ABSTRACT

Title of dissertation: SPEAKER-BASED VARIABILITY IN ROBOTIC SPOKEN LANGUAGE GROUNDING

Gaoussou Youssouf Kebe, Master of Science, 2022

Thesis directed by: Dr. Cynthia Matuszek
Department of Computer Science

Robots in human spaces need to be able to understand human-provided natural language instructions in the context of their physical environment. Learning to understand grounded language, which connects natural language to percepts, is a critical research area. However, the majority of existing efforts relies on highly curated text and ignores the noise and variance present in end-user speech. Existing speech-based grounded language learning works require an extensive amount of speech data. Additionally, variation in speech characteristics can cause challenges for grounding models, and prior works do not investigate the difference in performance between demographic groups. In this thesis, I train and evaluate language grounding models on collected spoken and textual descriptions of common household objects. I leverage recent work in self-supervised speech representation models to learn groundings without the interference of transcriptions as an intermediate representation. The goal is to eliminate the effects of off-the-shelf speech-to-text models as a potential source of bias. The experimental results suggest that this approach can make language grounding systems more inclusive towards accented speakers and increase general performance.

Curriculum Vitae

Name:	Gaoussou Youssouf Kebe
Degree:	Master of Science, 2022
Major:	Computer Science
Collegiate institutions attended:	University of Maryland, Baltimore County (UMBC), MS in Computer Science, 2022 Bursa Technical University (BTU), BS in Computer Engineering, 2018
Professional positions held:	Research Intern, Honda Research Institute, May 2022 - August 2022 Graduate Research Assistant, UMBC, Jan. 2020 - May 2022 Graduate Teaching Assistant, UMBC, Aug. 2019 - Dec. 2019 Research Assistant, UMBC, June 2018 - June 2019
Professional Publications:	*Bridging the Gap: Using Deep Acoustic Representations to Learn Grounded Language from Percepts and & Raw Speech.* **Gaoussou Youssouf Kebe**, Luke E. Richards, Edward Raff, Francis Ferraro, Cynthia Matuszek, **Accepted** at the 2022 AAAI Conference on Artificial Intelligence (**AAAI 2022**), Vancouver, BC, Canada, February 2022 *A Spoken Language Dataset of Descriptions for Speech-Based Grounded Language Learning.* **Gaoussou Youssouf Kebe**, Padraig Higgins, Patrick Jenkins, Kasra Darvish, Rishabh Sachdeva, Ryan Barron, John Winder, Donald Engel, Edward Raff, Francis Ferraro, Cynthia Matuszek, In the 35th Conference on Neural Information Processing Systems (**NeurIPS 2021**) Datasets and Benchmarks Track, December 2021

Practical Cross-modal Manifold Alignment for Robotic Grounded Language Learning. Andre T Nguyen, Luke Richards, **Gaoussou Youssouf Kebe**, Edward Raff, Kasra Darvish, Francis Ferraro, Cynthia Matuszek, In the 4th Multimodal Learning and Applications Workshop (MULA) at the Conference on Computer Vision and Pattern Recognition (**CVPR 2021**), June 2021

SPEECH-BASED GROUNDED LANGUAGE LEARNING WITH SPEAKER VARIABILITY

by

Gaoussou Youssouf Kebe

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, Baltimore County in partial fulfillment
of the requirements for the degree of
Master of Science
2022

Advisory Committee:
Dr. Cynthia Matuszek, Chair/Advisor
Dr. Francis Ferraro,
Dr. Edward Raff

© Copyright by
Gaoussou Youssouf Kebe
2022

Acknowledgments

I cannot fully express how lucky and appreciative I am to even be in the position to write this thesis. I owe my gratitude to all the people who have facilitated my academic journey and pursuit of graduate studies. This thesis would not have been possible without their invaluable support and guidance.

First and foremost I would like to thank my advisor, Dr. Cynthia Matuszek for being an ever-present source of compassion, encouragement, and advice throughout my time at UMBC. She has consistently been a praising and celebratory voice at the highest moments and an indispensable guide during the lowest ones. Beyond her role as a research mentor, she has played that of a therapist, life coach and career coach. I cannot even imagine how different my graduate school experience would have been without Dr. Matuszek's support.

I want to extend my sincere thanks to my committee member and (shadow advisor), Dr. Francis Ferraro for his hands-on mentorship and in-depth technical explanations on all the projects I was fortunate enough to work on under his guidance. I would also like to thank my committee member, Dr. Edward Raff for always finding the time in his busy schedule to answer my questions and provide me with valuable insights on all career and research-related matters. Thank you to all three of my committee members for their patience, faith and tolerance for my mistakes.

It has been an absolute pleasure to be a member of the IRAL lab. Thank you to all of my labmates for making the lab a safe and comfortable work environment. My thanks and appreciation to John Winder and Nisha Pillai for their support in my early days as a lab member; Patrick Jenkins for his hard work without which this thesis would not have been possible; Kasra Darvish for being a great partner during all the brainstroming and ranting sessions; Nadezhda Bzhilyanskaya for helping me face many of the walls in my last semester at UMBC –both metaphorical and virtual; to all of my co-authors and close collaborators (Luke Richards, Padraig Higgins, Andre Nguyen, Rishabh Sachdeva, Ryan Barron, Adam Berlier, Frank Serna, Don Engel, Aidan Newell) for their invaluable contributions.

I owe a debt of gratitude to too many at UMBC (Dr. Mohamed Younis, Lauren Mirzakhalili, etc...) and outside of it (Dr. Izzet Fatih Senturk, Dr. Orhan Ates, etc...) to list here. Thank you all!

Finally, my deepest thanks to my family for their unconditional support. To my mother, Aoua Sidibe, and my brother, Cheickna Youssouf Kebe, for being the rocks that I could always depend on; and to my aunts and uncles who have stepped up for me every time I needed them.

Lastly, thank God for this wonderful opportunity and experience.

Table of Contents

List of Figures

List of Abbreviations

AAAI	Association for the Advancement of Artificial Intelligence
ALFRED	Action Learning From Realistic Environments and Directive
AMT	Amazon Mechanical Turk
API	Application Programming Interface
ASR	Automatic Speech Recognition
AUC	Area Under the (ROC) Curve
BLEU	Bilingual Evaluation Understudy
BERT	Bidirectional Encoder Representations from Transformers
CLEVR	Compositional Language and Elementary Visual Reasoning diagnostics dataset
COCO	Common Objects in Context
CTC	Connectionist Temporal Classification
DeCoAR	Deep Contextualized Acoustic Representations
ELMo	Embeddings from Language Models
GoLD	Grounded Language Dataset
HRI	Human Robot Interaction
IRAL	Interactive Robotics and Language Lab
LSTM	Long Short-Term Memory
MFCC	Mel-Frequency Cepstral Coefficients
MRR	Mean Reciprocal Rank
NeurIPS	the conference on Neural Information Processing Systems
NLP	Natural Language Processing
PCC	Pearson Correlation Coefficient
RGB	Red, Green, and Blue
RGB-D	Red, Green, and Blue plus Depth
ROC	Receiver Operating Characteristic
STT	Speech To Text
WER	Word Error Rate

Chapter 1

Introduction

Learning to understand grounded language—learning the semantics of language that occurs in the context of, and refers to, the broader world—is a rich area of work that has engaged researchers from robotics [129], natural language processing [77], vision [30], and cognitive science [114], among others. In robotics, grounded language refers primarily to grounding human utterances in the perceived physical world of objects, actions, and the environment. Learning from grounded language is an intuitive choice for interacting with agents in a physical environment.

Current approaches to grounded language learning require data in both the perceptual ("grounded") and linguistic domains. Although existing datasets have been used for this purpose [58, 68, 28, 94, 131], the language component is almost always derived from textual input or manually transcribed speech [86, 128].

While language learning offers a clearly defined way for embodied agents to learn about changing environments and goals directly from a specific end user, with some exceptions, the majority of current work in this area still operates primarily on textual data. This approach significantly limits our ability to deploy agents in realistic human environments, where spoken inputs can be expected. Existing work on using speech directly typically relies on off-the-shelf speech-to-text systems. These systems are rarely developed in tandem with the robotics community, and so do not

take the unique challenges of robotic sensing into account [82]. In addition, current ASR systems work inconsistently across demographics [127, 50], which represents a problem in inclusive design. They are also "black box" systems that cannot improve their speech recognition from their environment or other perceptual clues. Since the grounding system does not have access to the information used by these models, it can only rely on their sometimes erroneous output.

1.1 Contributions

In this thesis, I extended the existing **Gro**unded **L**anguage **D**ataset (GoLD) [56], which contains images of common household objects and their description in multiple formats: text, speech (audio), and speech transcriptions (see fig. 3.1). Additionally, I bridge the gap between learning grounded language about the perceived world via text-based language, and directly learning to recognize speech without access to the physical context in which it occurs. I contribute a detailed analysis of natural language grounding from raw speech to robotic sensor data of everyday objects using state-of-the-art speech representation models. I then conduct an analysis of audio and speech qualities of individual participants, in which we demonstrate that learning directly from raw speech mitigates the performance difference between linguistic groups on a well-known grounded language learning problem.

The primary contributions of this thesis are as follows:

1. I extended a publicly available, multimodal, multi-labelled dataset of household objects, with image+depth data and textual and spoken descriptions with

automated transcription.

2. I demonstrate the feasibility of acquiring grounded language directly from end-user speech using a relatively small number of data points, without relying on intermediate textual representations.

3. I show that such learning improves performance on users with accented speech as compared to relying on automatic transcriptions.

1.2 Thesis Organization

The remainder of this thesis is organized as follows. Chapter 2 provides an overview of the field of grounded language and multimodal learning. It particularly describes related grounded language learning datasets and approaches and situates the contributions of this thesis within the current state of the literature. Chapter 3 describes the GoLD dataset, including the data collection process for all modalities, the speaker annotation process, and a qualititative evaluation of the data [60]. Chapter 4 describes the multimodal learning approach used for both spoken and textual language, and the featurization techniques used for the different modalities [61]. Chapter 5 describes the object selection task we use to determine whether language groundings have been learned successfully, compares the experimental results of learning from percepts and raw speech directly, vs. the traditional transcription-first approach, and provides an analysis of both approaches when learning from spoken language from different demographic groups present in the dataset.

Chapter 2

Related Work

This thesis falls into the general area of grounding — an interdisciplinary topic that has been the subject of extensive research across different communities, including Robotics [129, 26, 83], Natural Language Processing [16, 17, 8, 12], Computer Vision [105, 79, 30], Cognitive Science [15, 24, 125, 139].

Particularly, this thesis focuses on the grounded language learning task — semantically learning natural language by relating words and utterances to real world entities. On a robot, the grounded language learning task has a number of applications, including Vision & Language Task Completion — where a robotic agent translates natural language instructions into interaction actions [128, 10, 20, 120, 138, 119, 90, 104, 147], Vision & Language Navigation — where an agent follows a route to a target location or series of locations based on natural language descriptions [84, 65, 18, 118, 144, 38, 4, 136, 150], Human-Robot Dialog — where an embodied agent engages with a human in an interactive conversation [122, 66, 130, 133, 132, 32] and Robotic Object Retrieval — where a robot selects a target object or series of object among a set of candidates based on a descriptive natural language utterance [85, 86, 52, 2, 45, 101, 62, 103, 95, 109, 93].

The latter of these is used as the application area of this thesis. Retrieving objects based on human-provided descriptions [95] is a necessary component of care-

taking and domestic robots. The work described in this thesis has been published in two conference papers [60, 61].

This chapter goes over a subset of early and recent grounded language learning efforts and situates the work presented in this thesis within the broader context of the field.

2.1 Grounded Language Datasets

This section discusses a selection of text-based and spoken datasets for grounded language learning.

2.1.1 Text-based Datasets

Grounded language learning is a multimodal problem and, as such, requires datasets that combine language utterances with information from the visual modality. Datasets that were created for similarly situated problems such as image captioning [108, 75, 19, 143] and Visual Question Answering [80, 5, 58] can be used to train grounded language learning models. However, these datasets are handcrafted for their specific research problem and may lack some perceptual and language properties that are more relevant to the grounded language learning task.

For example, CLEVR [58] was designed as a benchmark for question answering tasks. Objects in CLEVR are limited to a small set of attributes which in turn limits the types of questions in both syntax and content. In comparison, GoLD contains more complex real-world objects and does not limit the scope of the annotations to

a fixed set of characteristics.

Similarly, the perceptual data in the image captioning datasets such as Flickr8k [108], COCO [75], and STAIR [143] is limited to flat RGB images, making them less suitable for grounding language concepts that require 3D sensors such as object shapes and distance. In addition to RGB images, GoLD includes depth images and point-clouds.

The perceptual data in GoLD is heavily influenced by the University of Washington RGB-D dataset [71]. Both datasets contain large numbers of everyday objects from multiple angles. GoLD is collected with a now state-of-the-art sensor which enables us to capture smaller objects at a finer level of detail (such as an Allen key, the diameter of which pushes the limits of the depth sensor when laid on the table). Furthermore, objects were selected based on their potential utility for specific human-robot interaction scenarios, such as things a person might find in a medicine cabinet or first aid kit, enabling learning research relevant to eldercare and emergency situations [11].

2.1.2 Speech-based Datasets

A number of speech-based datasets are derived from analogous text-based image captioning datasets [22]. This is often achieved by synthetically creating spoken captions. For example, Synthetically Spoken COCO [23] and Synthetically Spoken STAIR [47] generate their speech by feeding COCO [75] and STAIR [143] captions respectively into Google's Text-to-Speech system. A single synthetic voice

is used for all spoken captions, making both datasets lack speaker variability in their speech data.

