reverberations, multitude of actors talking in the background and diverse noise distributions. In the visual domain, there are challenges with lighting variances and speaker face occlusion [3]. When it comes to social robots, a variety of additional limitations present themselves. The robot's camera plays an important role in its visual modality, a common issue that arises from video recording using a robot's camera is that the frame rate is variable and demonstrates erratic behaviour. Furthermore, there is a balancing act between higher resolution and lower frame rate and vice versa. This factor can also influence the latency between input and output. Additionally, as social robots are not static, a moving camera can cause issues for video capture through the camera by introducing blurring and artifacting. Distance from the robot can have an affect on both visual and audio capture. In the visual domain, the capture of facial features can be affected by the distance from the robot and further impacted by the resolution. Furthermore, the robot's microphone(s) can have issues in capturing human speech due to hardware quality, distance or noise. However, it is important to note that this study seeks evidence that visual modality might be efficacious under controlled, noiseless and non-interactive environment.

This work addresses the aforementioned challenges by effectively combining the three modalities while taking potential limitations of social robots into account. We make the following contributions:

- *A pipeline for integrating an Audio-Visual Speech Recognition (AVSR) system, specifically the recent Auto-AVSR model [4], into a social robot.*
- *An evaluation of the pipeline with 26 participants and a Pepper social robot by assessing the three modalities for SR, namely ASR, VSR, and AVSR.*
- *Recommendations for implementing an HRI-compatible AVSR system for social robots.*

II. RELATED WORK

ASR is widely utilised in HRI contexts involving natural language interaction. Notably, in environments with low levels of noise, the word correct rate (WCR) for audio only modality can exceed 95% [5]. However, the effectiveness of ASR diminishes in noisy environments [6], a challenge particularly noticeable in dynamic settings where social robots operate. Addressing this, researchers have proposed incorporating visual modalities, specifically lip-reading methods, to enhance speech recognition robustness since the eighties. This approach stems from [7], who proposed that audio-visual approaches are more effective than methods based on singular audio or video inputs.

A. Visual modality and its impact on Speech Recognition

A variety of research has been done on complementing ASR with visual information. Ephrat et al. (2017) [8] used a model that turned silent videos into speech. This model helped improve speech sounds in noisy environments by using visual cues instead of the actual noisy audio. Other

[1] Vrije Universiteit Amsterdam, 1081HV Amsterdam, The Netherlands; corresponding author: k.v.hindriks@vu.nl

researchers developed a neural network that used computer vision to both improve speech sounds and recreate images of lip movements, highlighting the importance of using visual information to support speech [9]. Others have utilised Convolution Neural Networks (CNNs) to encode features from both video and audio, to unify the embedding between these two modalities before decoding audio [10]. The authors of [11] used temporal convolution networks (1D ResNet) to independently encode video and audio data. This allowed them to decode these into a mask, for the purposes of eliminating noisy components in audio. The aforementioned research illustrates the diversity in approaches researchers have taken to integrate visual information with audio modality. However, when it comes to HRI, this integration is not so transparent.

B. Visual Speech Recognition in HRI

To our knowledge, limited research has been dedicated to integrating VSR exclusively into robotic systems. While several studies have explored visual aspects of HRI, such as a robot's ability to understand intentions [12] or assess human attention levels [13], a comprehensive examination of VSR in this context is less common.

Among the few endeavors to incorporate VSR into robotic systems, researchers have aimed to enhance interactions between a RASA robot and hearing-impaired users in Iran through the development of neural network models (CNN-LSTM and 3D-CNN). These models, designed to facilitate automated lip-reading by robots, showed promise when assessed using the OuluVS2 dataset, demonstrating the potential of straightforward, efficiently trained networks for lip-reading in robotic applications [14]. However, the application of these models on a robotic system was tested with a dataset, suggesting room for further exploration, particularly in analyzing VSR in robotic systems through real-time camera feeds.

