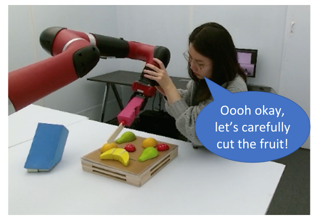
Understanding Acoustic Patterns of Human Teachers Demonstrating Manipulation Tasks to Robots

Akanksha Saran*[1], Kush Desai*[2], Mai Lee Chang[2], Rudolf Lioutikov[3], Andrea Thomaz[2] and Scott Niekum[4]

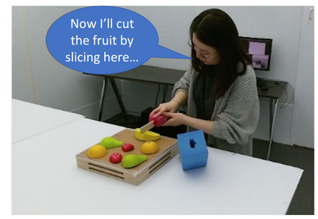
***Abstract*— Humans use audio signals in the form of spoken language or verbal reactions effectively when teaching new skills or tasks to other humans. While demonstrations allow humans to teach robots in a natural way, learning from trajectories alone does not leverage other available modalities including audio from human teachers. To effectively utilize audio cues accompanying human demonstrations, first it is important to understand what kind of information is present and conveyed by such cues. This work characterizes audio from human teachers demonstrating multi-step manipulation tasks to a situated Sawyer robot using three feature types: (1) duration of speech used, (2) expressiveness in speech or prosody, and (3) semantic content of speech. We analyze these features along four dimensions and find that teachers convey similar semantic concepts via spoken words for different conditions of (1) demonstration types, (2) audio usage instructions, (3) subtasks, and (4) errors during demonstrations. However, differentiating properties of speech in terms of duration and expressiveness are present along the four dimensions, highlighting that human audio carries rich information, potentially beneficial for technological advancement of robot learning from demonstration methods.**

I. Introduction

Human speech or audio is a natural, low-effort and rich channel of communication [1]. Typically, robot learning from demonstration (LfD) algorithms ignore information carried by a human teacher's audio cues, and only work with the state of the environment and the human teacher's actions [2]. A primary reason why incorporating this additional information has been challenging is the lack of understanding about how complex human audio signals are used and what they convey during demonstrations. Recent advancements in sensor technologies and speech processing algorithms [3] make it possible to extract informative features from speech. *Raw* human audio carries more information than is present in a transcribed narration of textual words produced by an automatic speech recognition (ASR) system [3]. Audio can



(a) Kinesthetic Demonstration



(b) Video Demonstration

Fig. 1: (a) Kinesthetic demonstrations and (b) Video demonstrations along with unrestricted audio signals of human teachers are collected for two manipulation tasks.

contain information via spoken language, as well as out-of-vocabulary words not part of a natural-language learning corpus, disfluencies (restarts, repetitions, and self-corrections), filled pauses ('um', 'uh'), hyperarticulations such as careful enunciation, slow speaking rate, increased pitch and loudness [4]. This work is a first step in analyzing how humans use unrestricted audio cues during demonstrations for multi-step manipulation tasks—characterizing duration, prosodic features, and spoken words.

We work in a limited-data regime with an in-house dataset of demonstration data and accompanying human audio, collected in a laboratory setting. With the help of human annotators, we characterize human speech via: (1) duration of speech in a demonstration, (2) annotated and computational prosodic features, (3) semantic content of words spoken during speech. We analyze these features across different dimensions: (a) type of demonstration used (kinesthetic teaching using the robot's arm versus the human performing the task themselves), (b) the instruction given to a teacher about usage of speech during demonstrations (explicit narration instructions versus implicit indication to use speech), (c) presence or absence of relevant subtasks being executed, (d) presence or absence of errors during demonstrations. We find that users convey similar semantic concepts through spoken words across all independent variable conditions. However, we also observe that teachers are more expressive but talk less densely during kinesthetic teaching compared to demonstrating themselves. Moreover, teachers are equally expressive across audio usage instructions but overall talk more when explicitly asked to do so. With some acoustic features, we see support for the fact that while teachers use speech more densely in the absence of errors and in gaps