SPEECH-COCO [46] improves upon these by using eight different synthetic voices, but still lacks the noise and variability present in natural speech. Flickr Audio Captions [41] and SpokenCOCO [51] solve this by having crowd workers read aloud the captions in the Flickr8K [108] and COCO datasets. This increases the naturalness of their speech data and incorporates the challenges of speech recognition in real world situations (background noise, accent variability, microphone quality, etc.). However, spontaneous speech differs perceptually and syntactically from scripted speech [70] and is more challenging for speech recognition models [39].

Places Audio Captions [42, 43], which uses the MIT Places 205 Database [149], is the only other major dataset in this area where the speech is collected directly from the spoken descriptions of crowd workers; however, the descriptions are of all the salient objects in an image instead of a single object.

All these datasets also only contain color images while GoLD extends this to include depth images and pointclouds.

Creating a dataset that includes speech has a high cost of collecting and transcribing audio. [111] presents a grounded language system that can generate descriptions of targets within a scene of colored rectangles. Visual data for this task is easily generated, but speech descriptions were recorded and transcribed over a long period of time. The manual audio transcription task can take four to ten hours per hour of audio [35, 140]. Such perfectly transcribed audio is also unrealistic for real-world usage, which must rely on automation. GoLD includes ASR-produced

transcriptions along with the raw audio. In addition, a subset of the ASR-produced transcriptions is evaluated for their quality.

2.2 Grounded Language Models

This section discusses a subset of existing multimodal learning approaches applied to the grounded language learning problem.

2.2.1 Text-based Models

A number of grounded language models, particularly in the context of robotic object retrieval, relied on using predefined natural language tokens such as "red", "round" as labels for visual attribute classifiers [14, 85, 102, 101, 62, 109]. While this approach was shown to perform well on tasks with a limited set of object attributes, the annotation and computational costs render it hardly scalable.

With the advance of deep learning, recent grounding models move away from these handcrafted features by learning more abstract semantic representations for both the language and vision modalities. Chandu et al. [16] identified three major approaches to learning these cross-modal representations: fusion, alignment, and projecting into a common space.

Multimodal fusion is generally used in prediction tasks that require information from both modalities, such as predicting individual low-level actions in vision & language navigation and task completion [20, 120, 104]. Multimodal alignment consists of synchronizing related visual and natural language sequences such as movie

scenes and screenplay [27] or cooking videos and recipes [81]. The last approach, and the one adopted in this thesis, is to project language and vision representations into a common space [121, 59, 25, 95, 93].

The grounded language learning approach used is in the thesis is the manifold alignment approach of Nguyen et al. [93], in which language and vision representations are projected into a shared manifold, which is used to retrieve relevant objects given a natural language description. The novelty of the work is not in the triplet loss learning method for multi-modal alignment but in the comparison of transcription-based versus raw speech methods, and analysis of performance for end-users.

2.2.2 Speech-based Models

Tellex et al. [129] surveyed the many machine learning methods used, possible applications, and the human-robotic interaction implications of grounded language learning on a robotic platform. Although some of these approaches use text derived from ASR, none use speech directly. The focus of this paper is on learning language groundings directly from speech. While the majority of existing grounded language learning is performed over text sources (either typed or transcribed), there are exceptions that demonstrate the importance of learning directly from speech [22]. In early work, Roy [110] presented a grounded speech learner that segments words from continuous speech. The problem discussed in this thesis is more complex, in that the aim is to ground full descriptions rather than words.

Projection into a common space has also been applied to the grounded spoken language models [124, 41, 42, 44, 43, 23, 21, 145, 16, 22]. These works use reccurrent or convolutional audiovisual neural networks to learn semantic similarity between single images and raw spoken utterances.

By contrast, the focus in this thesis is on multi-frame RGB-D percepts gathered from a sensor, aiming to identify individual objects rather than entire images. Additionally, as mentioned in 2.1.2 the speech corpora in many of these works consist of synthetically generated speech or read-aloud image captions. This may remove grammatical constructs, disfluencies, and speech repair, effectively gating the complexities of speech through written language. The dataset used in this thesis consists purely of people describing objects.

No previous work has compared grounding from raw speech with the widely used transcription-first approach. Additionally, I show how to create a speech-based grounding system based on complex perceptual data using a comparatively small number of data points, which is consistent with the requirements and available resources for implementing robotic systems. Compared to previous work, we leverage depth information and pretrained speech representation models to ground naturalistic spoken language in a model which converges with fewer data pairs; Harwath et al. [44] used 402,385 image-caption pairs and Chrupała et al.[23] specifically mention that the Flickr8K dataset of 40,000 image-caption pairs is small for the speech task.

Finally, the computational overhead of fine-tuning the model for specific domains is avoided, since they may change as the robot experiences new environments.

2.3 Language and Speech in Robotics

The role of language in robotics is wide-ranging [129], and the role of speech in particular is starting to receive significant attention [82]. Natural language is widely used in HRI tasks, for example in dialogue with assistive robots [69] or to facilitate human learning [72, 99, 115, 67, 106]. Speech-based HRI, in particular, has been applied to a wide variety of problems, such as emotion recognition [36, 141], social robotics [91, 1], and speech recognition. Mead and Matariç [87] determine how a robot should position itself for optimal speech and gesture recognition, complementing work on how people expect a robot to react when given instructions [92]. Speech is also an important source of insight into how different groups interact with robots, for example in assessing the communications of dementia patients through speech characteristics such as pitch [142].

2.4 Speech Processing Bias

While much previous work relies on ASR systems, these systems have known biases in their ability to recognize speech without errors. For now, most widely available ASR approaches depend on large-scale data [146]. These large datasets are usually derived from fairly heterogeneous groups [64]. Given this, we see gender [3, 127], race [13], disability status [37], and native language/dialect [50] disparities in successful ASR, reducing the accessibility of technology for underrepresented groups. Contemporaneous work [78] introduced a dataset to measure ASR performance by age, sex, and skin type. Technical remediation approaches remain scarce [88, 126],

although some work has been done in a multilingual grounded language [62]. Technical remediation methods have been introduced in a variety of ways such as datasets with expert labeled qualities [88] and adversarial training with inflectional perturbations of text [126]. However, work within the field of accessible natural language processing (NLP) tends to focus on text representations rather than speech itself. Within this work, we rather examine the remediation of bias towards demographics by removing the need for biased STT with direct-speech grounding within users interacting with a robotic system. To my knowledge, no previous work on grounded language learning has examined the impact of these factors on speech.

Chapter 3

GoLD: The Grounded Language Dataset

In this chapter, I describe the Grounded Language Dataset, or GoLD and its four modalities: RGB, depth, text, speech. The data is intended for use in training and evaluating multimodal grounded language models.

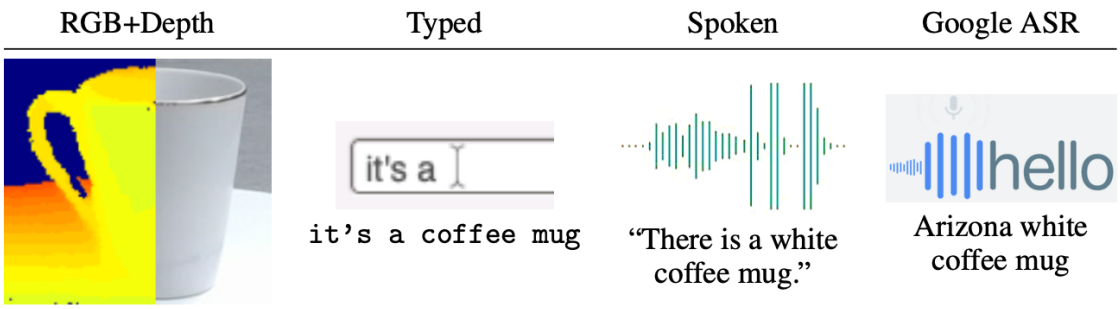

Figure 3.1: GoLD has RGB and depth point cloud images of 207 objects in 47 categories. It includes 16,500 text and 16,500 speech descriptions; all spoken descriptions include automatic transcriptions.

The GoLD dataset is the result of a collaboration between multiple members of the IRAL lab and was published at NeurIPS 2021 [60]. This dataset was initially recorded by Jenkins et al. [56] and developed in conjunction with Rishabh Sachdeva, myself, Padraig Higgins, Kasra Darvish, Edward Raff, Don Engel, John Winder, Francis Ferraro and Cynthia Matuszek. More information about the data collection can be found in the theses of Patrick Jenkins [57] and Rishabh Sachdeva [113]. In this work, I present an evaluation of an extended version of the existing dataset.

Table 3.1: Classes of objects in GoLD.

Topic	**Classes of Objects**
food	*potato, soda bottle, water bottle, apple, banana, bell pepper, food can, food jar, lemon, lime, onion*
home	*book, can opener, eye glasses, fork, shampoo, sponge, spoon, toothbrush, toothpaste, bowl, cap, cell phone, coffee mug, hand towel, tissue box, plate*
medical	*band aid, gauze, medicine bottle, pill cutter, prescription medicine bottle, syringe*
office	*mouse, pencil, picture frame, scissors, stapler, marker, notebook*
tool	*Allen wrench, hammer, measuring tape, pliers, screwdriver, lightbulb*

My novel contributions include: 13000 new spoken descriptions, 8500 new typed descriptions, and 552 speakers with voice annotations (in collaboration with Luke Richards and Ryan Barron).

GoLD is a collection of visual and English natural language data in five high-level groupings: *food*, *home*, *medical*, *office*, and *tools*. In these groups, 47 object classes (see table 3.1) contain 207 individual object instances. The dataset contains vision and depth images of each object from 450 different rotational views. From these, four representative 'keyframe' images were selected. These representative images were used to collect 16500 textual and 16500 spoken descriptions. The contents of the dataset are summarized in table 3.2.

Visual inputs were collected by rotating objects on a turntable in front of a

Table 3.2: Components of GoLD.

Categories (*e.g.*, medicine)	5	Images (vision + depth)	825
Classes (*e.g.*, apple)	47	Text descriptions	16500
Object instances (*e.g.*, `apple_3`)	207	Spoken descriptions	16500

commodity RGB-D (RGB + depth) video camera, as in [71]). For each object, four keyframes were manually selected to capture representative and diverse view angles for each object. Amazon Mechanical Turk workers were shown all four images and asked to provide either spoken or typed descriptions.

3.1 Vision + Depth Data Collection

Visual perception data were collected using a Microsoft Azure Kinect (i.e., a Kinect 3), a low-cost, high-fidelity commodity sensor that is widely used in robotics. For each object instance (*i.e.*, for each of the four staplers in the dataset), this sensor was used to collect raw image and point cloud videos.

The resulting dataset contains 207 90-second depth videos, one per instance, showing the object performing one complete rotation on a turntable. To ensure that each object has diverse views, e.g., examples of a mug with the handle occluded and visible, we manually selected 825 pairs of image and depth point cloud from 207 objects as representative frames, which we refer to as keyframes (examples are

shown in fig. 3.2).

Manually selecting keyframes avoids a known problem with many visual datasets: their tendency to show pictures of objects taken from a limited set of 'typical' angles [9]. For example, it is rare for a picture of a banana to be taken end-on. This aligns with our motivation of creating a dataset of household objects to support research on grounded language learning in an unstaged environment, as a robot looking at an object in a home may not see this typical view.

3.2 Text and Speech Description Collection

All descriptions were collected using Amazon Mechanical Turk (AMT).[1] Keyframes for randomly-chosen object instances were shown to the worker. They were asked to either type descriptions of objects in one or two short, complete sentences, or record descriptions using a microphone.

Collected speech was transcribed using Google's Speech to Text API, resulting in a spoken-language corpus of 16500 verbal descriptions. It should be noted that, although Mechanical Turk does not provide personally identifiable information about workers, it is possible that users may be identified by their voice or other side-channel information. For this reason, all collected language is limited to factual descriptions of simple household objects, and no value judgments, opinions, or emotional or potentially damaging subjects are discussed.

[1] See Ethical Considerations section, appendix.

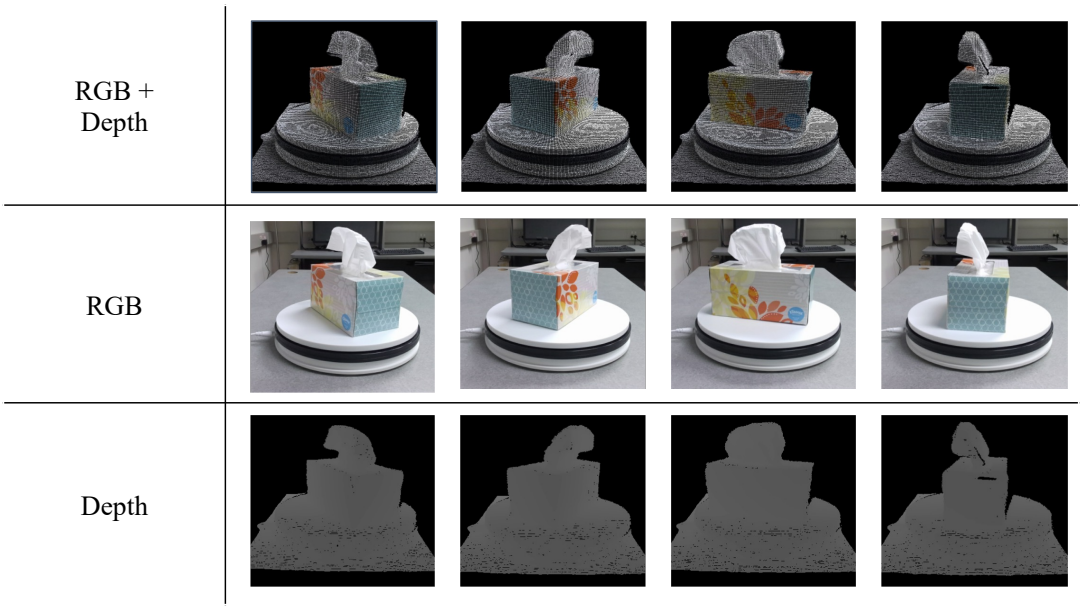

Figure 3.2: Samples showing keyframes in GoLD, along with the aligned 3D point cloud with depth information. Only the RGB image was shown to labelers.

