Multimodal Speech Recognition for Language-Guided Embodied Agents

Allen Chang[1,2], Xiaoyuan Zhu[1,2], Aarav Monga[1,2], Seoho Ahn[1,2],
Tejas Srinivasan[1], Jesse Thomason[1]

[1]Department of Computer Science, University of Southern California, USA
[2]Center for Artificial Intelligence in Society, University of Southern California, USA

`changall@usc.edu`

Abstract

Benchmarks for language-guided embodied agents typically assume text-based instructions, but deployed agents will encounter spoken instructions. While Automatic Speech Recognition (ASR) models can bridge the input gap, erroneous ASR transcripts can hurt the agents' ability to complete tasks. We propose training a multimodal ASR model that utilizes the accompanying visual context to reduce errors in spoken instruction transcripts. We train our model on a dataset of synthetic spoken instructions, derived from the ALFRED household task dataset, where we simulate acoustic noise by systematically masking spoken words. We find that utilizing visual observations facilitates masked word recovery, with multimodal ASR models recovering up to 30% more masked words than unimodal baselines. We also find that spoken instructions transcribed by multimodal ASR models result in higher task completion success rates for a language-guided embodied agent.
`github.com/Cylumn/embodied-multimodal-asr`

Index Terms: speech recognition, multimodal learning, human-robot interaction, embodied learning

1. Introduction

Several benchmarks aim to train simulation agents to complete household chores [1, 2, 3]. These agents are trained to receive and follow written instructions, neglecting the more realistic scenario of receiving spoken instructions. However, physical hardware agents and robots require the ability to understand spoken language for effective interaction with human users [4].

Automatic Speech Recognition (ASR) methods can transcribe spoken instructions into written text. However, erroneous ASR transcripts can hurt the embodied agents' ability to complete the required tasks. One proposed solution for improving ASR robustness is to use visual context, when available, to recover inaudible words. Previous work studied situations where spoken language described a single image [5]. We hypothesize that the visual observations made by embodied agents during task completion will be similarly helpful for disambiguating inaudible words. In particular, we test whether visual observations from an embodied agent can be utilized by multimodal ASR models to reason about and recover words that are masked in the audio signal. Consider transcribing *"Put the egg in the . . . "* where the destination word is too noisy to understand (Figure 1). A unimodal ASR model must use language priors alone to guess, for example, *"fridge."* By contrast, a multimodal ASR model can use the agent's visual observation to correctly reason that the user wants the egg to be put in the *"microwave."*

We apply the insight that visual scene information can inform noisy spoken language transcription in the context of an embodied, instruction-following agent. We synthesize a
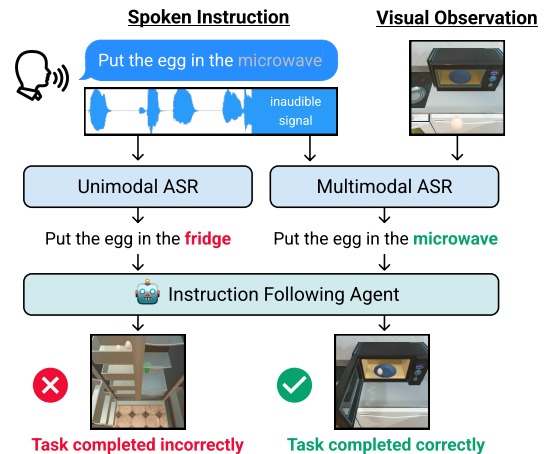

Figure 1: *Multimodal ASR systems can leverage an embodied agent's visual observation to better transcribe spoken instructions instead of relying on language priors alone.*

benchmark of text-to-speech (TTS) generated spoken instructions based on written instructions in the Action Learning From Realistic Environments and Directives (ALFRED) dataset [1]. To simulate acoustic noise, we systematically mask audio segments corresponding to words in the instructions. We train unimodal and multimodal ASR models on spoken instructions and evaluate their ability to recover masked words in seen versus unseen visual environments and heard versus unheard TTS speakers. We quantify the downstream impact of erroneous ASR transcriptions on an off-the-shelf ALFRED embodied agent.