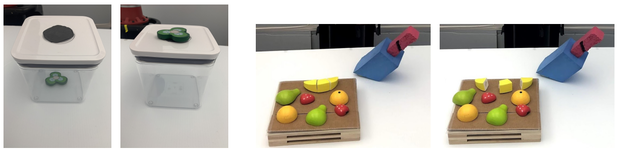
[1]Akanksha Saran is with Microsoft Research, New York, NY 10012, USA. Work done at the University of Texas at Austin. akanksha.saran@microsoft.com

[2]Kush Desai, Mai Lee Chang and Andrea Thomaz are with the Department of Electrical and Computer Engineering, University of Texas at Austin, Austin, TX 78712, USA. kushkdesai@utexas.edu, mlchang@utexas.edu, athomaz@ece.utexas.edu

[3]Rudolf Lioutikov is with the Karlsruhe Institute of Technology, Karlsruhe, BW 76131, Germany. lioutikov@kit.edu

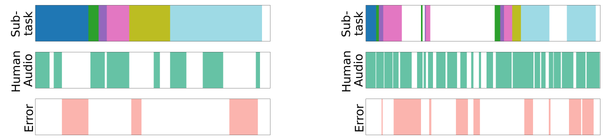
[4]Scott Niekum is with the Department of Computer Science, University of Texas at Austin, Austin, TX 78712, USA. sniekum@cs.utexas.edu

* Equal contribution.



(a) Box Opening (b) Fruit Cutting

Fig. 2: Start and end conditions of two manipulation tasks for which demonstrations are provided to a Sawyer robot.



(a) Natural Instruction Type (b) Narration Instruction Type

Fig. 3: Subtasks, human utterances, and errors during demonstrations are shown on the top, middle, and bottom rows respectively during kinesthetic demonstrations for the fruit cutting task (each subtask category represented with a different color, all error categories represented by the same color).

between relevant subtask executions, they emphasize their speech more *during* subtasks and errors.

Finally, with a proof-of-concept experiment, we show that human acoustic features are useful to detect presence of relevant subtasks and errors during demonstrations. Taken together, our findings highlight that human speech carries rich information about demonstrations, which can be beneficial for technological advancement of robot learning algorithms in the future.

II. Related Work

A. Human Audio-assisted Human-Robot Interaction

Scassellati et al. [5] showed that understanding emotion in the context of a conversation is an essential skill for keeping humans engaged when interacting with robots. They developed a service that helps a robot to recognize acoustic patterns in a human's tone of voice by classifying approving, neutral and prohibitive affects. Kim et al. [6] identified three phases during which humans use affective prosody during interactive reinforcement learning. These three phases are: direction (before the learner acts), guidance (as the learner indicates intent) and feedback (after the learner completes a task-action). Short et al. [7] used ratio of variances in raw speech features to detect contingent human responses to robot probes for open-world human-robot interaction. However, in comparison to the problem settings of these prior works, we focus on the learning from demonstration paradigm and characterize raw audio signals of human teachers to better understand the information present in this modality. Work on infant-directed demonstrations [8] as well as robotics work inspired from infant-directed speech [9] highlights that the way that people talk to a robot could possibly help it learn more easily if they treat it like it is an infant, than if they spoke to it like it is an adult. However, this may not always be possible with a variety of commercialized robots not exhibiting infant-like appearances or personalities. In our work, we study how human demonstrators use unrestricted speech to teach robots which are not infant-like in appearance or behavior, to better inform how human audio cues could be leveraged for learning by a variety of robots.

B. Human Audio-assisted Robot Learning

Prior research in learning from demonstration [2] has utilized human speech signals accompanying demonstrations, however with a restricted vocabulary of words used by the human teacher. Nicolescu et al. [10] demonstrated the role of verbal cues both during demonstrations and as feedback from the human teacher during the agent's learning process, to facilitate learning of navigation behaviors on a mobile robot. However, they restricted human teachers to use a limited vocabulary of words to indicate relevant parts of the workspace or actions that a robot must execute. Similarly, Pardowitz et al. [11] used a fixed set of seven vocal comments which are mapped one-to-one with features relevant to the task to augment subtask similarity detection and learning of the task model from demonstrations for a simple table setting task. Prior work in reinforcement learning has also utilized a fixed vocabulary of words to record voice-based feedback for reward shaping [12] and reinforcement learning [13] However, in our work, we use natural and unrestricted speech signals accompanying demonstrations to understand how human teachers convey information via audio.