3.3 Speaker Voice Qualities

We collected spoken descriptions from 552 Amazon Mechanical Turk workers. We labeled each of these workers based on perceived gender (man, woman, or undetermined),[2] accent (whether the speaker has a non-mid-American accent), creak (whether the user has a raspy, low-register voice), hoarseness (whether the speaker has a strained, breathy voice), muffled-ness (the level of distortion of the user's microphone, 1 to 3), volume (1 to 4), and level of background noise (1 to 4). Figure 3.3 shows the number of workers to whom each label has been attributed.

The labeling process was performed by randomly sampling five speech events

[2]Gender and sex are complex constructs. We asked annotators to choose the category that seemed to 'best describe' the speaker, but acknowledge the limitations of this approach.

Perceived Gender	Male	271	**Volume**	Low	10
	Female	**274**		Medium	157
	Undet	7		**High**	**331**
				Very High	54
Accented	**Yes**	**279**	**Background Noise**	**Low**	**366**
	No	273		Medium	143
Creak	Yes	194		High	39
	No	**358**		Very High	4
Hoarseness	Yes	48	**Muffledness**	**Low**	**393**
	No	**504**		Medium	119
				High	40

Figure 3.3: Number of workers labeled with each characteristic.

per user and annotating speakers based on the traits listed above. Some of these traits may vary for the same user from one example to the next. With that in mind, the annotations for those traits were done to reflect the majority case and may not apply to all examples provided by the user.

We intend for this data to be used as a test-bed for bias studies and other research into the performance of grounding models for different sub-populations. For example, a pilot study on this data has shown that accented users are particularly affected by the bias of speech-to-text models and that learning directly from raw speech can mitigate this bias.

Table 3.3: Human ratings of 250 automatic transcriptions. These ratings are designed strictly to assess the accuracy of the transcription, not the correctness of the spoken description with respect to the described object.

Rating	Transcription Quality Guidelines	#
1	wrong or gibberish / unusable sound file	28
2	slightly wrong (missing keywords / concepts)	21
3	pretty good (main object correctly defined)	33
4	perfect (accurate transcription and no errors)	168

3.4 Accuracy of Speech Transcriptions

Obtaining accurate transcriptions of speech in sometimes noisy environments is a significant obstacle to speech-based interfaces [73]. To understand the degree to which learning is affected by ASR errors, 250 randomly selected transcriptions were manually evaluated on a 4-point scale (see table 3.3). Of those, 80% are high quality ('perfect' or 'pretty good'), while only 11% are rated 'unusable.'

To get a more detailed understanding of transcription accuracy, we compare the ASR transcriptions and the human-provided transcriptions using the standard word error rate (WER) [98] and Bilingual Evaluation Understudy (BLEU) [97] scores. BLEU scores are widely used to measure the accuracy of language translations based on string similarity; we adopt this system to evaluate the goodness of transcriptions. BLEU is calculated by finding n-gram overlaps between machine translation and reference translations. We use tri-grams for our BLEU scores since

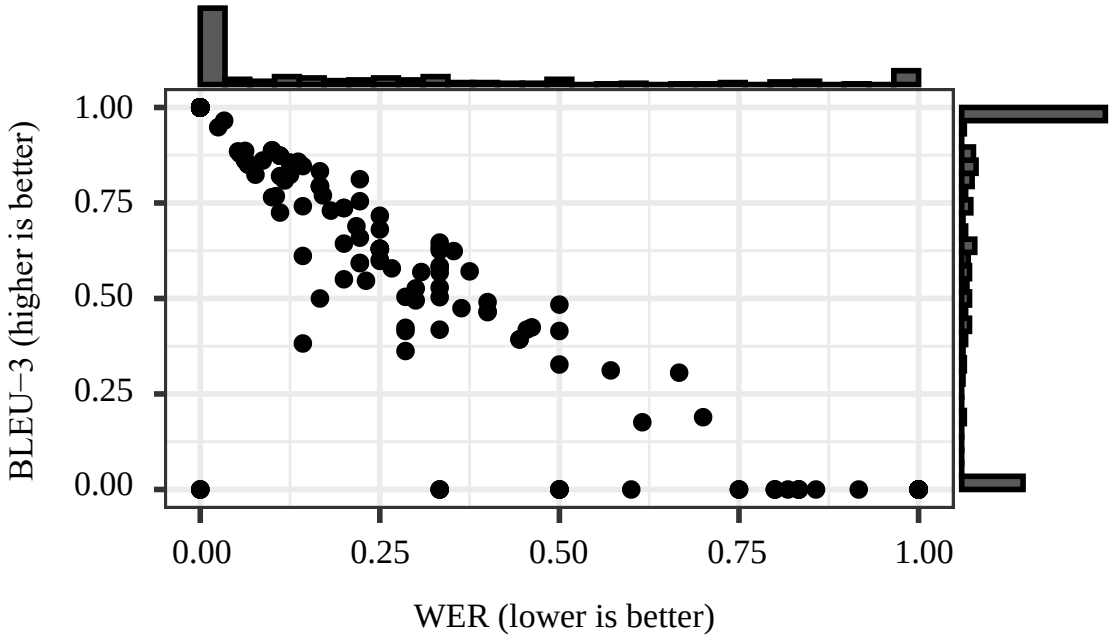

Figure 3.4: BLEU-3 and WER scores for 250 randomly selected speech transcriptions. A WER of 0 and a BLEU of 1 (top left corner) indicates perfect transcription. Marginal histograms show that some descriptions were perfectly transcribed.

some descriptions are shorter than four words such as "these are pliers", rendering a 4-gram BLEU score meaningless.

Figure 3.4 shows that many of the 250 manually transcribed descriptions were perfectly transcribed by automated speech-to-text. The marginal BLEU histogram shows more mistaken transcriptions (the second peak around 0) due to known problems with using BLEU to evaluate short sentences and tokens having mismatched capitalization or punctuation.

3.5 Comparative Analysis

Our initial hypothesis was that people would use more words when describing objects verbally than when typing, as it is a lower effort to talk than to type. Accordingly, we find that spoken descriptions are slightly longer than their textual counterparts ($p \geq 13.71$ using a Welch's t-test). While speech has more average words per description, 11.7, compared to text at 10.46, when stop words are removed the averages are 6.1 and 5.89 respectively. The larger mean decrease in speech descriptions is likely due to the tendency of ASR systems to interpret noise or murmur utterances as filler words, the inclusion of which has been shown to detract from meaning [123]. Text descriptions are more consistent in length than speech, with a standard deviation of 6.7 words for text, versus 9.51 words for transcribed speech. When we remove stop words, the standard deviation is 3.63 for text and 4.69 for speech.

Table 3.4 shows the top 20 most frequent words in both modalities. There is substantial overlap, as expected, since the same objects are being described, with colors dominating the lists. People use more filler words when describing objects using speech; for example, the word 'like' appears 889 times in speech data, whereas it was not significant in the text data.

Using the Stanford Part-of-Speech Tagger [134] to count the number of nouns, adjectives, and verbs between the two modalities yields no significant differences between the modalities. However, the word 'used' appears frequently (see table 3.4), typically to describe functionality. Developing grounded language models around

functionality for the analysis of affordances in objects [95] is an important research avenue that our dataset enables, which is not possible in prior datasets that do not contain the requisite modalities.

3.6 Dataset Distribution

The dataset is publicly available as a GitHub repository[3]. The repository contains three high-level datatypes: perception and language. The perceptual data is split into RGB-D images and depth data in the form of point clouds [112]. Each of these sets of data is subdivided by object class (e.g., "apple") and then further by instance (e.g., "apple #4"). The language is subdivided similarly, and for each object instance contains multiple speech descriptions (as .wav files) along with ASR transcriptions of that speech. Each instance also has multiple associated typed descriptions, which are not related to the spoken descriptions—they were provided by different workers at a separate time.

Each description of an instance also includes associated meta-data describing the data collection process. This includes: (1) a numeric identifier for the worker who provided each description; (2) the amount of time each description took to provide; and (3) the ground-truth category and instance label for each object.

The next chapters will demonstrate how the dataset enables grounded language learning research and investigation into the differences between spoken, transcribed and typed language.

[3]https://github.com/iral-lab/gold

Table 3.4: Top 20 most frequently used words in text (left) and speech (right) by percentage of occurrence in descriptions.

Token	**% Frequency**	**Token**	**% Frequency**
black	13.24	black	13.92
white	10.66	white	12.85
blue	9.97	blue	10.23
bottle	9.50	red	9.13
red	9.45	yellow	8.97
yellow	9.02	bottle	8.50
object	7.99	small	7.96
small	6.44	used	7.21
green	5.82	object	6.41
pair	5.27	green	5.85
used	5.21	plastic	5.30
handle	4.58	color	5.22
plastic	4.40	handle	4.85
silver	3.88	like	4.62
box	3.69	looks	3.99
label	2.92	silver	3.66
metal	2.79	turntable	3.33
pink	2.66	pair	3.32
light	2.44	box	3.21
scissors	2.43	label	3.01

Chapter 4

Learning Grounded Language from Percepts

This chapter provides an overview of the multimodal projection approach used for the grounded language learning experiments. Each experiment will combine RGB+depth images with a representation from one of the three language domains: typed text, transcribed speech, and speech audio. The approach described in this chapter was published at AAAI 2022 [61].

4.1 Perceptual Representations

In order to learn language from perceptual inputs, all modalities (RGB-D, typed text, transcribed text, and speech) must be featurized appropriately. In particular, because handling raw speech as perceptual input for grounding is a novel task, multiple speech representation models are experimented with. To avoid overfitting on this relatively small dataset, the feature extractor hyperparameters are not fine-tuned.

4.1.1 Visual Representations

Visual features are extracted using ResNet152 pre-trained on ImageNet [48], which achieves very strong results in image classification and object detection tasks (as a result, our system depends indirectly on labeled data by way of this pre-

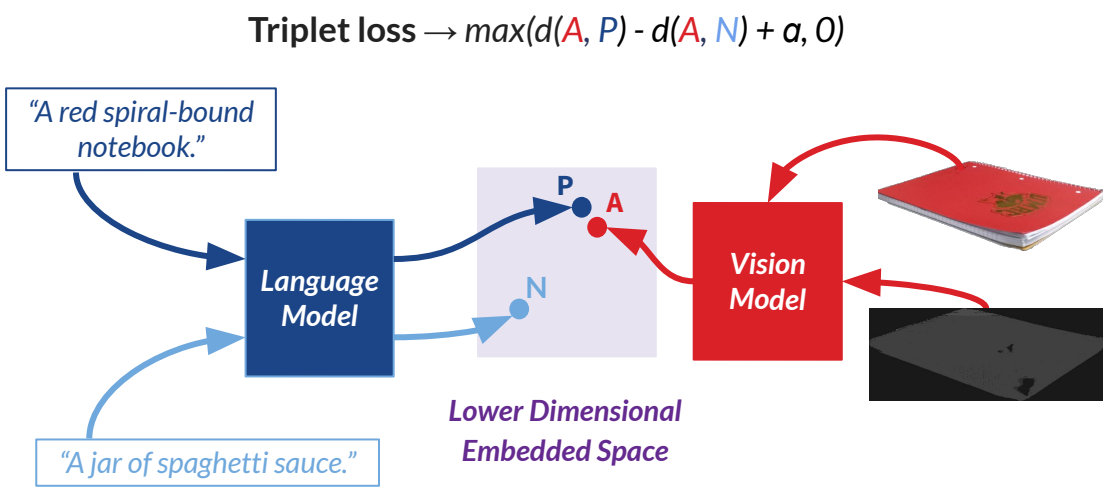

Figure 4.1: Our learning approach is to use manifold alignment in an attempt to capture a manifold between speech and visual perception. This approach is applied to grounded language acquisition by projecting visual and language representations into a shared latent space, where projections from both domains are closer to other projections of the same class. For example, the projection of the language utterance "I am seeing a white mug" should be close to the projection of the visual percepts of a white mug.

training). The last fully connected layer is removed to obtain the 2048-dimensional features used for classification. Both RGB and depth are processed through this network, the latter by colorizing depth images [109]. This yields two 2048-dimensional vectors, which are concatenated to create a multimodal object representation.

4.1.2 Text Representation

Language features for typed and transcribed text are obtained using BERT, a self-supervised bidirectional language model that achieves state-of-the-art perfor-

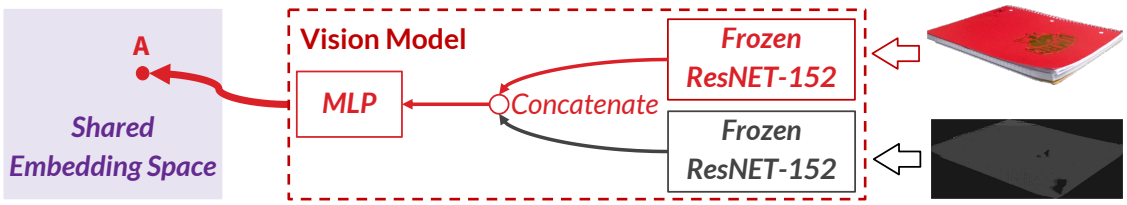

Figure 4.2: ResNet152 models pre-trained on ImageNet are used to featurize RGB and depth images and the resulting embeddings are concatenated and projected into the shared embedding space.

mance in multiple NLP tasks [31]. BERT's embeddings and linguistic performance make it useful for clustering [54], making BERT more appropriate for sentence-based language grounding than the commonly used words-as-classifiers approaches [116]. For a given natural language description, a 3,072 dimensional vector is obtained by extracting the average representation across the last four layers. In addition to the transcriptions from Google's speech-to-text API, transcriptions obtained from wav2vec 2.0 [7] are also considered as the model has been shown to achieve near state-of-the-art performance [96]. These transcriptions are directly comparable with our speech-based methods.