We demonstrate that multimodal models can effectively use visual observations to transcribe noisy spoken instructions. We find that the additional visual context is useful in both seen and unseen household environments. Visual context also can help the ASR model generalize better to new speakers. We demonstrate that multimodal ASR models can mitigate the effect of noise in spoken instructions on a text-trained embodied agent's ability to complete tasks. These findings are promising for building embodied agents that follow spoken instructions.

2. Background

This work addresses a gap in the embodied task completion literature. Most embodied agent training assumes clean text input, but human speech will be encountered in deployed settings. We leverage insights from existing ASR literature to develop a multimodal ASR model for noisy spoken instructions.

Embodied Task Completion. Existing benchmarks aim to train embodied agents to complete tasks by following instructions [6]. These benchmarks range from vision-language navigation [7, 8, 9, 10, 11], where an agent must follow the user's instructions by executing actions to reach the desired destination, to embodied task completion [1, 2, 12, 3], where agents interact with objects in the environment to complete the user's instructions. These benchmarks provide language as written text, and while a small number of benchmarks involve spoken instructions [13], almost all modeling attempts assume text-based instructions. There are some works at the intersection of embodied learning and acoustic signals [14, 15], but these do not involve spoken instructions. We explore multimodal ASR modeling for language-guided embodied agents.

Multimodal Speech Recognition. Previous works have explored augmenting speech recognition with visual information such as lip readings [16, 17], visual scenes [18, 19, 20], and task semantics [21]. When the audio signal is clear, the visual modality has been shown to regularize the model [22] rather than assisting with semantic disambiguation. In contrast, visual semantics have been shown to be helpful when the audio signal is degraded [5, 23]. When visually salient words are masked in the audio signal, multimodal ASR can utilize the visual input to recover the masked words [24]. We apply these findings to an embodied setting, where agents receiving instructions with noisy or degraded speech can leverage their visual observations to recover the instruction text and complete the requested task.

3. Methodology

We adapt the ALFRED [1] language-guided instruction benchmark to create synthesized speech commands, apply noising policies to that speech, and attempt to recover ground truth instructions using a novel, multimodal ASR model.

3.1. Preliminaries: ALFRED Instruction Following Task

ALFRED [1] is a benchmark in which an embodied agent must follow language instructions to complete tasks by navigating a room and interacting with objects. Tasks consist of a language goal $\mathcal{G}$ (e.g., *"Find an egg in the fridge and microwave it."*), a sequence of K sub-goal instructions $\mathcal{I}_{1...K}$ to achieve that goal, and target environment state conditions (e.g., an egg has been heated). Training and validation tasks include an annotated sequence of actions to accomplish the goal. At each timestep, the agent receives a single-frame visual observation $v \in \mathbb{R}^{W \times H \times 3}$ and executes a discrete navigation step (e.g., `RotateRight`) or object interaction (e.g., `PickUp(Egg)`) which updates the environment state. The task is successfully completed when the environment state satisfies the target state conditions.

3.2. Building a Dataset of Spoken Instructions

We create a dataset of synthetic spoken instructions because of the lack of embodied learning datasets with speech inputs.

We extract sub-goal text instructions $\mathcal{I}$ from the ALFRED dataset and apply off-the-shelf TTS models[1] to generate synthetic speech instructions $S(\mathcal{I})$. We pair each instruction with a visual context v, the agent's observation when the previous sub-goal is completed. For example, if the agent has navigated to a microwave as the previous subgoal, the observation accompanying *"Put the egg in the microwave"* will be of a kitchen counter with a microwave (Figure 1). Multimodal ASR models

---

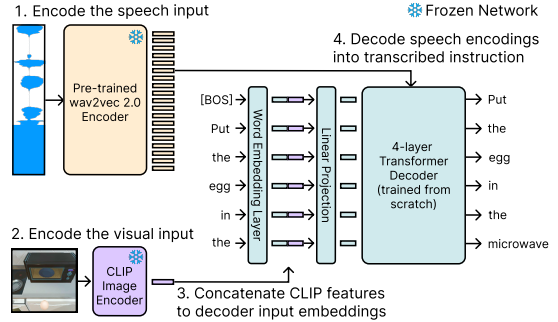[1]TTS models are sourced from the open-source Silero [25] library.