III. Methodology

To develop an understanding of complex audio signals from human teachers demonstrating manipulation tasks to an embodied robot, we aim to understand how three different acoustic properties (duration, prosody, semantic content) are used by human teachers across demonstration modalities, audio-usage instructions, relevant subtask segments, and errors during demonstrations. To analyze this information, we conducted a user study (Sec. III-C) where human teachers demonstrate two multi-step manipulation tasks (Sec. III-A) to a Sawyer robot. We recorded audio data along with visual state observations for two demonstration modalities — kinesthetic and video (Sec. III-B). To characterize audio utterances of human teachers accurately, we collected annotations for when teachers talk and what they conveyed through it. To understand how utterances relate to task segments and errors during demonstrations, we also collected annotations for (1) subtask segments, (2) demonstration errors, and (3) characterizations of human speech accompanying subtasks and errors. Details about our annotation procedure are explained in Sec III-D. Finally, in light of recent LfD methods which (1) highlight that learning a separate policy for each subtask in a multi-step task is more efficient [14] and (2) utilize suboptimal demonstrations [15], we analyze if human audio contains beneficial information for two related learning problems. Through proof-of-concept experiments, we demonstrate that random forest classifiers can detect subtasks & errors during demonstrations (Sec. III-F) with a varied set of human acoustic features (Sec. III-E).

A. Task Descriptions

The two household tasks relevant to personal robots that were used in our study are: (a) box opening and (b) fruit cutting (Fig. 2). Each task consists of multiple subtasks performed in sequence to accomplish the overall goal. There could be pauses and errors during and between the execution of the subtasks. Details of the two tasks are as follows:

1) Box Opening: A transparent box contains a green wooden object inside it. A circular button located in the center of the lid (covered with black paper) is pressed. This unlocks the lid from the box. The lifted circular part of the lid is then grasped to open the box. Then, the green wooden object is picked from inside the box and placed on the table. The lid is placed on top of the box. The circular button of the lid is pressed again to lock the box. Finally, the green object is picked from the table and placed on top of the black circular button of the lid. The goal is to perform the task without moving the box from its location on the table surface. This is a multi-step pick-and-place task with a challenging maneuver of lifting an object from inside the box. This task is representative of several goal-oriented household tasks.

2) Fruit Cutting: A flat surface contains different colored wooden fruit pieces on it, along with an uncut wooden banana whose three pieces are stuck together with velcro. A wooden knife is placed in a knife holder. The gripper of the knife is covered with foam so it can be easily grasped by the robot. A slit with a black marking indicates where the robot's gripper can firmly grasp the knife. The knife is taken out of the knife holder, and then used to cut the banana into three pieces by making two cuts through the two velcro attachments. Finally, the knife is placed back into the knife holder. The goal is to cut the banana without moving any other fruit on the wooden surface such that the cut pieces also do not fall off. This is a task which requires complex maneuvers to successfully cut the fruit pieces. This is representative of complex manipulation tasks where the trajectory path and forces applied can matter for successfully achieving the goal.

TABLE I: Ordered Subtask Categories of Box Opening Task

Index	Subtask Category
1	Move arm towards the lid on the closed box
2	Click button to open the lid of the box
3	Grasp button on the lid of the box
4	Transport the lid from the box towards table
5	Release the button of the lid on the table from the gripper
6	Move arm towards the green object that's inside the box
7	Grasp the green object
8	Transport the green object from inside the box towards the table
9	Release the green object from gripper
10	Move arm towards the lid on the table
11	Grasp the button of the lid placed on the table
12	Transport the lid from the table towards the box's body
13	Release the button of the lid on top of the box from the gripper
14	Push the button to lock lid
15	Move arm towards green object
16	Grasp green object on the table
17	Transport the green object towards lid button
18	Release the green object on top of the lid from gripper

TABLE II: Ordered Subtask Categories of Fruit Cutting Task

Index	Subtask Category
1	Move arm towards the knife
2	Grasp the knife from inside the knife holder
3	Move the knife out from knife holder
4	Transport knife from the knife holder towards the banana
5	Cut the first piece of the banana
6	Cut the second piece of the banana
7	Transport knife back from the cutting board to the knife holder
8	Place the knife back in knife holder
9	Release the knife from the gripper

B. Robot System

Demonstrations were provided to a 7 degree-of-freedom Sawyer manipulator with series elastic actuators and a parallel gripper. We used a Movo MC1000 Conference USB microphone to record audio during demonstrations. All participants were told that the robot can watch and listen to them. We used three static camera sensors (facing the human demonstrator and workspace) to collect high resolution video data at ~ 15 Hz for observing the demonstration. The recorded visual and audio data were synchronized and collected using ROS [16].