4.1.3 Speech Representation

Three different self-supervised speech models that have recently shown success in phoneme and speech recognition [76, 6, 7] are considered. The speech representations extracted from these models are intended to encode semantic information directly captured from raw speech; this is precisely the informational core that language acquisition seeks to capture. The hypothesis is that the process of mapping

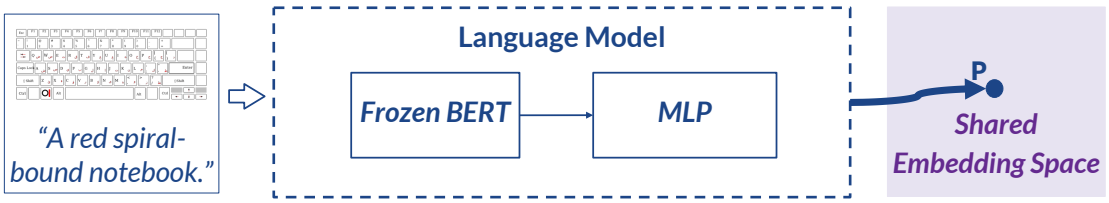

Figure 4.3: Typed descriptions are featurized using a pre-trained BERT model and the resulting embedding is projected into the shared embedding space.

raw speech representations to discrete transcriptions leads to a loss of information that may be detrimental to the performance of the grounding model. Therefore, learning directly from the representations extracted from these models is expected to reduce the effect of speech recognition errors on human-robot communication. A state-of-the-art model speech model, wav2vec 2.0 [7], and two other near state-of-the-art models in vq-wav2vec [6] and DeCoAR [76] are considered. The expectation is that performance of these three models will provide insights into the progress made and the overall direction of the field of acoustic representation learning.

4.1.3.1 Baseline: Mel-frequency cepstral coefficients

(MFCCs) [29] are a naïve baseline, which are widely used and easy to implement. MFCCs are inspired by the human auditory system and are extracted via a Discrete Fourier Transform analysis. They are frequently used in speech recognition systems, providing an effective comparison to using such systems directly. This baseline is used to evaluate how the highly pre-trained speech representation models compare to a simple speech feature extractor.

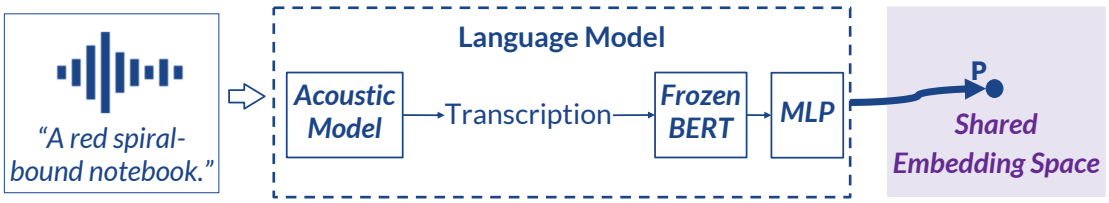

Figure 4.4: Spoken descriptions are fed into a pre-trained or off-the-shelf speech-to-text model and the textual output is featurized using a pre-trained BERT model. The resulting embedding is projected into the shared embedding space.

4.1.3.2 Model 1: DeCoAR

[76] is inspired by the vector-based word representation ELMo [100]. In ELMo, word vectors are learned from a contextualized bidirectional language model that is pre-trained on a large text corpus to predict the next word given a context. Unlike unidirectional language models, ELMo considers context from both directions. DeCoAR is an LSTM-based model that takes inspiration from ELMo's bidirectionality to learn deep contextualized acoustic representations, applying the same idea to speech by predicting a given slice of sound using past and future context through a backward and a forward LSTM. The sound is represented as sequential filterbank features. Combining these features, DeCoAR attempts to predict a given slice of sound by considering context from K steps ahead and behind. Existing pre-trained weights for DeCoAR are used.[1] DeCoAR was pre-trained on the LibriSpeech [96] dataset, which contains 960 hours of speech sampled at 16 kHz. The speech files from our dataset are sampled at 48kHz. Therefore, the first step was to downsample the speech files from 48kHz to 16kHz. The downsampled audio files are input into

[1] github.com/awslabs/speech-representations

Figure 4.5: Spoken descriptions are fed into a pre-trained speech model and the resulting embedding is projected into the shared embedding space.

the model and perform mean-pooling over the concatenated last layers from the forward and backward LSTMs, resulting in 2048-dimensional vectors.

4.1.3.3 Model 2: vq-wav2vec

[6] is based on wav2vec, a word2vec [89] inspired convolutional network, pre-trained on a context-prediction task to learn representations of audio data. The model outputs a series of 512 dimensional vectors representing 30ms of sound data with a stride of 10ms. The vq-wav2vec approach uses a two-step process: first, wav2vec is remodeled through quantization to output discrete units of speech, which are then fed through a BERT model pre-trained on speech signals to output final speech representations. This process results in two pre-trained models (vq-wav2vec and BERT). The authors' pre-trained weights are used for both models.[2]

The quantization process in vq-wav2vec involves replacing the continuous vector z_t at each timestep t of the wav2vec output Z with a discrete vector z'_t from a codebook with a fixed number of variable representations. This is done using k-means [137], where z'_t is chosen as the vector from the codebook which is closest to z_t using Euclidean distance. (vq-wav2vec is optimized using a context-prediction task,

as is as wav2vec). Audio data is then discretized through the pre-trained vq-wav2vec model and used to pre-train a BERT model. BERT's masked language modeling pre-training objective—in which random tokens are masked to achieve bidirectional language modeling—is modified to mask consecutive tokens in order to manage the brevity of speech tokens. vq-wav2vec was also pre-trained on the LibriSpeech ASR corpus. The downsampled audio files are provided as input to the pre-trained vq-wav2vec model, and the discrete tokens produced are then provided as input to the pre-trained BERT model. As with the text representations, 3072-dimensional vectors are extracted.

4.1.3.4 Model 3: wav2vec 2.0

[7] builds upon vq-wav2vec [6] in two key ways. First, the output of a wav2vec-like feature encoder is fed into a transformer. Second, quantization is used to discretize the feature extractor's output. However, the continuous representation produced by the feature extractor is fed into the transformer and the quantized representation is only used as target output during pre-training. Accordingly, the transformer learns over continuous representations rather than the discrete ones used in vq-wav2vec. The objective function is similar to BERT's [31] masked language modeling.

The speech encodings in wav2vec 2.0 are discretized through product quantization [55]. The continuous vector z_t at timestep t of the encoding Z is replaced with a vector z'_t such that z'_t is the concatenation of G discrete vectors chosen from

G codebooks using Gumbel Softmax [53]. In parallel, Z is also fed into the transformer with the random masking of a percentage of time steps and their subsequent N time steps. At each masked time step n the model attempts to identify the target representation z'_n among $K+1$ discrete representations including z'_n and K distractors.

Similarly to vq-wav2vec, pre-trained weights are used.[2] The pre-trained model was also fine-tuned for speech recognition using Connectionist Temporal Classification (CTC) [40]. The same process is used to extract the transcriptions mentioned earlier in the transcribed text representation section..

4.2 Approach

The problem of learning groundings from unconstrained speech is approached in an unconstrained environment. The visual percepts are RGB-D point clouds obtained from a mounted Kinect 3. RGB and depth are encoded using a sensor fusion convolution neural network for both RGB and depth [34, 109]. I experiment with learning from various pre-trained speech representation models, as well as from transcriptions. Our learning approach is to use manifold alignment with triplet loss [93] in an attempt to capture a manifold between speech and visual perception. This manifold represents the grounding between query language and objects in a selection task.

[2] github.com/pytorch/fairseq/tree/master/examples/wav2vec

4.2.1 Manifold Alignment

The manifold alignment approach of Nguyen et al. [93] is used in this thesis. Given two heterogeneous representations, the goal is to learn mappings to a shared latent space. This approach is applied to grounded language acquisition by projecting visual and language representations into a shared high-dimensional latent space, in which the projection of a language utterance that describes an object o of class c should be 'close' to the projections of other language utterances and visual percepts belonging to o, and to a broader degree other objects of class c, as seen in Fig. 4.1.

Triplet Loss is a popular geometric approach that has shown success in learning metric embeddings [117, 49, 33]. Learning uses triplets of the form (a, p, n), where a is an 'anchor' point, p is a positive instance of the same class as the anchor (e.g., mug), and n is a negative instance from a different class (e.g., apple). For each triplet, the embedding function f is learned so that the distance between a and n is maximized while the distance between a and p is minimized. This is achieved via the following loss function:

$$L = \max(d(f(a, m_a), f(p, m_p)) - d(f(a, m_a), f(n, m_n)) + \alpha, 0) \tag{4.1}$$

where d is a distance metric, m_x is the modality of point x, and α is a margin imposed between positive and negative instances. This approach lends itself well to a human learning scenario, in which a person could provide positive and negative examples of a given description or object.

Due to the heterogeneous nature of our problem, the embedding function f is the encoder that projects instances of a given modality into the shared manifold. A

different encoder is implemented for each modality as the input size and type are different. Each member of the triplet (a, p, n) can be selected from the vision or language domain. The domain is randomly selected. Cosine distance is used as the distance metric and a margin $\alpha = 0.4$.

Training. The 16,500 pairs of RGB-D data and descriptions from GoLD are split into training, validation and testing sets of respectively 13,040; 1,620; and 1,840 instances. Alignment models are trained with five different language representations: BERT embeddings for wav2vec 2.0 transcriptions in addition to MFCCs, vq-wav2vec, wav2vec 2.0 and DeCoAR embeddings for raw speech. All five are aligned with the visual features. All pre-trained feature extraction models are fixed during training. Only alignment models are optimized. Figure 4.2 provides an outline of the vision sub-network. Figure 4.3, 4.4, and 4.5 describe the language sub-network for respectively the typed-text, transcription and speech modalities. The default architecture of our alignment model is comprised of language and vision sub-networks that both consist of an input layer, two hidden layers with rectified linear units (ReLu) as activation functions and an output layer to obtain a final 1024 dimensional projection into the shared manifold. This architecture is used for the BERT, wav2vec 2.0, vq-wav2vec and DeCoAR embeddings.

Because of the low-dimensional nature of MFCCs, an LSTM-based language network is also considered. Instead of mean pooling, sequential MFCCs are input into a LSTM with 64 dimensional hidden states and concatenate the last 32 hidden states together resulting in a 2048 dimensional vector, which is input to a fully connected layer to obtain the final projection. All five methods are trained for a

total of 300 epochs using Adam [63] with a learning rate of 0.001 that is reduced by a factor of 10 after each hundred epochs.

4.2.2 User Trait-based Analysis

One of the goals of speech-based and other machine learning technologies is that they should be accessible, fair, and unbiased towards various demographics of users. In the following section, I outline how the differences in outcomes between transcription-based text versus raw speech approaches are analyzed for a variety of speaker traits.

4.2.2.1 Individual User Analysis

For embodied learning systems to be deployed effectively, they must be able to learn from individual users with a variety of speaker characteristics. To analyze the ability of the system to learn from individuals with a variety of speaker traits, speaker trait labels from the GoLD dataset are used. These labels are based on qualities in which speaker variance is known to affect the success of speech recognition models, *e.g.*, accented *vs.* unaccented speech. The results obtained when using wav2vec 2.0 speech representations are compared to those from wav2vec 2.0 transcriptions.

The dataset provides the ability to analyze individual users in the context of providing spoken learning examples. For testing the effectiveness of learning from individuals, this first evaluation is restricted to users who provide sufficient exemplars (described in 5.3). Each user contributed a variety of examples, each

with idiosyncrasies and unique perceptions of the description task. This offers a diverse set of user interactions. Each individual user's data is split into training and test sets and the learning system is evaluated based on its ability to learn successfully from individual speakers.

4.2.2.2 User-Group Analysis

To better understand the extent that user-specific traits affect the performance of the learned model, in a second evaluation, the training and evaluation is done over groups of multiple users with shared characteristics. In this analysis, speakers are split based on perceived gender, accent, muffled-ness, background noise and volume. The accessibility hurdles faced by members of minority populations in learning systems are well documented [50, 127, 64, 3]. These hurdles can be attributed in part to lack of representation of minority groups in large datasets, but other factors also come into play, especially with smaller datasets and feature-engineered methods. For each trait, the split is done such that each split has the same amount of training and testing data.

Chapter 5

Experimental Results and Discussion

To evaluate the trained models, we simulate object retrieval tasks with objects found in a sensed environment. The system is given a description and is responsible for selecting the correct objects from a subset of objects in GoLD. In a real-world setting, descriptions will often match multiple objects in an environment; in the dataset, natural language utterances describe the image they are associated with but they are also likely to describe other images of the same object or different objects of the same class. One of the main advantages of the manifold alignment approach is the possibility of retrieving multiple visual embeddings that are within a threshold of a language embedding in the shared manifold. While the model should be able to select the target object given a description, it should also be able to separate negative and positive instances using this threshold. In order to evaluate the model against these two goals, Mean Reciprocal Rank (MRR)-based and threshold-based evaluation tasks are considered. Each experiment will combine the RGB+depth images with language data from one of the three language domains: typed text, transcribed speech, and speech audio. Additional experiments are performed on combined typed text and transcribed speech data to test how the combination of the two might boost learning.