Figure 2: *Our multimodal ASR model architecture. For the unimodal ASR model, the visual features are not concatenated to the word embeddings and there is no linear projection layer.*

trained on our data are tasked with transcribing the speech input $S(\mathcal{I})$ into the word sequence $\mathcal{I}$ using the paired visual context.

Since the test examples in the ALFRED dataset are not paired with annotated visual observations, we use the ALFRED validation data as our test data. Our test set is partitioned into environments that were *seen* and *unseen* during training, enabling us to study the effects of in- versus out-of-distribution visual contexts on multimodal ASR. We partition the ALFRED training data in a 90:10 split to generate our train and validation sets. Training instructions with exact text matches in our validation or test sets are removed. In total, there are $41,474$ spoken instructions in the training set, $5,222$ in validation, $5,570$ in *seen* test environments, and $5,140$ in *unseen* test environments.

To examine the effects of multimodal ASR when generalizing across speakers, we create two sets of spoken instructions: (1) we generate spoken instructions for training, validation, and test folds using a single American English speaker TTS model, denoted as $S_A(\mathcal{I})$; (2) we use 10 Indic English speaker TTS models for training and validation, the same 10 speakers for a *heard* ASR test instruction split, $S_I^{\text{heard}}(\mathcal{I})$, and 5 different Indic English speakers for an *unheard* ASR test instruction split, $S_I^{\text{unheard}}(\mathcal{I})$. In all cases, TTS outputs for all sub-goal instructions within a single task are generated from the same speaker.

3.3. Injecting Noise into Spoken Instructions

Prior studies in machine translation [26] and ASR [5] have shown that visual context is helpful when the primary modality is degraded. Building on these insights, we corrupt the audio in our dataset by masking audio segments in the instructions.

Following prior work [5], we apply masking at the word-level. After segmenting instructions with wav2vec 2.0 [27] to identify word boundaries, we mask words by substituting their speech segments with Gaussian noise. In our controlled setting, we can mask out different sets of words to evaluate the utility of the visual context under different conditions.

3.4. Modeling

We train unimodal and multimodal ASR models and evaluate their ability to transcribe noised speech. The unimodal ASR model consists of a speech encoder and a language decoder. The speech encoder is a frozen wav2vec 2.0 [27], pre-trained on Librispeech [28], which encodes the spoken instruction $S(\mathcal{I})$ into a sequence of speech encodings. The decoder is a 4-layer Transformer trained from scratch, which jointly attends over the speech encodings and decoded words.

Table 1: *WER (%) and RR (%) for ASR models trained on the American English TTS speaker across masking policies.*

Test Set	ASR Model	No Mask	Only Nouns			All Words	
			20%	40%	100%	20%	40%
Word Error Rate ↓							
Seen	Unimodal	12.6	20.0	26.4	34.0	35.2	49.9
Seen	Multimodal	**11.9**	**19.9**	**24.5**	**30.4**	**29.3**	**46.4**
Unseen	Unimodal	13.8	**21.4**	27.6	34.7	36.3	51.7
Unseen	Multimodal	**12.6**	21.5	**26.2**	**33.1**	**32.3**	**50.2**
Recovery Rate ↑							
Seen	Unimodal	–	61.1	56.1	48.0	51.4	38.8
Seen	Multimodal	–	**64.3**	**60.5**	**56.8**	**57.7**	**45.9**
Unseen	Unimodal	–	60.6	55.9	48.4	52.0	37.5
Unseen	Multimodal	–	**62.2**	**57.5**	**52.0**	**53.4**	**40.4**

To decouple the role of the visual modality from model architecture in ASR performance, our multimodal ASR model has a near-identical architecture (Figure 2) to the unimodal ASR model. The visual context v is encoded using a frozen CLIP-ViT image encoder [29] into a feature vector. At every timestep of generation, the CLIP feature vector is concatenated to the decoder's input word embedding, linearly projected back to the original embedding dimension, and passed through the Transformer for generating the next word.