C. User Study

1) Independent Variables (IVs): We evaluated the effects of two independent variables (demonstration type and instruction type) on different acoustic features. For demonstration type (within-users), we focus on two modalities of LfD for robot manipulation [14]: (1) learning via kinesthetic teaching (KT) in which the joints of a robot are moved along a trajectory in order to accomplish the task [17], and (2) learning from observation, specifically video demonstrations, in which a robot passively observes a human performing the task. For instruction type (between-users), we focused on the narration and natural conditions. Half the participants were part of the narration condition, in which they were explicitly asked to use speech to communicate their intentions as they demonstrate. The remaining half of the participants were in the natural condition, in which they were only told that the robot can listen to them but not asked to use speech specifically. To analyze data for subtasks and errors during demonstrations, an IV indicating their presence or absence is used (example of presence/absence shown in Fig. 3).

2) Data Collection Procedure: We collected demonstration data from 20 participants (8 males, 12 females) for each of the two tasks (Fig. 2). Participants were graduate or undergraduate students, recruited from a university campus. All participants used English speech and were proficient at speaking the language. Each participant was allowed one practice round for each demonstration type on the task they were assigned to do first. After one round of practicing, participants completed 2 demonstrations (1 KT, 1 video) for the box opening task and 2 demonstrations (1 KT, 1 video) for the fruit cutting task. The order of tasks and demonstration types were counterbalanced across all participants. We discarded data from two participants due to network issues during recording and perform our analysis on the remaining 18 participants. This amounted to a total of ∼24 minutes of video demonstration data and ∼119 minutes of KT demonstration data.

3) Dependent Measures (DMs): Our dependent measures are the (1) duration of utterances during demonstrations (also measured as speech density, i.e. fraction of a demonstration accompanied by speech), (2) acoustic features capturing the prosody of speech, and (3) richness of the content conveyed by utterances. Richness of the content refers to the diversity of the concepts conveyed via the spoken words.

4) Research Hypotheses: Our research hypotheses are:

- **H1**: Prior work has shown that humans rely more on additional channels of communicating intent, such as audio, during challenging tasks [1]. Based on this finding and given the complexity of our manipulation tasks, we hypothesize that teachers would naturally rely on using equivalent utterances (in terms of speech density and expressiveness) to the condition when they are explicitly told to do so (narration).
- **H2**: Following the previous motivation from [1]—since kinesthetic demonstrations require more physical effort from the human teacher compared to video demonstrations [18], we hypothesize that more pronounced acoustic features and higher density of speech would be present during kinesthetic demonstrations.
- **H3**: Prior work which studied unrestricted audio signals in human tutoring from parents to children [19], established that human speech from the demonstrator binds to action events, with structured pauses between events. Given these findings, we hypothesize that human teachers' speech would be more pronounced in terms of density and prosodic features *during* relevant subtask executions in a demonstration (presence), compared to periods of gaps between subtasks (absence).
- **H4**: Human utterances have been shown to be beneficial for improving the error prediction of ASR systems [20], with more emphatic uses of speech when errors occur. Thus, we hypothesize that human teachers' speech would be more expressive and emphatic during presence of error segments versus in their absence.

D. Data Annotation

Automated speech detection and recognition is an active and challenging research area [3]. Average word error rates on our dataset (computed using the publicly available Google Speech to Text API [21]) are 0.37 (video demos) and 0.59 (KT demos). To accurately characterize human speech, subtasks, and errors during demonstrations, we collected detailed annotations from three different human annotators. All annotators were undergraduate students recruited at a university campus. The following annotations were collected for each utterance (separated by a significant pause, as judged by the annotators.): start time, stop time, and speech transcription. Each annotation was provided via a GUI interface[1] with information from audio and all three camera views available to be played back at any time.