Figure 5.1: In the Triplet setting, the MRR is calculated from a set of 3 objects: the target object, an object from the same class and an object from a totally different class.

5.1 Downstream Object Retrieval

A robotic object retrieval task in which the goal is to retrieve the correct target vision instance i for a given language utterance is simulated. The system has N chances to pick an object given a description. The metric Mean Reciprocal Rank (MRR) measures how many tries are necessary for the correct object to be selected. The models are evaluated on the average of the reciprocal rank $\frac{1}{n_i}$ across all testing instances. The reciprocal rank is the inverse of the rank at which the target object was retrieved.

$$MRR = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{n_i} \tag{5.1}$$

The first retrieval setting is inspired by the triplets used for training. The MRR is calculated from a set of 3 objects: the target object, an object from the same class (but different instance) and an object from a different class. A real world example of this would be a robot picking between a green apple, a red apple and an orange when a description of a green apple is given. We will refer to the MRR performance

Figure 5.2: In the Subset setting, the MRR is calculated from a set of 5 objects: the target object and 4 randomly selected objects from different classes.

from this setting as Triplet MRR. A limitation of this metric is that in cases of multiple positive examples (e.g., two green apples), the system is over-penalized. To counter the effects of this limitation, we consider a second setting that involves 5 objects: the target and 4 randomly selected objects from different classes. The MRR performance from this setting is referred to as Subset MRR.

Table 5.1 compares the Google Transcriptions-based model to the typed text model. A combined "T + TS" model that is trained on a combination of typed text and transcribed speech is evaluated three separate times. First, it is tested individually on held-out sets where L is drawn first from text, then from speech. It is then evaluated on the combination of the two held-out sets. From the F1 evaluation, the transcribed speech model performs better than the other models, including the typed text model. These results seem to indicate that, despite potential errors in the transcription process, spoken input may be more linguistically meaningful than typed input. In all testing scenarios, there is little difference between the transcribed speech model and the combined text and transcription model.

	Triplet MRR	Subset MRR	F1
Google Transcriptions (GT)	**0.87 (±0.002)**	**0.96 (±0.004)**	**0.94 (±0.005)**
Typed Text (TT)	0.85 (±0.002)	0.89 (±0.004)	0.84 (±0.002)
TT + GT	0.87 (±0.001)	0.94 (±0.003)	0.92 (±0.003)
(Test on TT)	0.87 (±0.002)	0.96 (±0.004)	—
(Test on GT)	0.87 (±0.001)	0.94 (±0.001)	—
Random Baseline	0.61	0.46	—

Table 5.1: MRR & F1 results (higher is better). Transcriptions from Google's speech-to-text API and Typed Text with standard deviation over 5 runs. In addition to the queried object, the triplet setting includes an object from the same class and an object from a different class. The subset setting includes 4 objects from other classes.

Additionally, 5 speech-based methods including the wav2vec 2.0 transcriptions-based model, are evaluated using the two object retrieval tasks and the results are reported in table 5.2. In these results, wav2vec 2.0 represents the state-of-the-art in speech featurizations, and outperforms other approaches. As expected, learning a grounded language model from wav2vec 2.0 speech featurization approach outperforms text transcriptions from the same model. This confirms the hypothesis that the information lost in the transcription process negatively affects the performance of the grounding model. It is worth noting that adding another model for transcriptions to embeddings causes more latency between speech act and robot response.

Google's Speech to Text system is one of the more commonly used automatic speech recognition tools available for off-the-shelf ASR. While the overall results show an improvement as compared to wav2vec 2.0-based speech results, these results come with the costs associated with using a cloud-based service, including latency and dependence on a strong network connection during human-agent interaction. Perhaps more critically, because the underlying speech model and training data is not available, it is not possible to compare Google language model transcriptions to speech-based grounding using the same language model. The results shown in Table 5.2 suggest that if this experiment were to be performed, speech-based learning would likely still outperform transcribed text.The baseline featurization, an MFCC, performs no better than chance; DeCoAR and the vq-wav2vec methods both achieve reliable results but are outperformed by wav2vec 2.0.

5.2 Classification by threshold

The most intuitive way to deploy the alignment models is to define a fixed threshold t such that any object within radius t of a language utterance in the learned manifold is predicted to be described by that utterance. This can be defined as a binary classification task where an object falling within a radius t of a language utterance is a positive prediction. The held-out data is used to simulate this task by considering every visual percept of the same class as a language description to be a positive instance and randomly sampling the same number of percepts from other classes to be negative instances. The F1 measure of this task is reported, as

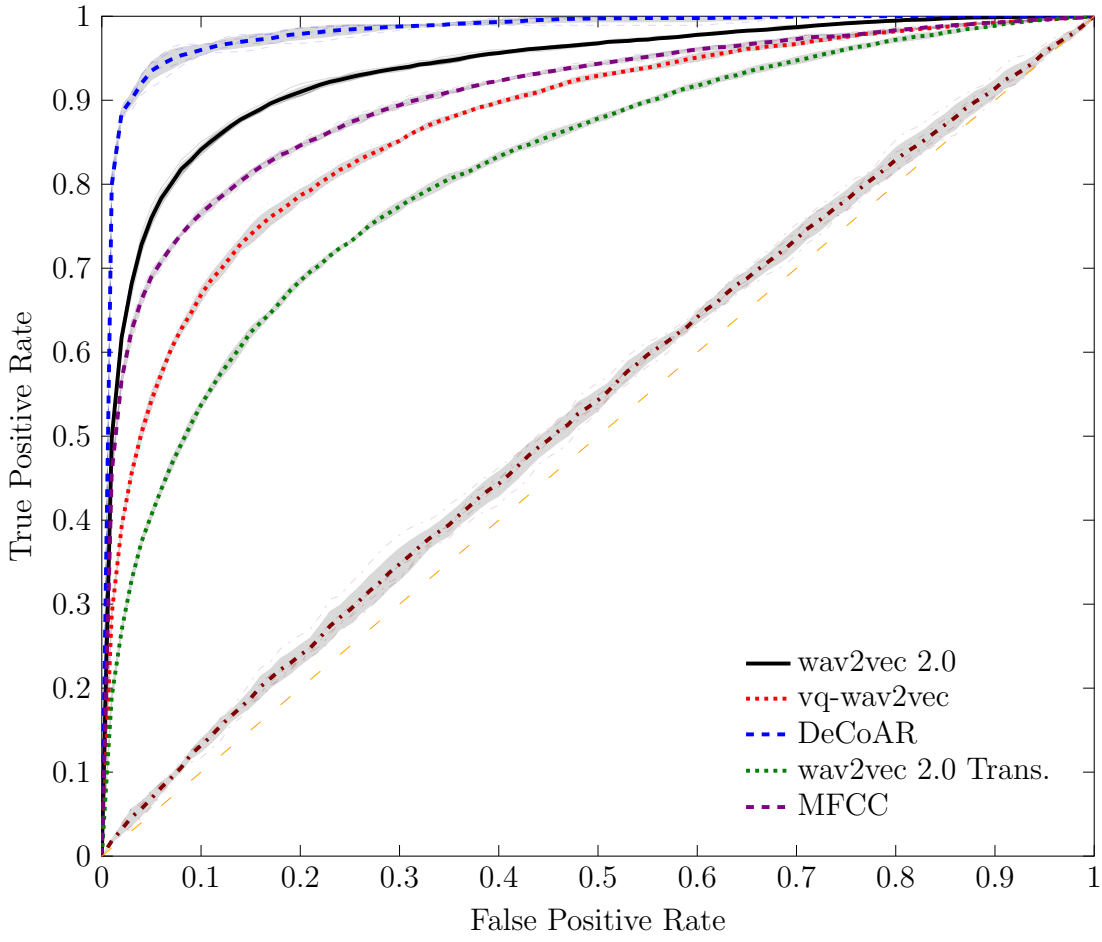

Figure 5.3: The ROC curves of each model on the validation set. The gray area around each curve represents the standard deviation of the model's performance over 5 runs. Higher AUC and closeness of the ROC curve to the top left corner mean that the model is better at discriminating between positive and negative examples of a given language description. The performance of the MFCC approach approximates that of a model with no skill.

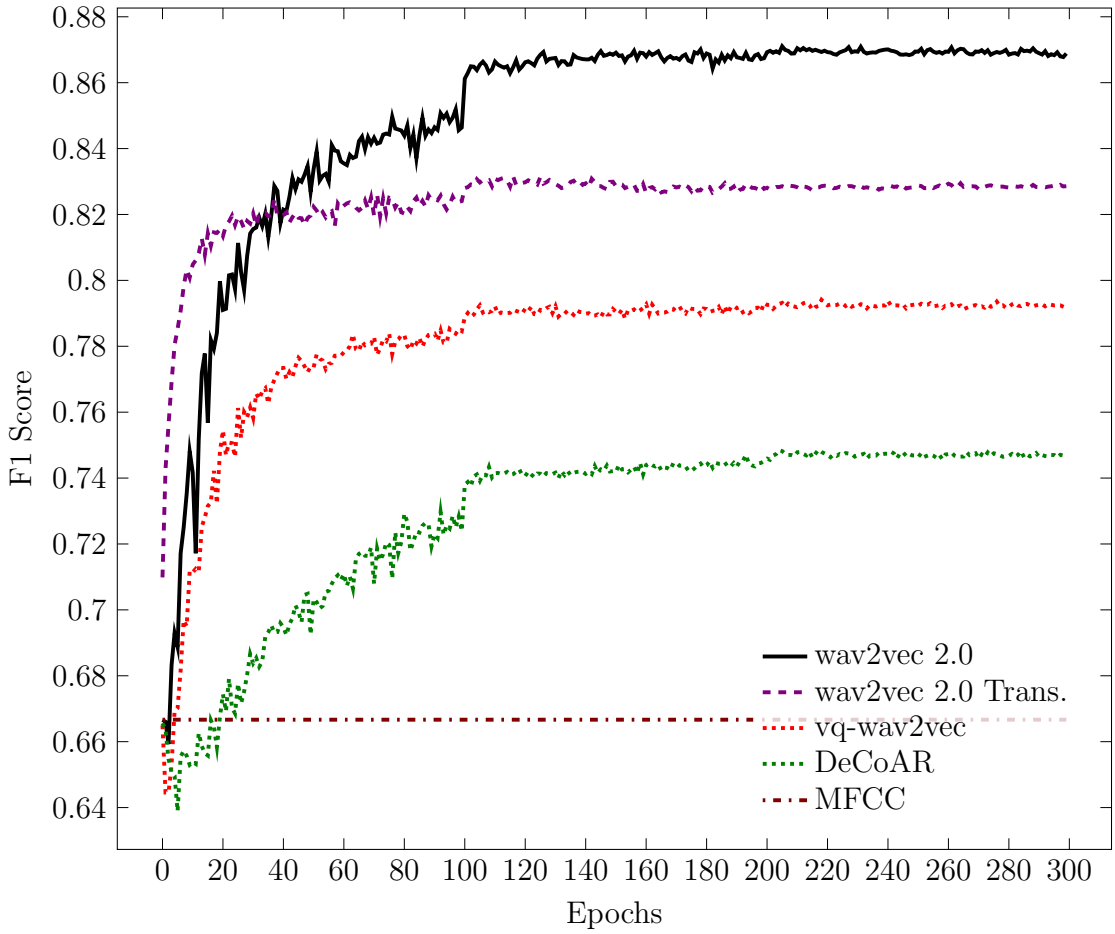

Figure 5.4: The convergence of F1 performance on the validation set as a function of training time (measured by training epochs). Each model's performance is averaged over 5 runs. Notice that the wav2vec 2.0 speech approach resulted in the best performance, followed by the transcriptions from wav2vec 2.0, while DeCoAR converges slightly faster than vq-wav2vec. The slight bumps at epoch 100 and 200 are due to a decrease in learning rate. MFCCs consistently underperformed.

	Triplet MRR	Subset MRR	F1
wav2vec 2.0 Speech	**0.85 (±0.002)**	**0.86 (±0.002)**	**0.83 (±0.003)**
wav2vec 2.0 Trans.	0.83 (±0.002)	0.83 (±0.002)	0.79 (±0.003)
vq-wav2vec	0.82 (±0.004)	0.78 (±0.004)	0.76 (±0.004)
DeCoAR	0.80 (±0.003)	0.72 (±0.004)	0.71 (±0.003)
MFCC	0.69 (±0.004)	0.49 (±0.01)	0.67 (±0)
Random Baseline	0.61	0.46	—

Table 5.2: MRR & F1 Results (higher is better) for speech representation models and transcriptions from wav2vec 2.0 with standard deviation over 5 runs. In addition to the queried object, the triplet setting includes an object from the same class and an object from a different class. The subset setting includes 4 objects from other classes. The wav2vec 2.0 Speech approach achieves the strongest performance.

we value both the model's precision and its recall.

A key aspect of this problem is to determine the value of t. The threshold t is tuned for every model on the validation set. The cosine distance is divided by 2 to normalize the values between 0 and 1. We find that a threshold of 0.35 for the text-based approaches; 0.4 for the wav2vec 2.0 approaches; and 0.45 for the vq-wav2vec and DeCoAR approaches achieve peak performance. The MFCC approach achieves peak performance at a threshold of 1, which results in the model only making true predictions, since every instance—positive or negative—will be within a radius of 1 of the description. This indicates that the MFCC approach

performs poorly regardless of threshold. The ROC curves in Fig. 5.3 confirm this assessment and show that all speech approaches except the MFCC approach learn to discriminate between positive and negative examples of language descriptions.