C. Combined Audio-Visual Speech Recognition

Integration of lip-reading into non-robotic ASR has been approached from various angles. Key developments include LipNet [15], which produced 4.8% word error rate (WER) on GRID corpus [16]. Higher accuracy of 3.0% WER was achieved on GRID corpus with the encoder-to-decoder architecture [17]. Furthermore, [18] examined the performance of transformer self-attention architecture, utilising encoder-to-decoder mechanism, a common concept in speech recognition and translation [19]. In this case this approach was tailored to suit multi-modal integration of data from multiple modalities [18]. This approach resulted in 7.2% WER on LRS3 dataset and 8.5% WER on LRS2 dataset. However, this approach demonstrated increased difficulty with noisy audio signals, with WERs rising to 42.5% and 34.2% on these datasets. Finally, a novel approach comes from a recent study that tackles the challenge of transcribing continuous spoken sentences from both audio and visual streams [4]. This research acknowledges the complementary nature of visual and audio information, particularly in environments with high levels of noise. This Auto-AVSR model stands out for its approach to overcoming one of the primary challenges in AVSR: the dependency on large-scale, well-labelled datasets.

D. Auto-AVSR

The Auto-AVSR model, notable for recent advancements in VSR and AVSR benchmarks, focused on improving its utility in automatic contexts [20] [4]. This is accomplished by training the model with labeled data, which is then utilized to independently label data that is not annotated. In this case, 'automatic' refers to the automated process of converting spoken language into written text using audio and visual inputs, in other words, the transcription process. However, achieving a real-time effect is a difficult task. Due to this, it is important for the model to be able to deal with incoming information in a smart way. The model either utilizes pre-existing data, such as visual or audiovisual records, or gathers incoming information through audio-visual equipment like microphones and cameras, depending on the implementation used. The pre-processing application does data cleanup and then the information is passed on to the model for inferencing [4]. The Auto-AVSR model stands out for its advancements in AVSR, primarily attributed to the utilization of large models and extensive training sets.

As a result, facilitating the training of ASR, VSR, and AVSR models led to significant improvements in performance, with the Auto-AVSR model achieving a WER of 0.9% on the LRS3 dataset [4]. Notably, the VSR modality, focusing solely on lip-reading, recorded a WER of 19.1% on the LRS3 dataset.

III. AVSR PIPELINE

To deploy the Auto-AVSR [4] inferencing model onto a social robot, we build a pipeline that consists of several stages; (1) Visual Speech Detection using lip movement, (2) the recording of audio and video stream when speech is detected, (3) post-processing of the audio and video chunks after speech detection, and (4) the inferencing of the post-processed chunks by the Auto-AVSR [4] model. A visual representation of the pipeline is shown in figure 1. We implemented this pipeline within the Social Interaction Cloud (SIC) framework [21] on the Pepper robot [22].

A. Visual Speech Detection (VSD)

Building upon Mediapipe's landmark detection [23], several key additions have been implemented to facilitate the detection and calculation of lip movements. We utilize both the upper and lower lip landmarks with their outer and inner indices for more robust lip movement identification. For our pipeline we make the assumption that when there is significant lip movement, the person is speaking. Another modification is the introduction of a temporal lip movement buffer, which retains the data over a series of frames. This buffer, with a configurable size, allows for temporal analysis of lip movements, accommodating the detection of active speaking phases. The VSD calculates the centre point of the

lips for each frame and assesses the relative movement by comparing current and previous frames of lip positions. This visual approach enables detection of speech while avoiding potential pitfalls of audio based speech detection, such as noise, which is common in HRI. Our VSD is designed to continuously analyze a video feed and to start and stop recording audio and video automatically based on detection of speaking activity from the human. We calibrated this based on the results of a pilot study with 6 participants.

B. Audio-video recording

When the system detects significant lip movement (indicating speech), it begins the recording of both the video and audio streams; conversely, recording stops when speech is no longer detected. The video feed from Pepper is variable and low considering the requirements of the Auto-AVSR model. As we need a stable frame rate, we record the timestamps of the first and last frames. To achieve a stable frame rate we divide the total amount of frames by the duration (in seconds) between the two timestamps which results in approximately 9 to 11 frames fer second (fps). An obvious solution would be to use a more capable camera. However, this would defeat the purpose of our feasibility study for standard social robot platforms and affect reproducibility.