TABLE III: Unordered Error Categories of Box Opening and Fruit Cutting Tasks

Index	Error Category
1	Teacher forgets to perform a step of the task
2	Teacher struggles to move the robot's arm in a certain way (KT only)
3	Teacher struggles to grasp an object
4	Robot/Human arm collides with, knocks off or moves items from their original position
5	Teacher performs a step of the task in the wrong order
6	Teacher unintentionally uses two hands instead of one to perform the task (video only)
7	Transport knife back from the cutting board to the knife holder
8	Teacher accidentally drops an object already grasped by the robot's gripper (KT only)
9	Teacher intentionally re-strategizes or re-attempts a step of the task
10	Unsuccessful step execution e.g. not applying enough force with the knife to completely cut the banana
11	Other

Annotators also separately marked the start and end times of predefined error and subtask categories for each demonstration (Fig. 3), along with labels of predefined speech categories present with each identified subtask and error. The predefined subtask categories for the box opening task are listed in Table I and for the fruit cutting task are listed in Table II. Each subtask category represents a primitive step that a robot can learn a policy for. The predefined error categories for the user for both tasks are listed in Table III. The predefined speech categories that were labeled along with subtask and error instances were: (1) frustration, (2) encouragement, (3) speech pauses (explicit pause or filler words between two speech utterances), (4) laughter, (5) surprise, (6) no variation in manner of speech (normal). All predefined categories were determined following the methodology of grounded theory analysis [22].

E. Acoustic Feature Computation

We compute several acoustic features over the annotated utterances to characterize prosody and content of human speech. We first processed raw audio signals of human demonstrators with a speech enhancement model [23] to filter out the effect of environmental object-interaction sounds and robot motor noises. We then compute 10 hand-crafted prosodic features (Audio II) to capture prosody and affect in terms of variation in the loudness or speed with which a speaker can alter their pronouncement of an utterance [20], [24]— maximum pitch, energy, and loudness; mean pitch, energy, and loudness; total energy; pause prior to current speech act; total word count and word rate. Pitch, loudness, and energy features represent measures of intonation and

TABLE IV: Percentage of a demonstration duration accompanied by utterances (speech density), subtasks, and errors. Aggregate statistics (mean and standard error) are presented for each of the four subsets within two IVs.

Demo Type/ Instruction Type	%Utterances	%Subtasks	%Errors
Video/Natural	57.97 $\pm$14.12	79.56 $\pm$1.80	9.09 $\pm$3.27
Video/Narration	89.89 $\pm$3.26	90.67 $\pm$2.28	4.84 $\pm$1.87
KT/Natural	27.90 $\pm$11.86	91.43 $\pm$2.41	20.01 $\pm$3.87
KT/Narration	71.60 $\pm$6.09	85.33 $\pm$2.95	24.57 $\pm$2.74

TABLE V: Means and standard errors of speech density (% of demonstration accompanied by speech utterances of a specific type) for annotated acoustic features (Audio I).

	Frustration	Surprise	Speech Pauses	Normal Speech	Laughter	Encourage -ment
KT	0.44$\pm$0.44	2.95$\pm$1.51	8.2$\pm$2.91	62.94$\pm$8.09	0.59$\pm$0.59	1.89$\pm$1.49
Video	0$\pm$0	0$\pm$0	3.66$\pm$3.03	63.29$\pm$7.57	0$\pm$0	0$\pm$0
F(1,16)	1	4.63	1.71	0.003	1	1.67
p	0.33	**<0.05**	0.21	0.96	0.33	0.21
Narration	0.44$\pm$0.44	2.95$\pm$1.51	2.93$\pm$1.83	81.46$\pm$4.24	0.59$\pm$0.59	1.89$\pm$1.49
Natural	0$\pm$0	0$\pm$0	8.92$\pm$3.72	44.77$\pm$8.08	0$\pm$0	0$\pm$0
F(1,16)	1	4.63	1.57	9.95	1	1.67
p	0.33	**<0.05**	0.23	**<0.01**	0.33	0.21

enunciation, whereas duration, pause, word count, word rate capture timing information. These hand-crafted acoustic features have been shown to enhance semantic parsing [24], understand speech recognition failures in dialogue systems [20], and widely used for applications in human-robot interaction [6], [7] and speech recognition [3] communities.