The obtained thresholds are applied to the testing set and the results are reported in table 5.1 and table 5.2. Figure 5.4 shows the evolution of the F1 scores of the 5 speech-based approaches on the validation set over the course of training. Again, the wav2vec 2.0 Speech approach performed better than wav2vec 2.0 transcriptions, and the DeCoAR and vq-wav2vec methods achieve reasonable F1 scores of 0.71 and 0.76. The F1 for the typed text, Google Transcriptions, and combined models are .84, .94, and .92, respectively as shown in Table 5.1 These results further support the conclusions reached in Section 5.1, confirming that grounded language acquisition from raw speech can lead to tangible results that surpass those of the traditional "transcription-first" method.

5.3 Speaker Traits Study

For each user, a language alignment model and a vision alignment model are both trained with the manifold alignment learning objective described above. We exclude users from the dataset who did not provide at least 2 examples for at least 5 object classes. Furthermore, the remaining users' examples from the object classes for which they have provided less than 2 examples are excluded. These constraints both guarantee that only users who have provided sufficient examples for meaningful evaluation are included, and ensure there are 5 object classes for the subset MRR

evaluation. This leaves us with 87 users sub-selected from the complete dataset. These speakers provide an average count of 61.5 examples each, with a median of 35.

On average, training was performed on 40.7 examples. Even though the models take advantage of domain encoding for speech, language, and vision, this small amount of training data per user is still a challenge. On average there are 20.8 testing examples per user. This is taken into account when analyzing the end performance of the model. Of the 87 speakers, 50.5% had accents. For gender, 39.1% were annotated as men, 57.4% as women, and 3.5% as undetermined. 24.1% of the users had creak, 4.6% had hoarseness, 11.5% had high levels of muffled-ness. 2.3% of users had low volume, 82.8% had medium volume, and 14.9% had high volume. 90.8% of users had low background noise and 9.2% of users had high background noise. Multiple kinds of background noise were heard in the samples, including alarms, children, and fans. These noises contribute to real-world situation representation in the data. Due to the low amount of users with hoarse voices, the hoarseness trait is excluded from the individual user study.

5.3.1 Individual User-based Model Performance

For each user, two models are trained using the transcriptions and speech embeddings from the wav2vec 2.0 model. The Pearson correlation coefficient (PCC) between each of the labeled voice traits and MRR scores is used to see which factors cause variation in both methods and which groups are most affected by the loss of

information that occurs during transcription.

$$PCC(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2(y_i - \overline{y})^2}} \tag{5.2}$$

where X is the set of the values of a speech trait for all considered individual users and Y is the set of MRR scores for the same users.

The correlation results for the subset MRR are shown in Fig. 5.5. The results are mostly similar across both MRR metrics.

As expected, we find that for both methods, performance is negatively correlated with accent, creak and background noise and positively correlated with volume and the number of examples provided by the user. Background noise has a slight decreased correlation with the speech method compared to the correlation with the transcription method. This may be a benefit of not strictly mapping to a language token but rather a discretized high-dimensional value.

A key takeaway from the experiment is the significant gap between the correlation of the two models' performance with accented language, in which language models learned directly from accented speech are less negatively affected than those learned from transcriptions of that speech. In terms of the overall effects, the difference in subset MRR between the accented speakers and the non-accented ones for the transcription-first approach is triple the difference of the raw speech approach (approximately 6% vs 2%). This gap suggests that accented users are the most affected by the information loss of the transcription process. It was expected that the noisier nature of the raw speech representations would help the learner in alleviating bias. The less negative correlation with accent supports this claim.

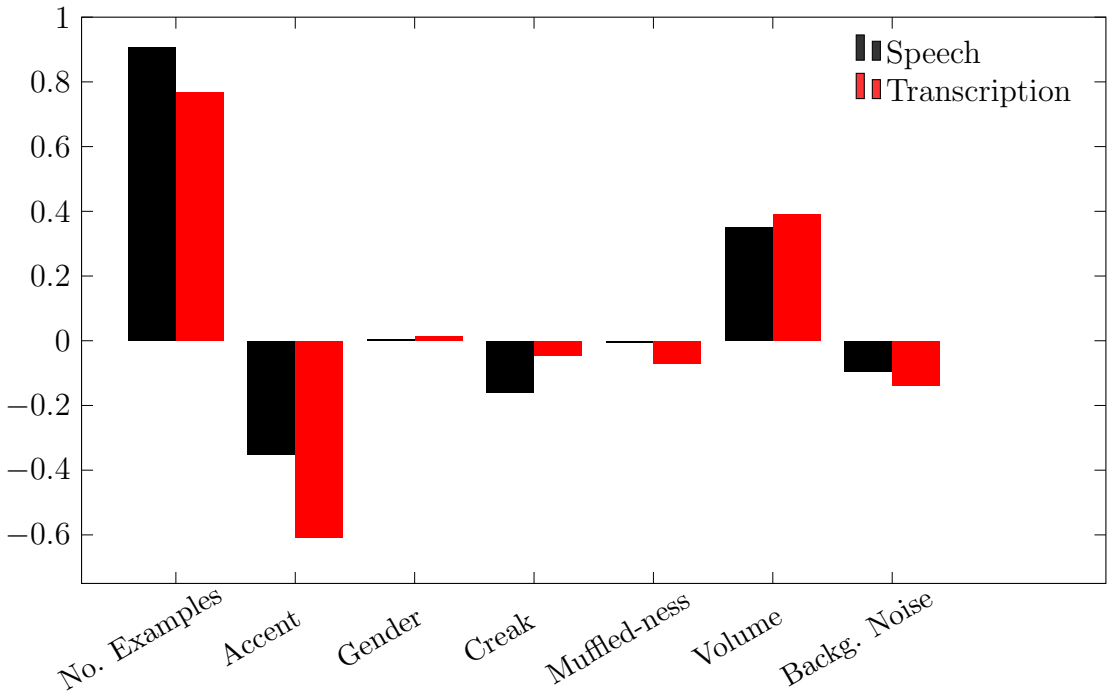

Figure 5.5: The correlation between Subset MRR performance and different user qualities for the wav2vec2.0 speech and transcription methods. Accent is negatively correlated with performance in both, but the correlation is stronger when using transcriptions. The difference in performance is less pronounced for other speaker traits.

5.3.2 Group-based Model Performance

Finally, the speech data is split based on groups of users. Workers who were excluded from the individual-user study are included in these experiments, as aggregating the user data allows for proper training and test splits. Data is split between accented and unaccented users, users with creak and without creak, perceived men and women, low (1), medium (2,3), and high volume (4), low (1,2) and high (3,4) background noise, and low (1,2) and high (3) muffled-ness. For each trait, the number training and testing examples is controlled by splitting the data into equally sized groups with the same number of training and testing examples. As with the individual-user study, it is ensured that, for each group, the model is tested on object classes seen in training.

wav2vec 2.0 speech and transcriptions methods are trained and tested on these splits. Results are reported in table 5.3. For the volume splits, there was a steady increase from low to high at each tier. For both models, the decrease in performance for accented users noticed in the individual-user study is absent. In the individual-user study, the performance is captured through statistics that utilize linear relationships and show that each participant has unique combinations of analyzed and non-analyzed characteristics that play a factor on each axis but when grouping, the variance is addressed. We see this analysis as a critical step for further investigations into analysis and technical methods to support the study of bias within individual user understanding.

		Transcriptions		Speech	
	No. examples	Subset	Triplet	Subset	Triplet
non-accent	8191	0.82	0.85	0.84	0.85
accent	8191	**0.92**	**0.89**	**0.93**	**0.90**
non-creak	4932	0.83	0.85	0.85	0.85
creak	4932	**0.85**	**0.86**	**0.87**	**0.87**
low volume	350	0.62	0.75	0.55	0.73
med. volume	350	0.64	0.77	0.57	0.72
high volume	350	**0.66**	**0.78**	**0.58**	**0.74**
men	7897	0.84	0.85	0.85	0.86
women	7897	**0.89**	**0.87**	**0.91**	**0.89**
low background noise	1352	**0.74**	**0.81**	**0.68**	**0.79**
high background noise	1352	0.73	**0.81**	**0.68**	**0.79**
low muffledness	1610	**0.74**	**0.81**	**0.72**	**0.80**
high muffledness	1610	0.73	0.80	0.70	0.79

Table 5.3: MRR scores for User Group-splits with wav2vec 2.0 transcriptions and speech methods. Higher is better. For both models, performance increases as volume goes from low to high at each tier. However, there is no decrease in performance for accented users.

Chapter 6

Conclusion

This thesis focused on the problem of speech-based grounded language learning. Particularly, I tackled the issue of speaker-based variability. As robots move into human spaces, they will participate in spoken interactions with human users from different backgrounds and with different speech characteristics. The goal of this research was to study the differences between learning from typed utterances and spontaneous spoken utterances (with or without intermediate transcriptions) and analyze the effects of learning from direct speech representations (vs. intermediate transcriptions) for individuals and demographics.

6.1 Future Work

A limitation of the GoLD dataset is that the post-data-collection speaker trait annotations cannot be assumed to be the ground truth. While this process allows for interesting early investigation into the effects of the annotated speaker characteristics, ground-truth data would support more conclusive experiments. For example, voice characteristics like creak and hoarseness are hard to measure in a subjective annotation process. In contrast, collecting self-reported biographical information such as age would allow for more informative and reliable experiments. Additionally, the accent annotation process in GoLD only accounts for the presence or absence of

Standard American Accent as it is not trivial to pinpoint a speaker's accent. Collecting demographic data like race, nationality and first-language would allow for more fine-grained studies into the effects of accent on grounded language learning performance. Obtaining demographic data directly from the speakers would also help to control over- and under-representation of different groups in the collected dataset.

While object retrieval is a well-suited task for these initial experiments, a good next step would be to expand this approach to more challenging robotic language grounding tasks. Text-based grounding models are performing increasingly well on the popular ALFRED [120] benchmark, where models are trained to navigate in various environments, interact with a large diverse set of objects and complete tasks based on human-provided natural language instructions. Tasks like ALFRED require a better and more sequential understanding of language; and analogous speech-based benchmarks would create interesting challenges for speech models.

Finally, the work presented in this thesis focuses on investigating bias in language grounding models. A necessary next step is the development of bias mitigation techniques. It would be interesting to apply existing bias mitigation strategies such as speaker embeddings [148, 135, 50] and multi-task learning[107, 74] to the grounding problem.

As interest in spoken language grounding continues to grow, it will be of vital importance to ensure that state-of-the-art models are inclusive and work for the widest demographic.

6.2 Contributions

In chapter 3, I presented the GoLD dataset and its four modalities of input (text, speech, RGB and depth) that allows us to tackle new challenges in grounded language learning such as learning directly from speech audio.

In chapter 4, I introduced a triplet-loss based multimodal learning approach that leverages pretrained self-supervised speech representation models to efficiently learn spoken natural language groundings.

In chapter 5, the quality of the data in GoLD was investigated by performing grounded language learning experiments from typed text, transcriptions and raw speech. Additionally, the experiments showed that it is both possible and effective to directly learn natural language groundings from raw speech data to visual percepts, without having to rely on the intermediate textual representations of prior work. The results demonstrate that direct grounding of speech to vision can minimize information loss and enable more reliable human-agent communication. The investigation into direct grounding of speech includes a user study that identifies speaker/audio traits that historically affect speech recognition. The study showed that accented users are most affected by this information loss; direct grounding to raw speech has the potential benefit of reducing systems performance bias toward these and potentially other populations of users. While identifying and assigning these traits is preliminary, these initial results are relevant to bias and effectiveness in deployed, real-world systems.

Bibliography

[1] Samer Al Moubayed, Jonas Beskow, Bajibabu Bollepalli, Joakim Gustafson, Ahmed Hussen-Abdelaziz, Martin Johansson, Maria Koutsombogera, José David Lopes, Jekaterina Novikova, Catharine Oertel, et al. Human-robot collaborative tutoring using multiparty multimodal spoken dialogue. In *HRI*, 2014.

[2] Muhannad Alomari, Paul Duckworth, David C Hogg, and Anthony G Cohn. Natural language acquisition and grounding for embodied robotic systems. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2017.

[3] Eiman Alsharhan and Allan Ramsay. Investigating the effects of gender, dialect, and training size on the performance of Arabic speech recognition. *Language Resources and Evaluation*, 54(4):975–998, 2020.

[4] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-Language Navigation: Interpreting Visually-Grounded Navigation Instructions in Real Environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3674–3683, 2018.

[5] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433, 2015.

[6] Alexei Baevski, Steffen Schneider, and Michael Auli. Vq-wav2vec: Self-Supervised Learning of Discrete Speech Representations. In *ICLR*, 2020.

[7] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *NeurIPS*, 2020.

[8] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, February 2019. ISSN 1939-3539.

[9] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Advances in Neural Information Processing Systems*, pages 9453–9463, 2019.

[10] Emanuele Bastianelli, Danilo Croce, Andrea Vanzo, Roberto Basili, and Daniele Nardi. A Discriminative Approach to Grounded Spoken Language Understanding in Interactive Robotics. In *IJCAI*, pages 2747–2753, 2016.

[11] Philipp Beckerle, Gionata Salvietti, Ramazan Unal, Domenico Prattichizzo, Simone Rossi, Claudio Castellini, Sandra Hirche, Satoshi Endo, Heni Ben Amor, Matei Ciocarlie, et al. A Human–robot interaction Perspective on Assistive and rehabilitation robotics. *Frontiers in Neurorobotics*, 11(24):1, 2017.

[12] Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. Experience Grounds Language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, Online, November 2020. Association for Computational Linguistics.