4. Experiments

We train unimodal and multimodal ASR models on our spoken instructions dataset (Section 4.1) across several audio masking policies (Section 4.2). We evaluate the added benefit of the visual modality when the audio signal is noised (Section 4.3).

4.1. Training and Inference Implementation Details

Across all versions of our spoken instructions dataset, we train unimodal and multimodal ASR models with cross-entropy loss. Models are trained for 50 epochs using the Adam optimizer with a 10^{-4} learning rate. During inference, transcriptions are generated using beam search with a beam width of 5.

4.2. Audio Masking Policies

To evaluate the benefit of the visual modality with respect to the degree of acoustic noise, we conduct experiments across six noise policies for word masking. Three masking policies consist of masking only nouns, since these words are the most visually salient. We identify the 100 most frequent nouns in our training set using the Natural Language Toolkit [30] and create three versions of our dataset where 20%, 40% and 100% of these nouns in each spoken instruction $S(\mathcal{I})$ are masked. Following [31], we use two noise policies that mask words at random, where 20% and 40% of words in each $S(\mathcal{I})$ are masked. Finally, a trivial masking policy is to use the original, un-noised audio. We apply each policy at both training and inference time.

4.3. Evaluation Metrics

We evaluate ASR models on their Word Error Rate (WER), as well as their Recovery Rate (RR), which measures the percentage of correctly transcribed noised words in the dataset [5]. Further, to quantify the added benefit from the visual modality, we introduce two new metrics.
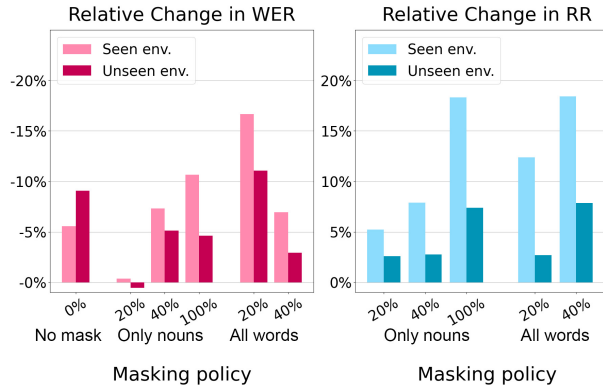

Figure 3: *Relative Change in WER (negative is better) and Relative Change in RR (positive is better) between the unimodal and multimodal ASR models for the American English TTS speaker across masking policies between seen and unseen environments.*

Relative Change in WER (Δ_{WER}): The relative change in Word Error Rate for the *M*ultimodal model over *U*nimodal.

$$\Delta_{WER} = \frac{WER_M - WER_U}{WER_U} \times 100\%.$$

Relative Change in RR (Δ_{RR}): The relative change in Recovery Rate for the *M*ultimodal model over *U*nimodal.

$$\Delta_{RR} = \frac{RR_M - RR_U}{RR_U} \times 100\%.$$

5. Results and Discussion

5.1. Multimodal ASR Improves Masked Word Recovery in both Seen and Unseen Environments

We test whether incorporating visual information improves ASR for spoken instructions $S_A(\mathcal{I})$ using a single, American English TTS model in *seen* versus *unseen* visual environments. While both unimodal and multimodal ASR models suffer from higher WER and lower RR as the level of audio masking increases, multimodal ASR models consistently outperform their unimodal counterparts across all masking policies (Table 1).

We further observe that the utility of multimodal ASR is directly proportional to the level of audio degradation, evidenced by increases in Δ_{WER} and Δ_{RR} as the proportion of masked words increases (Figure 3). This trend suggests that visual observations are more advantageous as speech signals become more degraded, and multimodal ASR may be more beneficial to agents when performing tasks in noisier environments. The multimodal models' WER and RR improvements generalize to unseen environments, though the improvements are less pronounced than in seen environments. These results demonstrate the viability of this approach for embodied learning, where training and evaluation environments are often different.