We also analyze annotated prosodic features (Audio I) in the form of emotion labels present in speech (Sec. III-D). In addition to the annotated (Audio I) and hand-crafted acoustic features (Audio II), we also compute 256-dimensional deep acoustic features (PASE [25]) for our learning experiments (Sec. III-F). PASE features are obtained from the last layer of a problem-agnostic speech encoder, trained in a self-supervised manner, and beneficial for downstream tasks leveraging speech features. To understand the richness of semantic content conveyed via audio utterances, we used variance of GloVe embeddings [26] (first principal component from PCA analysis) for words in the annotated speech transcriptions. GloVe embeddings are word representations that capture fine-grained semantic and syntactic regularities of words using vector arithmetic. The variance of such embeddings can quantify the variety of concepts communicated.

F. Learning Classifiers for Subtask and Error Detection

To further understand if different acoustic features in human speech can identify subtasks and errors during demonstrations, we use random forest classifiers for two binary classification tasks: (1) subtask detection and (2) error detection. We use 100 trees for each experiment. From the human facing camera, we also sample every 16^{th} frame to compute video features (penultimate layer output of I3D [27] pretrained on the Kinetics activity recognition dataset [28]) and corresponding audio features (Sec. III-E) for a 1 second window around this frame. This provides about 2000 (cutting) and 2500 (box) samples for the kinesthetic demonstrations, and 370 (cutting) and 510 (box) samples for the video demonstrations. Audio features consist of annotated speech labels encoded as one-hot vectors (Audio I) described in Sec. III-D, hand-crafted prosodic features (Audio II) described in Sec. III-E, as well as deep acoustic features (PASE [25]).

IV. Results and Discussion

We tested the reliability across annotators for consistency of timing and content information they labeled. The Interclass Correlation (ICC) test shows high agreement on timing information between the annotators for utterances, segments and mistakes with $ICC(1) = 0.94, F(35, 10) = 41.6, p < 0.001$ (utterances); $ICC(1) = 0.82, F(971, 971) = 10.1, p < 0.001$ (segments); $ICC(1) = 0.89, F(35, 33) = 18.4, p < 0.001$ (mistakes). Annotations for the error categories, speech categories, and subtask categories also have good reliability with Cohen's $\kappa > 0.72$. Due to the high agreement among annotators, we used data from a single annotator to present the results going forward.

We report various findings about utterances (Sec. IV-A), subtasks and errors (Sec. IV-B), and along the way address if our hypotheses are supported by the results. Unless otherwise noted, a statistical model based on the 2 x 2 mixed design with instruction type as the between-subjects factor and demonstration type as the within-subjects factor was used in the analyses of variance (ANOVA). We also report findings from our learning experiments in Sec. IV-C.

*A. Human Audio Analysis during Demonstrations (**H1**, **H2**)*

1) Quantification of Utterances: The total duration of the demonstrations are 4069.37 sec (Narration/KT), 3080.34 sec (Natural/KT), 893.95 sec (Narration/Video), 521.86 (Natural/Video). The total human audio duration are 2910.15 sec (Narration/KT), 885.30 sec (Natural/KT), 811.11 sec (Narration/Video), 306.56 sec (Natural/Video). The ANOVA results reveal that the percent duration of a demonstration accompanied by utterances (speech density) has a significant main effect along both IVs: instruction type ($F(1, 16) = 8.41, p < 0.05$) and demonstration type ($F(1, 16) = 10.86, p < 0.01$), as shown in column 2 of Table IV. Speech density is significantly more in the narration instruction conditions ($M = 76.02, SD = 16.49$) in comparison to the natural instruction conditions ($M = 37.35, SD = 39.70$). Thus, the results for speech density do not provide support for H1. Speech density is significantly more during video demonstrations ($M = 64.76, SD = 36.81$) compared to KT demonstrations ($M = 48.61, SD = 33.87$). This result can be explained by the fact that there are longer pauses between utterances (normalized by demonstration duration) during KT compared to video demonstrations ($F(1, 16) = 26.89, p < 0.01$). Often users struggle with moving the robot arm in the right configuration during KT demos, and focus on executing subtasks instead of simultaneously talking during such periods. Thus, speech density does not provide support for H2.