C. Post-processing and inferencing

After recording, the video and audio chunks are then placed into a queue for post processing. As the Auto-AVSR [4] inferencing model requires a stable frame rate between 25 and 30 fps, our next step involves interpolation using motion vectors between the frames of the video chunks, using FFmpeg [24]. This process generates the missing frames between the existing frames to achieve the required frame rate of 25 fps. Finally, the video and audio chunks are merged into a singular chunk which is ready for inferencing by Auto-AVSR [4].

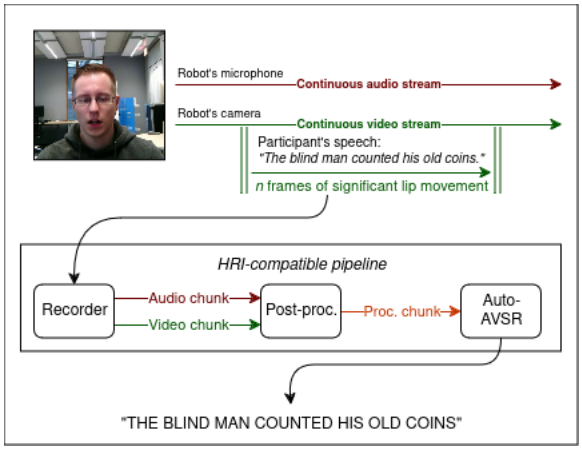

Fig. 1. Flow diagram of our multi-stage AVSR pipeline

IV. FEASIBILITY STUDY

Our study was conducted in accordance with the faculty guidelines on ethical research and was exempted from a full review according to an online self-check. Audio and video recordings are not retained, ensuring data anonymity through the removal of identifiable information.

A. Participants

We used convenience sampling at the Vrije Universiteit Amsterdam to collect data to evaluate our pipeline. We recruited 26 participants (14 females, 12 males, 0 non-binary and 0 that did not disclose their gender; participants were aged between 18 and 34 years). Two participants disclosed a diagnosis of dyslexia. All other participants indicated they had no impairments related to hearing, reading, or speech abilities. Participants reported their native language: Dutch (15), German (2), Romanian (2), English, Italian, Mandarin, Polish, Portuguese, Telugu, and Turkish. Participants were also asked to self-rate English language proficiency on a scale of 1 (not proficient) to 5 (native-level skills). All participants self-rated as having at least intermediate level skills, capable of handling standard language situations with some ease. Several participants rated themselves as either advanced or possessing native-level competence. Advanced proficiency indicates being able to interact in English with fluency and spontaneity, facilitating regular interaction with native speakers without strain for either party.

B. Design

The study followed a within-subject design, where each participant was asked to read a set of unique Semantically Unrelated Sentences (SUS) [25], also called Harvard sentences which is the standard for measuring speech inference quality [26]. These sentences are phonetically balanced and are developed to have low semantic predictability. The participants were asked to read 10 sentences under two different approaches. For the first approach the participants were not informed on the robots capabilities. During the second approach the participants were informed about the robots lip-reading capabilities, and were asked to enunciate. This design allows for comparing the performance of the speech recognition for natural speech and enunciated speech. For this feasibility study, we hypothesize that the accuracy of speech inference will be better for the second approach compared to the first approach. Additionally, we hypothesize that the AVSR model will outperform both VSR and ASR models.