5.2. Multimodal ASR Generalizes Better to New Speakers

Next, we test whether incorporating multimodal visual information improves ASR for spoken instructions of multiple speakers when those speakers are heard $S_I^{heard}(\mathcal{I})$ or unheard $S_I^{unheard}(\mathcal{I})$ during training. Multimodal ASR is more helpful when evaluated on 5 unheard Indic English speakers, compared to the 10 Indic English speakers present in the training data (Figure 4).
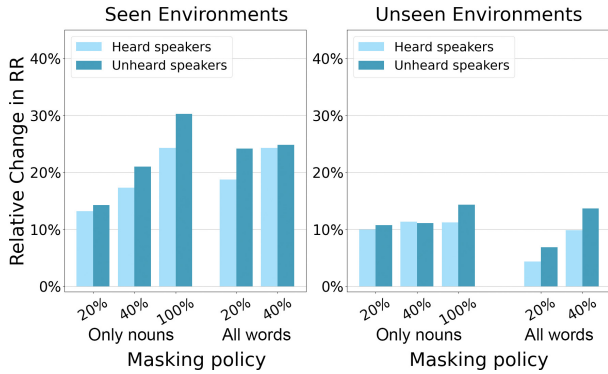
Figure 4: Δ_{RR} *between the unimodal and multimodal models across masking policies in both seen and unseen environments, evaluated on heard and unheard Indic English speakers.*

Table 2: Δ_{RR} *on the subset of words corresponding to nouns and non-nouns from the random all-words masking policies.*

Speaker(s)	POS	20% Masking		40% Masking	
		Seen	*Unseen*	*Seen*	*Unseen*
American	Noun	+12.7%	+03.0%	+31.0%	+14.6%
American	Other	+04.4%	−02.2%	−00.1%	−01.5%
Indic (Heard)	Noun	+20.1%	+05.5%	+40.1%	+18.8%
Indic (Heard)	Other	−04.3%	−11.9%	+02.5%	−01.8%
Indic (Unheard)	Noun	+25.7%	+07.9%	+45.5%	+24.7%
Indic (Unheard)	Other	−00.7%	−08.1%	−01.6%	−00.2%

These relationships indicate that visual signals additionally regularize model training to adapt to audio from unheard speakers. Importantly, this relationship is observed in the joint case of *unseen environments and unheard speakers*, as will be the case for newly deployed embodied agents in the world.

5.3. Multimodal ASR is Helpful For Visually Salient Words

We investigate whether multimodal ASR is helpful for the right reasons by evaluating whether it recovers masked words that are more visually observable. For the masking policies that mask words at random, we evaluate Δ_{RR} on the subsets of the masked words corresponding to nouns and non-nouns. In Table 2, we observe that masked nouns are much more likely to be recovered than other masked words by multimodal ASR. Δ_{RR} between nouns and non-nouns are most different in the 40% random all-words masking policy for seen environments, suggesting that multimodal ASR is most helpful when audio is heavily perturbed but visual observations are still familiar. The inverse is true with 20% random all-words masking in unseen environments, which has the lowest Δ_{RR}. These trends reveal that multimodal models are effective when masked words have strong visual salience, particularly in familiar environments.

5.4. Multimodal ASR Helps Agents Complete Tasks Better

We now investigate the impact of ASR transcription quality on an embodied agent's ability to complete tasks and whether leveraging visual observations for ASR mitigates the agent's performance degradation in the presence of noisy speech. We perform our analysis on the Episodic Transformer (E.T.) [32], an off-the-shelf ALFRED agent that is trained to receive a goal text $\mathcal{G}$ and K sub-goal instruction texts $\mathcal{I}_{1...K}$.
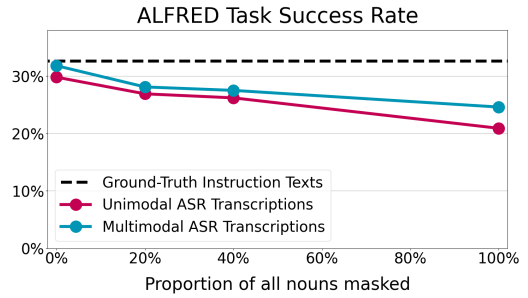

Figure 5: *Text-trained ALFRED agents that observe instructions transcribed by multimodal ASR achieve higher downstream task success than when using unimodal ASR.*

We begin by extracting transcripts for the spoken instruction $S(\mathcal{I}_i)$ corresponding to every sub-goal instruction $\mathcal{I}_i$ for each episode. Transcripts are extracted for the American English speaker from unimodal and multimodal ASR models, across four different masking policies (no masking and 20%, 40%, and 100% noun masking). The ground-truth goal text $\mathcal{G}$ and the transcribed sub-goal instructions are passed to E.T., and the agent attempts to complete the task using the transcribed instructions. We evaluate the agent's Task Success Rate on the ALFRED validation set, in the seen environments.