TABLE VI: Means and standard errors of speech duration (sec), hand-crafted acoustic features (Audio II), GloVe embeddings.

	Speech Duration	Mean Pitch	Max Pitch	Mean Energy(1E-4)	Max Energy	Total Energy	Mean Loudness(1E-4)	Max Loudness	Word Count	Word Rate	Pause (norm)	GloVe (PC Variance)
KT	210.9±0.08	120.18±14.61	264.48±41.56	2.92±0.58	0.21±0.07	983.52±305.05	56.50±9.06	0.37±0.07	34.40±9.10	1.1±0.16	0.25±0.05	2.85 ±0.39
Video	62.41±0.09	121.18±17.9	227.33±39.73	3.65±0.69	0.08±0.02	374.5±138.83	73.09±11.45	0.24±0.04	15.28±4.76	1.75±0.25	0.06±0.02	2.77±0.39
F (1,16)	23.94	0.027	3.75	6.91	5.4	7.15	19.39	8.62	8.52	45.21	26.89	0.28
p	**<0.01**	0.87	0.07	**<0.05**	**<0.05**	**<0.05**	**<0.01**	**<0.01**	**<0.05**	**<0.01**	**<0.01**	0.60
Narration	206.86±49.93	145.16±8.61	312.25±32.93	3.95±0.55	0.23±0.06	941.22±281.56	77.69±7.00	0.43±0.06	35.67±8.76	1.73±0.15	0.19±0.04	3.49±0.1
Natural	66.45±23.4	96.21±19.73	179.56±41.76	2.62±0.69	0.06±0.02	416.8±189.4	51.91±12.36	0.19±0.04	14.01±5.09	1.11±0.26	0.12±0.05	2.13±0.49
F (1,16)	5.31	2.54	3.18	1.2	7.81	1.65	1.73	9.87	3.23	2.69	1.69	3.62
p	**<0.05**	0.13	0.09	0.29	0.01	0.22	0.21	**<0.01**	0.09	0.12	0.21	0.08

2) Speech Prosody: The annotated (Audio I) and hand-crafted (Audio II) prosodic feature values computed across both IVs are listed in Table V and columns $3 - 12$ of Table VI respectively. The feature values are accumulated for utterances in a single demonstration and these accumulated values are averaged across demonstrations. First, we observe that most demonstrations are accompanied with normal speech and speech pauses (Table V). Surprise is conveyed significantly more during KT demos ($F(1,16) = 4.63, p < 0.05$) and during narration instructions ($F(1,16) = 4.63, p < 0.05$). Since, more errors occur during KT demos and narrations (Table IV, Sec. IV-B.2), the presence of surprise also indicated most errors are unintentional as the users navigated maneuvering of the robot arm (getting stuck in singular configurations) and grasp on the light-weight gripper (objects falling from the grasp of the gripper). This finding does not support H1 and partially supports H2.

For hand-crafted annotated features (Table VI), we find only max loudness to be significantly higher for the narration versus natural condition ($F(1,16) = 9.87, p < 0.01$). The other 9 features are not significantly different across instruction types, thus providing partial support for H1. Across demonstration types, 8 out of 10 features are significantly different. With KT comprising of more errors and pauses, the behavior and speech patterns of teachers are different compared to video demos, providing partial support for H2.