C. Materials and Measures

The study was conducted in a controlled lab setting. The lab provided a well lit environment, with a light right above Pepper and the participant, to facilitate reading the sentences from the tablet. We used the robot Pepper V1.8a. To guarantee a non-interactive environment, the camera is always directed at the face of a participant. To maintain consistent and standardized conditions for evaluating the robot's lip-reading performance, the robot was configured to remain stationary, by disabling movement of its limbs, body, and head. Because the tablet is positioned in close proximity to the camera, there was no need for participants to tilt their

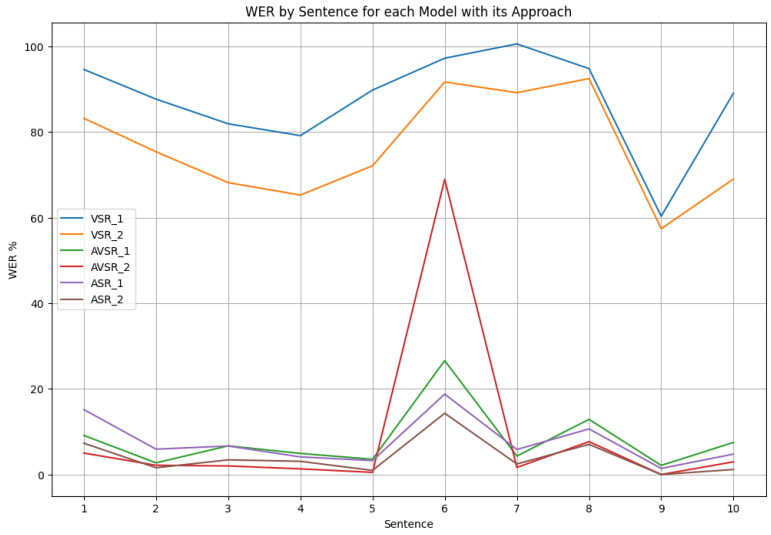

Fig. 2. Sentence based line-plot across modalities.

head when reading. The tablet on the chest of the robot was used as an autocue to display Harvard sentences, specifically the sentences from list 72 [26].

We used Word Error Rate (WER) as our measure for evaluating the speech recognition performance of the robot. WER is a common metric in speech recognition tasks that quantifies errors by calculating the sum of substitutions (S), insertions (I), and deletions (D), normalized by the total number of spoken words (N) [27]:

$$\text{WER} = \frac{S + I + D}{N} \times 100$$

For the latency metric, we used the CPU computation time for each stage of the pipeline, which is in seconds. Additionally, at the end of the experiment a survey with three open-ended questions was used to collect more qualitative data alongside five sample informative questions. The open-ended questions were focusing on the self-perceived enunciation of participant's own speech.

D. Procedure

Participants were greeted by one of the researchers in the lobby and escorted to the lab one at a time. They were seated in a designated chair positioned one meter away from Pepper. After a brief introduction the participants were instructed to read the first 10 sentences displayed on the robot's tablet. After finishing the first set of sentences, a second briefing commenced and participants read the second set of 10 sentences. Upon finishing both sets, participants were asked to fill in a survey, debriefed about the purpose of the study and asked to provide their consent again to use the data collected in the session.

V. RESULTS

To analyze issues in how participants read sentences, we employed methods similar to those in [28]. We meticulously transcribed all video recordings and used a cross-validation process for any ambiguous instances, involving multiple team members to verify the interpretation of participant utterances. This was done to address the issue of potential misreadings by the participants, which would in turn give a wrong impression of the models interpretation performance.

	Including Sentence 6		Excluding Sentence 6	
Model	M (%)	SD	M (%)	SD
AVSR_1	8.1	13.8	6.0	11.5
AVSR_2	9.5	82.7	2.6	7.2
VSR_1	87.8	25.8	86.7	26.1
VSR_2	76.6	29.0	74.9	29.1
ASR_1	7.6	14.1	6.4	13.1
ASR_2	4.2	9.0	3.1	7.9

TABLE I
MEANS AND STANDARD DEVIATION FOR WER

The results, as detailed in Table V, show quite some variability in WER across the three modalities. Note that we use subscripts to indicate the data obtained from the first, uninformed reading of sentences (1) and the data obtained from the second, informed reading of sentences (2). A notable aspect of the collected data, is that for reasons not clear to us, the AVSR inferencing for sentence six lead to a significant increase in standard deviation only in the informed case. This sentence in particular contributed significantly to errors in AVSR speech-to-text inferencing and proved challenging across all modalities, as shown in line-plot in Fig. 2. Specifically, in the case of informed AVSR, it led to a marked increase in WER, often ending in 'the grass and bushes were we' without continuation, despite the video and audio evidence of participants completing the sentence. This discrepancy marks the sentence as an anomaly in our data set. Consequently, to maintain the integrity of our analysis, we opted to present the results both with anomaly and without.