We observe that transcribed instructions from both unimodal and multimodal ASR models lead to lower model performance than the ground-truth instructions (Figure 5). However, the multimodal ASR models' transcriptions lead to lower performance degradation. We further observe that as the level of audio masking increases, the performance gap between the unimodal and multimodal ASR increases. These findings demonstrate that not only does multimodal ASR lead to more accurate transcription of spoken instructions, it also results in better downstream task completion for the embodied agent.

6. Conclusions

In this work, we address the challenge of embodied task completion by following spoken instructions. We demonstrate that embodied agents can use their visual observations to improve ASR when transcribing spoken instructions. ASR models using visual observations achieve better WER and RR across home environments, speakers of varied demographics, and levels of audio degradation. These models also show higher improvement when spoken instructions become noisier and when transcribing speech from unheard speakers, improving the robustness of ASR. Finally, we demonstrate that multimodal ASR can improve a pre-trained embodied agent's ability to complete tasks successfully from spoken instructions. These findings motivate the use of visual observations in the implementation of ASR for language-guided embodied agents.

Our work presents a proof-of-concept for spoken instruction following by creating a synthetic speech dataset and systematically masking words in the audio. Future work should investigate how our findings generalize to real human speakers and more realistic audio degradation. Further, while we use only a single visual observation, future work should explore utilizing 3D scene representations built by embodied agents [33, 34], as well as other strategies to actively search the visual environment, to resolve ambiguities in ASR.

7. References

[1] M. Shridhar, J. Thomason, D. Gordon, Y. Bisk, W. Han, R. Mottaghi, L. Zettlemoyer, and D. Fox, "ALFRED: A benchmark for interpreting grounded instructions for everyday tasks," in *Computer Vision and Pattern Recognition (CVPR)*, 2020.

[2] X. Puig, K. Ra, M. Boben, J. Li, T. Wang, S. Fidler, and A. Torralba, "Virtualhome: Simulating household activities via programs," in *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[3] A. Padmakumar, J. Thomason, A. Shrivastava, P. Lange, A. Narayan-Chen, S. Gella, R. Piramuthu, G. Tur, and D. Hakkani-Tur, "TEACh: Task-driven embodied agents that chat," in *Conference on Artificial Intelligence (AAAI)*, 2022.

[4] M. Marge, C. Espy-Wilson, N. G. Ward, A. Alwan, Y. Artzi, M. Bansal, G. Blankenship, J. Chai, H. Daumé III, D. Dey *et al.*, "Spoken language interaction with robots: Recommendations for future research," *Computer Speech & Language*, 2022.

[5] T. Srinivasan, R. Sanabria, and F. Metze, "Looking enhances listening: Recovering missing speech using images," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[6] J. Gu, E. Stefani, Q. Wu, J. Thomason, and X. Wang, "Vision-and-language navigation: A survey of tasks, methods, and future directions," in *Association for Computational Linguistics (ACL)*, 2022.

[7] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, and A. Van Den Hengel, "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments," in *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[8] Y. Qi, Q. Wu, P. Anderson, X. Wang, W. Y. Wang, C. Shen, and A. v. d. Hengel, "REVERIE: Remote embodied visual referring expression in real indoor environments," in *Computer Vision and Pattern Recognition (CVPR)*, 2020.

[9] H. Chen, A. Suhr, D. Misra, N. Snavely, and Y. Artzi, "Touchdown: Natural language navigation and spatial reasoning in visual street environments," in *CVPR*, 2019.

[10] J. Thomason, M. Murray, M. Cakmak, and L. Zettlemoyer, "Vision-and-dialog navigation," in *Conference on Robot Learning (CoRL)*, 2019.