[13] Su Lin Blodgett and Brendan O'Connor. Racial disparity in natural language processing: A case study of social media African-American English. *arXiv preprint arXiv:1707.00061*, 2017.

[14] Liefeng Bo, Xiaofeng Ren, and Dieter Fox. Depth kernel descriptors for object recognition. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 821–826, September 2011.

[15] Angelo Cangelosi. Grounding language in action and perception: From cognitive agents to humanoid robots. *Physics of Life Reviews*, 7(2):139–151, June 2010. ISSN 1571-0645.

[16] Khyathi Raghavi Chandu, Yonatan Bisk, and Alan W Black. Grounding 'Grounding' in NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4283–4305, Online, August 2021. Association for Computational Linguistics.

[17] David L. Chen and Raymond J. Mooney. Learning to sportscast: A test of grounded language acquisition. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 128–135, New York, NY, USA, July 2008. Association for Computing Machinery. ISBN 978-1-60558-205-4.

[18] David L Chen and Raymond J Mooney. Learning to interpret natural language navigation instructions from observations. In *Proceedings of the National Conference on Artificial Intelligence AAAI*, 2011.

[19] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. Microsoft COCO Captions: Data Collection and Evaluation Server, April 2015.

[20] Maxime Chevalier-Boisvert, Dzmitry Bahdanau, Salem Lahlou, Lucas Willems, Chitwan Saharia, Thien Huu Nguyen, and Yoshua Bengio. BabyAI: First Steps Towards Grounded Language Learning With a Human In the Loop. In *ICLR*, 2019.

[21] Grzegorz Chrupa\la. Symbolic Inductive Bias for Visually Grounded Learning of Spoken Language. In *ACL*, 2019.

[22] Grzegorz Chrupała. Visually Grounded Models of Spoken Language: A Survey of Datasets, Architectures and Evaluation Techniques. *Journal of Artificial Intelligence Research*, 73:673–707, February 2022. ISSN 1076-9757.

[23] Grzegorz Chrupa\la, Lieke Gelderloos, and Afra Alishahi. Representations of language in a model of visually grounded speech signal. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 613–622, 2017.

[24] Herbert H. Clark and Susan E. Brennan. Grounding in communication. In *Perspectives on Socially Shared Cognition*, pages 127–149. American Psychological Association, Washington, DC, US, 1991. ISBN 978-1-55798-121-9.

[25] Vanya Cohen, Benjamin Burchfiel, Thao Nguyen, Nakul Gopalan, Stefanie Tellex, and George Konidaris. Grounding Language Attributes to Objects using Bayesian Eigenobjects. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1187–1194, Macau, China, November 2019. IEEE. ISBN 978-1-72814-004-9.

[26] Silvia Coradeschi, Amy Loutfi, and Britta Wrede. A Short Review of Symbol Grounding in Robotic and Intelligent Systems. *KI - Künstliche Intelligenz*, 27 (2):129–136, May 2013. ISSN 1610-1987.

[27] Timothee Cour, Chris Jordan, Eleni Miltsakaki, and Ben Taskar. Movie/Script: Alignment and Parsing of Video and Text Transcription. In *Proceedings of the 10th European Conference on Computer Vision: Part IV*, ECCV '08, pages 158–171, Berlin, Heidelberg, October 2008. Springer-Verlag. ISBN 978-3-540-88692-1.

[28] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2054–2063, 2018.

[29] Steven Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4), 1980.

[30] Chaorui Deng, Qi Wu, Qingyao Wu, Fuyuan Hu, Fan Lyu, and Mingkui Tan. Visual Grounding via Accumulated Attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7746–7755, 2018.

[31] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT)*, 2019.

[32] Fethiye Irmak Doğan, Ilaria Torre, and Iolanda Leite. Asking Follow-Up Clarifications to Resolve Ambiguities in Human-Robot Conversation. In *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '22, pages 461–469, Sapporo, Hokkaido, Japan, March 2022. IEEE Press.

[33] Xingping Dong and Jianbing Shen. Triplet Loss in Siamese Network for Object Tracking. In *ECCV*, September 2018.

[34] Andreas Eitel, Jost Tobias Springenberg, Luciano Spinello, Martin A. Riedmiller, and Wolfram Burgard. Multimodal deep learning for robust RGB-D object recognition. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2015.

[35] Jeanine C Evers. From the past into the future. How technological developments change our ways of data collection, transcription and analysis. In *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research*, 2011.

[36] Kerstin Fischer, Malte Jung, Lars Christian Jensen, and Maria Vanessa aus der Wieschen. Emotion expression in HRI–when and why. In *HRI*, 2019.

[37] Raymond Fok, Harmanpreet Kaur, Skanda Palani, Martez E Mott, and Walter S Lasecki. Towards more robust speech interactions for deaf and hard of hearing users. In *Proc. of the ACM Conf. on Computers and Accessibility*, 2018.

[38] Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. Speaker-Follower Models for Vision-and-Language Navigation. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[39] Sadaoki Furui, Masanobu Nakamura, Tomohisa Ichiba, and Koji Iwano. Why Is the Recognition of Spontaneous Speech so Hard? In Václav Matoušek, Pavel Mautner, and Tomáš Pavelka, editors, *International Conference on Text, Speech and Dialogue*, Lecture Notes in Computer Science, pages 9–22, Berlin, Heidelberg, 2005. Springer. ISBN 978-3-540-31817-0.

[40] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *ICML*, 2006.

[41] David Harwath and James Glass. Deep multimodal semantic embeddings for speech and images. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 237–244, 2015.

[42] David Harwath, Antonio Torralba, and James R. Glass. Unsupervised Learning of Spoken Language with Visual Context. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 1866–1874, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 978-1-5108-3881-9.

[43] David Harwath, Galen Chuang, and James Glass. Vision as an Interlingua: Learning Multilingual Semantic Embeddings of Untranscribed Speech. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4969–4973, Calgary, AB, April 2018. IEEE. ISBN 978-1-5386-4658-8.

[44] David Harwath, Adria Recasens, Didac Suris, Galen Chuang, Antonio Torralba, and James Glass. Jointly Discovering Visual Objects and Spoken Words from Raw Sensory Input. In *ECCV*, 2018.

[45] Jun Hatori, Yuta Kikuchi, Sosuke Kobayashi, Kuniyuki Takahashi, Yuta Tsuboi, Yuya Unno, Wilson Ko, and Jethro Tan. Interactively Picking Real-World Objects with Unconstrained Spoken Language Instructions. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3774–3781, May 2018.

[46] William Havard, Laurent Besacier, and Olivier Rosec. SPEECH-COCO: 600k Visually Grounded Spoken Captions Aligned to MSCOCO Data Set. *GLU 2017 International Workshop on Grounding Language Understanding*, August 2017.

[47] William N. Havard, Jean-Pierre Chevrot, and Laurent Besacier. Models of Visually Grounded Speech Signal Pay Attention to Nouns: A Bilingual Experiment on English and Japanese. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8618–8622, 2019.

[48] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[49] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[50] Arthur Hinsvark, Natalie Delworth, Miguel Del Rio, Quinten McNamara, Joshua Dong, Ryan Westerman, Michelle Huang, Joseph Palakapilly, Jennifer Drexler, Ilya Pirkin, et al. Accented Speech Recognition: A Survey. *arXiv preprint arXiv:2104.10747*, 2021.

[51] Wei-Ning Hsu, David Harwath, Tyler Miller, Christopher Song, and James Glass. Text-Free Image-to-Speech Synthesis Using Learned Segmental Units. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural*

Language Processing (Volume 1: Long Papers), pages 5284–5300, Online, August 2021. Association for Computational Linguistics.

[52] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. Natural language object retrieval. In *CVPR*, 2016.

[53] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *5th International Conference on Learning Representations, ICLR*, 2017.

[54] Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. What Does BERT Learn about the Structure of Language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy, July 2019. Association for Computational Linguistics.

[55] Herve Jegou, Matthijs Douze, and Cordelia Schmid. Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1):117–128, 2010.

[56] Patrick Jenkins, Rishabh Sachdeva, Gaoussou Youssouf Kebe, Padraig Higgins, Kasra Darvish, Edward Raff, Don Engel, John Winder, Francis Ferraro, and Cynthia Matuszek. Presentation and Analysis of a Multimodal Dataset for Grounded Language Learning, September 2020.

[57] Patrick D. Jenkins. *Transfer Learning of Grounded Language Models for Use in Robotic Systems*. M.S., University of Maryland, Baltimore County, United States – Maryland, 2020.

[58] Johanna E. Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1988–1997, 2016.

[59] Andrej Karpathy, Armand Joulin, and Li F Fei-Fei. Deep Fragment Embeddings for Bidirectional Image Sentence Mapping. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[60] Gaoussou Youssouf Kebe, Padraig Higgins, Patrick Jenkins, Kasra Darvish, Rishabh Sachdeva, Ryan Barron, John Winder, Don Engel, Edward Raff, Francis Ferraro, and Cynthia Matuszek. A Spoken Language Dataset of Descriptions for Speech-Based Grounded Language Learning. In *NeurIPS*, 2021.

[61] Gaoussou Youssouf Kebe, Luke E. Richards, Edward Raff, Francis Ferraro, and Cynthia Matuszek. Bridging the Gap: Using Deep Acoustic Representations to Learn Grounded Language from Percepts and Raw Speech. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI-22)*, 2022.

[62] Caroline Kery, Nisha Pillai, Cynthia Matuszek, and Francis Ferraro. Building Language-Agnostic Grounded Language Learning Systems. In *Robot and Human Interactive Communication*, 2019.

[63] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[64] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. Racial disparities in automated speech recognition. *Proc. of the National Academy of Sciences (PNAS)*, 117(14), 2020.

[65] Thomas Kollar, Stefanie Tellex, Deb Roy, and Nicholas Roy. Toward understanding natural language directions. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 259–266. IEEE, 2010.

[66] Thomas Kollar, Vittorio Perera, Daniele Nardi, and Manuela Veloso. Learning environmental knowledge from task-based human-robot dialog. In *2013 IEEE International Conference on Robotics and Automation*, pages 4304–4309, May 2013.

[67] Hatice Köse, Pınar Uluer, Neziha Akalın, Rabia Yorgancı, Ahmet Özkul, and Gökhan Ince. The effect of embodiment in sign language tutoring with assistive humanoid robots. *Int. Journal of Social Robotics*, 7(4):537–548, 2015.

[68] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1): 32–73, 2017.

[69] Vladimir A Kulyukin. On natural language dialogue with assistive robots. In *HRI*, pages 164–171, 2006.

[70] Gitta P M Laan. PERCEPTUAL DIFFERENCES BETWEEN SPONTANEOUS AND READ ALOUD SPEECH. *Proc. of the Institute of Phonetic Sciences Amsterdam.*, 16:65–79, 1992.

[71] Kevin Lai, Liefeng Bo, Xiaofeng Ren, and Dieter Fox. A Large-Scale Hierarchical Multi-View RGB-D Object Dataset. In *2011 IEEE International Conference on Robotics and Automation*, pages 1817–1824, May 2011.

[72] Sungjin Lee, Hyungjong Noh, Jonghoon Lee, Kyusong Lee, Gary Geunbae Lee, Seongdae Sagong, and Munsang Kim. On the effectiveness of robot-assisted language learning. *ReCALL: the Journal of EUROCALL*, 23(1):25, 2011.

[73] Bo Li, Yu Tsao, and Khe Chai Sim. An investigation of spectral restoration algorithms for deep neural networks based noise robust speech recognition. In *Annual Conference of the International Speech Communication Association*, 2013.

[74] Bo Li, Tara N. Sainath, Khe Chai Sim, Michiel Bacchiani, Eugene Weinstein, Patrick Nguyen, Zhifeng Chen, Yanghui Wu, and Kanishka Rao. Multi-Dialect Speech Recognition with a Single Sequence-to-Sequence Model. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4749–4753, Calgary, AB, Canada, April 2018. IEEE Press.

[75] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, Lecture Notes in Computer Science, pages 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1.

[76] Shaoshi Ling, Yuzong Liu, Julian Salazar, and Katrin Kirchhoff. Deep Contextualized Acoustic Representations for Semi-Supervised Speech Recognition. *Conf. on Acoustics, Speech, and Signal Processing*, 2020.

[77] Changsong Liu, Shaohua Yang, Sari Saba-Sadiya, Nishant Shukla, Yunzhong He, Song-Chun Zhu, and Joyce Chai. Jointly learning grounded task structures from language instruction and visual demonstration. In *EMNLP*, 2016.

[78] Chunxi Liu, Michael Picheny, Leda Sarı, Pooja Chitkara, Alex Xiao, Xiaohui Zhang, Mark Chou, Andres Alvarado, Caner Hazirbas, and Yatharth Saraf. Towards Measuring Fairness in Speech Recognition: Casual Conversations Dataset Transcriptions. *arXiv preprint arXiv:2111.09983*, 2021.

[79] Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. Federated Learning for Vision-and-Language Grounding Problems. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):11572–11579, April 2020. ISSN 2374-3468.

[80] Mateusz Malinowski and Mario Fritz. A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[81] Jonathan Malmaud, Earl Wagner, Nancy Chang, and Kevin Murphy. Cooking with Semantics. In *Proceedings of the ACL 2014 Workshop on Semantic Parsing*, pages 33–38, Baltimore, MD, June 2014. Association for Computational Linguistics.

[82] Matthew Marge, Carol Espy-Wilson, Nigel G. Ward, Abeer Alwan, Yoav Artzi, Mohit Bansal, Gil Blankenship, Joyce Chai, Hal Daumé, Debadeepta Dey, Mary Harper, Thomas Howard, Casey Kennington, Ivana Kruijff-Korbayová, Dinesh Manocha, Cynthia Matuszek, Ross Mead, Raymond Mooney, Roger K. Moore, Mari Ostendorf, Heather Pon-Barry, Alexander I. Rudnicky, Matthias Scheutz, Robert St Amant, Tong Sun, Stefanie Tellex, David Traum, and Zhou Yu. Spoken language interaction with robots: Recommendations for future research. *Computer Speech & Language*, 71:101255, 2022. ISSN 0885-2308.