To compare the different modalities and evaluate which group (informed, or uninformed) for each modality per-

formed significantly better, we first needed to determine whether the collected data followed a normal distribution. We plotted the data points of each modality in QQ-plots and conducted Shapiro-Wilk tests to assess if the normality assumption was violated. The results indicated that the assumption of normality was indeed violated for all datasets. Therefore, we employed Wilcoxon signed-rank tests to compare the combined modality against the individual modalities. Our hypothesis was that the WER of AVSR would be lower than that of VSR and ASR.

When considering all sentences, the results indicate that there are significant differences between AVSR and VSR for both the uninformed ($W = 0.0$, $p < .001$) and informed ($W = 243.0$, $p < .001$) groups. However, there are no significant differences between AVSR and ASR in either group ($W = 1234.0$, $p = .855$, $W = 500.5$, $p = .730$).

The results excluding sentence 6 indicate that there are significant differences between AVSR and VSR for both the uninformed ($W = 0.0$, $p < .001$) and informed ($W = 0.0$, $p < .001$) groups. However, there are no significant differences between AVSR and ASR in either group ($W = 527.5$, $p = .349$, $W = 133.0$, $p = .213$).

Furthermore, in comparing each modality, we hypothesized that the WER of the uninformed group would be greater than that of the informed group. To test this hypothesis, we conducted Wilcoxon signed-rank tests for AVSR, VSR, and ASR.

These tests yielded the following statistics: $W = 3275.5$, 16332.0, 2958.5 for the datasets including sentence 6, and $W = 2060.0$, 13163.5, 1877.0 for the datasets excluding sentence 6 across the three modalities tested. The consistent finding of $p < .001$ across all results confirmed our hypothesis, indicating that the WER of the uninformed group surpasses that of the informed group across all modalities, irrespective of whether sentence 6 is considered or not.

Regarding latency metrics, each component of the pipeline exhibits distinct average latency values: the recording stage has an average latency of 0.09 seconds, whereas the post-processing phase for each data chunk incurs an average latency of 1.03 seconds. Within the Auto-AVSR framework, the latency for AVSR stands at 2.12 seconds, for VSR at 1.86 seconds, and for ASR at 0.53 seconds per chunk. It is noteworthy that improvements in hardware specifications are directly correlated with reductions in latency durations.

The analysis of open-ended qualitative responses revealed noteworthy patterns in speech modification strategies. During the uninformed approach, the majority of the 26 participants indicated a slight degree of adjustment in their speech. Some changes included speaking louder, and at a slower pace. Most noted no significant change in their speech style, suggesting a natural approach to the task.

Study Specific Limitations: The study aimed at performing a first step to investigate the feasibility of applying audiovisual speech recognition for social robotics. Perhaps one of the most notable limitations is that our study was performed in the lab, and we designed our study to maintain a consistent as possible setup for each participant. As a result, we did not take into account, for example, variable participant heights, which need to be dealt with in natural HRI settings and brings challenges of its own.

VI. DISCUSSION

In our study, AVSR performance compared to individual ASR and VSR modalities aligns with multi-modal speech recognition in Auto-AVSR models. Auto-AVSR achieved WERs of 0.9% for AVSR and 1% for ASR on the LRS3 dataset [4], while we recorded 2.6% for AVSR and 3.1% for ASR (Informed). Although Auto-AVSR excels on benchmark datasets, our real-world HRI context faced unique challenges, resulting in higher WERs. This highlights the difficulty of adapting these technologies from controlled datasets to dynamic HRI environments. The performance gap between AVSR and ASR may widen further in noisy conditions.