[11] S. Banerjee, J. Thomason, and J. Corso, "The RobotSlang benchmark: Dialog-guided robot localization and navigation," in *Conference on Robot Learning (CoRL)*, 2021.

[12] A. Suhr, C. Yan, J. Schluger, S. Yu, H. Khader, M. Mouallem, I. Zhang, and Y. Artzi, "Executing instructions in situated collaborative interactions," in *Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

[13] A. Ku, P. Anderson, R. Patel, E. Ie, and J. Baldridge, "Room-Across-Room: Multilingual vision-and-language navigation with dense spatiotemporal grounding," in *Empirical Methods for Natural Language Processing (EMNLP)*, 2020.

[14] C. Chen, U. Jain, C. Schissler, S. V. A. Gari, Z. Al-Halah, V. K. Ithapu, P. Robinson, and K. Grauman, "Soundspaces: Audio-visual navigation in 3d environments," in *European Conference on Computer Vision (ECCV)*, 2020.

[15] S. Paul, A. Roy-Chowdhury, and A. Cherian, "AVLEN: Audio-visual-language embodied navigation in 3d environments," in *Neural Information Processing Systems (NeurIPS)*, 2022.

[16] C. Bregler, H. Hild, S. Manke, and A. Waibel, "Improving connected letter recognition by lipreading," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1993.

[17] P. Duchnowski, U. Meier, and A. Waibel, "See me, hear me: integrating automatic speech recognition and lip-reading," in *International Conference on Spoken Language Processing (ICSLP)*, 1994.

[18] Y. Miao and F. Metze, "Open-domain audio-visual speech recognition: A deep learning approach." in *INTERSPEECH*, 2016.

[19] S. Palaskar, R. Sanabria, and F. Metze, "End-to-end multimodal speech recognition," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[20] R. Sanabria, O. Caglayan, S. Palaskar, D. Elliott, L. Barrault, L. Specia, and F. Metze, "How2: A large-scale dataset for multimodal language understanding," in *Visually Grounded Interaction and Language (ViGIL) Workshop, Neural Information in Processing Systems (NeurIPS)*, 2018.

[21] R. Corona, J. Thomason, and R. J. Mooney, "Improving black-box speech recognition using semantic parsing," in *International Joint Conference on Natural Language Processing (IJCNLP)*, 2017.

[22] O. Caglayan, R. Sanabria, S. Palaskar, L. Barraul, and F. Metze, "Multimodal grounding for sequence-to-sequence speech recognition," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[23] T. Srinivasan, R. Sanabria, and F. Metze, "Analyzing utility of visual context in multimodal speech recognition under noisy conditions," in *The How2 Workshop at the International Conference on Machine Learning (ICML)*, 2019.

[24] T. Srinivasan, R. Sanabria, F. Metze, and D. Elliott, "Fine-grained grounding for multimodal speech recognition," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020.

[25] S. Team, "Silero models: pre-trained enterprise-grade stt/tts models and benchmarks," *GitHub*, 2021.

[26] O. Caglayan, P. S. Madhyastha, L. Specia, and L. Barrault, "Probing the need for visual context in multimodal machine translation," in *North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019.

[27] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Neural Information Processing Systems (NeurIPS)*, 2020.

[28] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[29] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning (ICML)*, 2021.

[30] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, 2009.

[31] T. Srinivasan, R. Sanabria, F. Metze, and D. Elliott, "Multimodal speech recognition with unstructured audio masking," in *NLP Beyond Text Workshop, Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[32] A. Pashevich, C. Schmid, and C. Sun, "Episodic Transformer for Vision-and-Language Navigation," in *International Conference on Computer Vision (ICCV)*, 2021.

[33] S. Y. Min, D. S. Chaplot, P. Ravikumar, Y. Bisk, and R. Salakhutdinov, "FILM: Following Instructions in Language with Modular Methods," in *International Conference on Learning Representations (ICLR)*, 2022.

[34] V. Blukis, C. Paxton, D. Fox, A. Garg, and Y. Artzi, "A persistent spatial semantic representation for high-level natural language instruction execution," in *Conference on Robot Learning (CoRL)*, 2021.