[83] Cynthia Matuszek. Grounded Language Learning: Where Robotics and NLP Meet. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 5687–5691, Stockholm, Sweden, July 2018. International Joint Conferences on Artificial Intelligence Organization. ISBN 978-0-9992411-2-7.

[84] Cynthia Matuszek, Dieter Fox, and Karl Koscher. Following Directions Using Statistical Machine Translation. In *HRI*, 2010.

[85] Cynthia Matuszek, Nicholas FitzGerald, Luke Zettlemoyer, Liefeng Bo, and Dieter Fox. A joint model of language and perception for grounded attribute learning. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, pages 1435–1442, Madison, WI, USA, June 2012. Omnipress. ISBN 978-1-4503-1285-1.

[86] Cynthia Matuszek, Liefeng Bo, Luke Zettlemoyer, and Dieter Fox. Learning from unscripted deictic gesture and language for human-robot interactions. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

[87] Ross Mead and Maja J Matarić. Autonomous human-robot proxemics: A robot-centered approach. In *HRI*, pages 573–573, 2016.

[88] Josh Meyer, Lindy Rauchenstein, Joshua D Eisenberg, and Nicholas Howell. Artie Bias Corpus: An Open Dataset for Detecting Demographic Bias in Speech Applications. In *LREC*, 2020.

[89] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NeurIPS*, 2013.

[90] So Yeon Min, Devendra Singh Chaplot, Pradeep Kumar Ravikumar, Yonatan Bisk, and Ruslan Salakhutdinov. FILM: Following Instructions in Language with Modular Methods. In *International Conference on Learning Representations*, September 2021.

[91] Christophe Mollaret, Alhayat Ali Mekonnen, Julien Pinquier, Frédéric Lerasle, and Isabelle Ferrané. A multi-modal perception based architecture for a non-intrusive domestic assistant robot. In *HRI*, 2016.

[92] Pooja Moolchandani, Cory J. Hayes, and Matthew Marge. Evaluating Robot Behavior in Response to Natural Language. In *HRI*, pages 197–198, 2018.

[93] Andre T. Nguyen, Luke E. Richards, Gaoussou Youssouf Kebe, Edward Raff, Kasra Darvish, Frank Ferraro, and Cynthia Matuszek. Practical Cross-Modal Manifold Alignment for Robotic Grounded Language Learning. In *CVPR Workshops*, pages 1613–1622, June 2021.

[94] Khanh Nguyen and Hal Daumé III. Help, Anna! Visual Navigation with Natural Multimodal Assistance via Retrospective Curiosity-Encouraging Imitation Learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 684–695, 2019.

[95] Thao Nguyen, Nakul Gopalan, Roma Patel, Matthew Corsaro, Ellie Pavlick, and Stefanie Tellex. Robot Object Retrieval with Contextual Natural Language Queries. In *Proceedings of Robotics: Science and Systems*, Corvalis, Oregon, USA, July 2020.

[96] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. Librispeech: An ASR corpus based on public domain audio books. In *Conf. on Acoustics, Speech, and Signal Processing*, 2015.

[97] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[98] Youngja Park, Siddharth Patwardhan, Karthik Visweswariah, and Stephen C Gates. An empirical analysis of word error rate and keyword error rate. In *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[99] Askarbek Pazylbekov, Daryn Kalym, Anuar Otynshin, and Anara Sandygulova. Similarity attraction for robot's dialect in language learning using social robots. In *HRI*, 2019.

[100] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep Contextualized Word Representations. In *North American ACL*, 2018.

[101] Nisha Pillai and Cynthia Matuszek. Unsupervised selection of negative examples for grounded language learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[102] Nisha Pillai, Karan K Budhraja, and Cynthia Matuszek. Improving grounded language acquisition efficiency using interactive labeling. *UMBC Student Collection \textcolorredno \textcolorredno \textcolorredno \textcolorredno \textcolorredno \textcolorredno \textcolorredno*, 2016.

[103] Nisha Pillai, Cynthia Matuszek, and Francis Ferraro. Deep Learning for Category-Free Grounded Language Acquisition. In *Proc. of the NAACL Combined Workshop on Spatial Language Understanding and Grounded Communication for Robotics (NAACL-SpLU-RoboNLP)*, Minneapolis, MI, USA, June 2019.

[104] Kunal Pratap Singh, Suvaansh Bhambri, Byeonghwi Kim, Roozbeh Mottaghi, and Jonghyun Choi. Factorizing Perception and Policy for Interactive Instruction Following. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1868–1877, Montreal, QC, Canada, October 2021. IEEE. ISBN 978-1-66542-812-5.

[105] Yanyuan Qiao, Chaorui Deng, and Qi Wu. Referring Expression Comprehension: A Survey of Methods and Datasets. *IEEE Transactions on Multimedia*, 23:4426–4440, 2021. ISSN 1941-0077.

[106] Aditi Ramachandran, Chien-Ming Huang, Edward Gartland, and Brian Scassellati. Thinking Aloud with a Tutoring Robot to Enhance Learning. In *HRI*, pages 59–68, 2018.

[107] Kanishka Rao and Haşim Sak. Multi-accent speech recognition with hierarchical grapheme based models. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4815–4819, New Orleans, LA, USA, March 2017. IEEE Press.

[108] Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. Collecting Image Annotations Using Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 139–147, Los Angeles, June 2010. Association for Computational Linguistics.

[109] Luke E Richards, Kasra Darvish, and Cynthia Matuszek. Learning Object Attributes with Category-Free Grounded Language from Deep Featurization. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020.

[110] Deb Roy. Grounded spoken language acquisition: Experiments in word learning. *IEEE Transactions on Multimedia*, 5(2):197–209, 2003.

[111] Deb K. Roy. Learning visually grounded words and syntax for a scene description task. *Computer Speech & Language*, 16:353–385, 2002.

[112] Radu Bogdan Rusu and Steve Cousins. 3D is here: Point Cloud Library (PCL). In *2011 IEEE International Conference on Robotics and Automation*, Shanghai, China, May 2011.

[113] Rishabh Sachdeva. *Speech Vs Textual Data for Grounded Language Learning*. M.S., University of Maryland, Baltimore County, United States – Maryland, 2020.

[114] Dario D Salvucci. Interactive Grounding and Inference in Instruction Following. *Topics in Cognitive Science*, 2021.

[115] Brian Scassellati, Jake Brawer, Katherine Tsui, Setareh Nasihati Gilani, Melissa Malzkuhn, Barbara Manini, Adam Stone, Geo Kartheiser, Arcangelo Merla, Ari Shapiro, et al. Teaching language to deaf infants with a robot and a virtual human. In *HCI*, pages 1–13, 2018.

[116] David Schlangen, Sina Zarrieß, and Casey Kennington. Resolving References to Objects in Photographs using the Words-As-Classifiers Model. In *ACL*, 2016.

[117] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, June 2015.

[118] Pararth Shah, Marek Fiser, Aleksandra Faust, J. Kew, and Dilek Z. Hakkani-Tür. FollowNet: Robot Navigation by Following Natural Language Directions with Deep Reinforcement Learning. *ArXiv*, abs/1805.06150, 2018.

[119] Mohit Shridhar, Dixant Mittal, and David Hsu. INGRESS: Interactive visual grounding of referring expressions. *Int'l Journal of Robotics Research*, 39(2-3), 2020.

[120] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks. In *CVPR*, 2020.

[121] Carina Silberer and Mirella Lapata. Learning Grounded Meaning Representations with Autoencoders. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 721–732, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[122] M. Skubic, D. Perzanowski, A. Schultz, and W. Adams. Using spatial language in a human-robot dialog. In *Proceedings 2002 IEEE International Conference on Robotics and Automation (Cat. No.02CH37292)*, volume 4, pages 4143–4148 vol.4, May 2002.

[123] Andreas Stolcke and Jasha Droppo. Comparing human and machine errors in conversational speech transcription. *Annual Conference of the International Speech Communication Association*, 2017.

[124] Gabriel Synnaeve, Maarten Versteegh, and Emmanuel Dupoux. Learning words from images and speech. In *In NIPS Workshop on Learning Semantics*, 2014.

[125] Mariarosaria Taddeo and Luciano Floridi. Solving the symbol grounding problem: A critical review of fifteen years of research. *Journal of Experimental & Theoretical Artificial Intelligence*, 17(4):419–445, December 2005. ISSN 0952-813X.

[126] Samson Tan, Shafiq Joty, Min-Yen Kan, and Richard Socher. It's Morphin' Time! Combating Linguistic Discrimination with Inflectional Perturbations. In *58th Conf. of the ACL*, 2020.

[127] Rachael Tatman. Gender and dialect bias in YouTube's automatic captions. In *ACL Workshop on Ethics in Natural Language Processing*, 2017.

[128] Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew Walter, Ashis Banerjee, Seth Teller, and Nicholas Roy. Understanding natural language commands for robotic navigation and mobile manipulation. In *AAAI*, 2011.

[129] Stefanie Tellex, Nakul Gopalan, Hadas Kress-Gazit, and Cynthia Matuszek. Robots that use language. *Annual Review of Control, Robotics, and Autonomous Systems*, 3:25–55, 2020.

[130] Jesse Thomason, Shiqi Zhang, Raymond J Mooney, and Peter Stone. Learning to interpret natural language commands through human-robot dialog. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[131] Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. Vision-and-dialog navigation. In *Conference on Robot Learning*, pages 394–406. PMLR, 2020.

[132] Jesse Thomason, Aishwarya Padmakumar, Jivko Sinapov, Nick Walker, Yuqian Jiang, Harel Yedidsion, Justin Hart, Peter Stone, and Raymond Mooney. Jointly improving parsing and perception for natural language commands through human-robot dialog. *Journal of Artificial Intelligence Research*, 67:327–374, 2020.

[133] Ye Kyaw Thu, Takuya Ishida, Naoto Iwahashi, Tomoaki Nakamura, and Takayuki Nagai. Symbol grounding from natural conversation for human-robot communication. In *5th Int'l Conf. on Human Agent Interaction*, 2017.

[134] Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 173–180. ACL, 2003.

[135] Mehmet Ali Tuğtekin Turan, Emmanuel Vincent, and Denis Jouvet. Achieving Multi-Accent ASR via Unsupervised Acoustic Model Adaptation. In *INTERSPEECH 2020*, October 2020.

[136] Emre Ünal, Ozan Arkan Can, and Yücel Yemez. Visually Grounded Language Learning For Robot Navigation. In *1st International Workshop on Multimodal Understanding and Learning for Embodied Applications*, pages 27–32, 2019.

[137] Aäron van den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *NeurIPS*, 2017.

[138] Andrea Vanzo, Danilo Croce, Emanuele Bastianelli, Roberto Basili, and Daniele Nardi. Grounded language interpretation of robotic commands through structured learning. *Artificial Intelligence*, 278, 2020.

[139] Paul Vogt. The physical symbol grounding problem. *Cognitive Systems Research*, 3(3):429–457, September 2002. ISSN 1389-0417.

[140] Ayelet Weizs. How Long it Really Takes to Transcribe (Accurate) Audio, July 2019.

[141] Andrew B Williams, Rosa M Williams, Ronald E Moore, and Matthias McFarlane. Aida: A social co-robot to uplift workers with intellectual and developmental disabilities. In *HRI*, 2019.

[142] Toshiki Yamanaka, Yutaka Takase, and Yukiko I Nakano. Assessing the communication attitude of the elderly using prosodic information and head motions. In *HRI*, 2016.

[143] Yuya Yoshikawa, Yutaro Shigeto, and Akikazu Takeuchi. STAIR Captions: Constructing a Large-Scale Japanese Image Caption Dataset. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 417–421, Vancouver, Canada, 2017. Association for Computational Linguistics.

[144] Xiaoxue Zang, Ashwini Pokle, Marynel Vázquez, Kevin Chen, Juan Carlos Niebles, Alvaro Soto, and Silvio Savarese. Translating Navigation Instructions in Natural Language to a High-Level Plan for Behavioral Robot Navigation. In *EMNLP*, 2018.

[145] Mingxin Zhang, Tomohiro Tanaka, Wenxin Hou, Shengzhou Gao, and Takahiro Shinozaki. Sound-Image Grounding Based Focusing Mechanism for Efficient Automatic Spoken Language Acquisition. In *INTERSPEECH*, pages 4183–4187, 2020.

[146] Yaodong Zhang and James R Glass. Unsupervised spoken keyword spotting via segmental DTW on Gaussian posteriorgrams. In *2009 IEEE Workshop on Automatic Speech Recognition & Understanding*, pages 398–403. IEEE, 2009.

[147] Yichi Zhang and Joyce Chai. Hierarchical Task Learning from Language Instructions with Unified Transformers and Self-Monitoring. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4202–4213, Online, August 2021. Association for Computational Linguistics.

[148] Yong Zhao, Jinyu Li, Shixiong Zhang, Liping Chen, and Yifan Gong. Domain and Speaker Adaptation for Cortana Speech Recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5984–5988, Calgary, AB, Canada, April 2018. IEEE Press.

[149] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning Deep Features for Scene Recognition Using Places Database. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'14, pages 487–495, Cambridge, MA, USA, 2014. MIT Press.

[150] Fengda Zhu, Yi Zhu, Xiaojun Chang, and Xiaodan Liang. Vision-Language Navigation With Self-Supervised Auxiliary Reasoning Tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10012–10022, 2020.