Our VSR system performed poorly on selected Harvard sentences, with a notable disparity: Auto-AVSR recorded a 19.1% WER, while we recorded 74.9% in VSR (Informed). Factors include the scarcity of native English speakers, interpolated footage, and model generalization limitations, affecting its ability to process semantically ambiguous sentences.

For sentence six, "The grass and bushes were wet with dew," we hypothesize model weight issues affecting AVSR inferencing, with unclear modality influence. Notably, native speakers correctly inferred this sentence, indicating potential training data bias.

Among the 26 participants, only one was a native English speaker, raising concerns about model bias related to accent impact. Non-native accents likely affected WER results, influenced by the Auto-AVSR model's English-dominant training data. This factor contributed to poor VSR performance. The LRS3 dataset, including TED and TEDx talks, emphasizes clear speech. Our results show that emphasizing lip movement in the informed approach enhances performance, but a significant gap remains compared to Auto-AVSR on standard datasets.

VII. CONCLUSION AND FUTURE WORK

Our results show that asking participants to enunciate significantly improved text-to-speech inference results across all modalities (speech, visual, or combined). This suggests that clearer user communication can enhance speech recognition accuracy in social robots, especially in visually dependent modalities.

Despite advances, a gap remains between the performance of state-of-the-art models in benchmarks and their performance in social robots. Key limiting factors for social robots using visual input include variable and low frame rates and low-resolution video capture, which significantly impact visual speech recognition. Addressing these challenges requires balancing technical constraints and user experience, a task for future developments in social robotics and AVSR technology.

Our work identifies several areas for improvement in lip-reading technology and AVSR in human-robot interaction (HRI):

1) Pipeline Design: We recommend developing a modular pipeline that supports interchangeable models for active speech detection and speech recognition based on the HRI environment. This flexibility enhances system adaptability. Key areas for future work include:

- *Aligning video and audio*: Implementing systems to accurately timestamp video frames and corresponding audio fragments to ensure temporal alignment.
- *Using ASD for segmentation*: Accurate active speaker detection is needed to obtain precise timestamps for AVSR inference.
- *Building a bridge between hardware limitations and model requirements*: There is often a mismatch between available hardware capabilities and the requirements of models. A more modular approach is necessary to effectively bridge these gaps, ensuring that robotics hardware can support evolving model demands.

2) Training on HRI-specific data: Training AVSR models on HRI-specific data that includes unstable and low frame rate scenarios can enhance text-to-speech performance and reduce latency for more effective social robot deployment. Utilizing data captured directly from social robots can reduce dependency on stable 25 fps data and eliminate the need for interpolation. Moreover, incorporating diverse, natural conversational speech, including various accents, can better prepare models for real-world interactions.

In conclusion, while AVSR holds promise for HRI applications, attention must be given to participant diversity, speech types, synchronization challenges, and hardware capabilities. Future research should focus on training models with diverse, real-world HRI data and developing flexible, modular systems that can adapt to dynamic HRI conditions. Additionally, it is essential to explore the impact of background noise and interactivity on performance when using the visual modality for speech recognition. This study provided initial insights into the visual modality's potential to improve ASR in controlled environments, with future work to include the addition of background noise.

REFERENCES

[1] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the acoustical society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[2] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–748, 1976.

[3] I. D. Gebru, "Audio-visual analysis in the framework of humans interacting with robots," Ph.D. dissertation, Université Grenoble Alpes, 2018.

[4] P. Ma, A. Haliassos, A. Fernandez-Lopez, H. Chen, S. Petridis, and M. Pantic, "Auto-avsr: Audio-visual speech recognition with automatic labels," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[5] J. Shin, J. Lee, and D. Kim, "Real-time lip reading system for isolated korean word recognition," *Pattern Recognition*, vol. 44, no. 3, pp. 559–571, 2011.

[6] I. Q. Habeeb, T. Z. Fadhil, Y. N. Jurn, Z. Q. Habeeb, and H. N. Abdulkhudhur, "An ensemble technique for speech recognition in noisy environments," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 18, no. 2, pp. 835–842, 2020.

[7] E. D. Petajan, *Automatic lipreading to enhance speech recognition (speech reading)*. University of Illinois at Urbana-Champaign, 1984.

[8] A. Ephrat, T. Halperin, and S. Peleg, "Improved speech reconstruction from silent video," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 455–462.

[9] J.-C. Hou, S.-S. Wang, Y.-H. Lai, Y. Tsao, H.-W. Chang, and H.-M. Wang, "Audio-visual speech enhancement using multimodal deep convolutional neural networks," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 117–128, 2018.

[10] A. Gabbay, A. Shamir, and S. Peleg, "Visual speech enhancement," *arXiv preprint arXiv:1711.08789*, 2017.

[11] T. Afouras, J. S. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement," *arXiv preprint arXiv:1804.04121*, 2018.

[12] Z. Li, Y. Mu, Z. Sun, S. Song, J. Su, and J. Zhang, "Intention understanding in human–robot interaction based on visual-nlp semantics," *Frontiers in Neurorobotics*, vol. 14, p. 610139, 2021.

[13] P. Chakraborty, S. Ahmed, M. A. Yousuf, A. Azad, S. A. Alyami, and M. A. Moni, "A human-robot interaction system calculating visual focus of human's attention level," *IEEE Access*, vol. 9, pp. 93 409–93 421, 2021.

[14] A. Gholipour, A. Taheri, and H. Mohammadzade, "Automated lip-reading robotic system based on convolutional neural network and long short-term memory," in *Social Robotics: 13th International Conference, ICSR 2021, Singapore, Singapore, November 10–13, 2021, Proceedings 13*. Springer, 2021, pp. 73–84.

[15] Y. M. Assael, B. Shillingford, S. Whiteson, and N. De Freitas, "Lipnet: End-to-end sentence-level lipreading," *arXiv preprint arXiv:1611.01599*, 2016.

[16] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[17] J. Son Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6447–6456.

[18] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 12, pp. 8717–8727, 2018.

[19] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[20] Y. Shi, Y. Wang, C. Wu, C.-F. Yeh, J. Chan, F. Zhang, D. Le, and M. Seltzer, "Emformer: Efficient memory transformer based acoustic model for low latency streaming speech recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6783–6787.

[21] Social AI Lab, Vrije Universiteit Amsterdam, "Social Interaction Cloud (SIC) V2," https://socialrobotics.atlassian.net/wiki/spaces/CBSR, 2023, accessed: 20-03-2024.

[22] A. K. Pandey and R. Gelin, "A mass-produced sociable humanoid robot: Pepper: The first machine of its kind," *IEEE Robotics & Automation Magazine*, vol. 25, no. 3, pp. 40–48, 2018.

[23] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee *et al.*, "Mediapipe: A framework for building perception pipelines," *arXiv preprint arXiv:1906.08172*, 2019.

[24] F. Developers, "Ffmpeg," http://ffmpeg.org/, 2024.

[25] C. Benoît, M. Grice, and V. Hazan, "The sus test: A method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences," *Speech communication*, vol. 18, no. 4, pp. 381–392, 1996.

[26] E. Rothauser, "Ieee recommended practice for speech quality measurements," *IEEE Transactions on Audio and Electroacoustics*, vol. 17, no. 3, pp. 225–246, 1969.

[27] A. Ali and S. Renals, "Word error rate estimation for speech recognition: e-WER," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 20–24. [Online]. Available: https://aclanthology.org/P18-2004

[28] G. Krynicki, K. Dziubalska-Kołaczyk, J. Weckwerth, G. Michalski, K. Kaźmierski, B. Maciejewska, B. Wiskirska-Woźnica, M. Żygis, W. Kuczko, and A. Sekuła, "Automatic english phoneme recognition from articulatory data generated by epg systems with grid and anatomical layout of contact sensors," 2019.