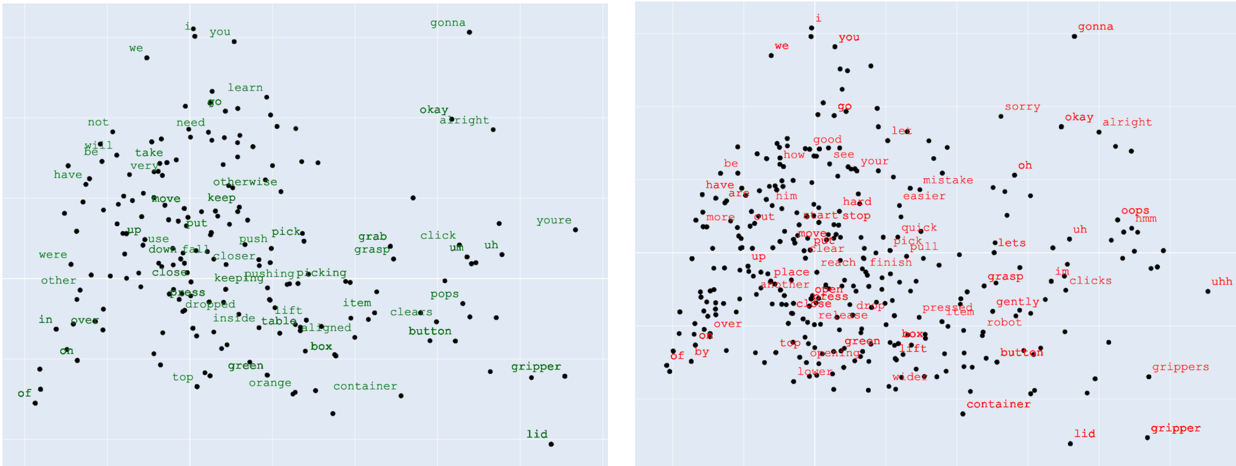
3) Information Conveyed via Spoken Words: We analyzed the word distributions used in teachers' utterances via variance of GloVe vectors across the first principal component (Table VI). The semantic concepts conveyed via spoken words are not significantly different for either demonstration type ($F(1,16) = 0.28, p = 0.60$) or instruction type ($F(1,16) = 3.62, p = 0.08$). PCA projections of GloVe word embeddings for the box task are shown in Fig. 4, with a similar trend observed for the fruit cutting task. The overall word count for KT is higher than videos ($F(1,16) = 8.52, p < 0.05$) with a lot more paraphrasing, prepositions, gerunds, and noun modifiers, adn words specific to robot parts (such as 'closing', 'closed', 'kind of', 'picking', 'picked', 'grip', 'gripper','grasp','keyframe' etc.) compared to video demos.

*B. Human Audio Analysis in relation with Subtasks and Errors during Demonstrations (**H3, H4**)*

The percentage of demonstrations that consist of subtask segments and errors for both IVs are shown in the last 2 columns of Table IV. The percentage for subtasks is significantly different across demonstration type ($F(1,16) = 8.54, p < 0.01$) but not across instruction type ($F(1,16) = 0.31, p = 0.59$). The former effect can be explained by the presence of more errors, pauses, and gaps during KT demonstrations (Sec. IV-A.2). The latter effect indicates that the execution of subtasks does not vary based on the instruction for using audio. The percentage of errors is very low for video demonstrations, resulting in a negligible sample size of errors. Hence, we didn't analyze errors for video demonstrations any further. For kinesthetic demonstrations, the percentage of errors are not significantly different across instruction types ($F(1,16) = 0.6, p = 0.45$). For further analysis, we perform 1-way ANOVA analyses using four subsets of our data independently. The four subsets are generated using the values of demonstration type and instruction type. A single IV for this setting is a binary category depicting the presence or absence of subtasks/errors. The DMs are same as before (Sec. III-C.3).

1) Quantification of Utterances: For each of the four subsets of data, we observe that the duration and density of speech is higher during absence of errors (KT/Narration: $F(1,8) = 124.57, p < 0.01$); KT/Natural: $F(1,8) = 6.48, p < 0.05$). Talking more during absence of errors can be explained by the fact that majority of the demonstration data does not comprise of errors (Table IV). A similar trend of longer duration and higher density of speech is observed during absence of subtasks (KT/Narration: $F(1,8) = 9.88, p < 0.05$; KT/Natural: $F(1,8) = 2.08, p = 0.19$; Video/Narration: $F(1,8) = 10.21, p < 0.05$; Video/Natural: $F(1,8) = 3.97, p = 0.08$). Teachers talking more during gaps between subtask executions does not provide support for H3 and is similar to findings of Nagai et al. [8], where demonstrators use infant-directed speech and explain the subtask to be performed before they begin the execution.

2) Speech Prosody: For analysis of subtasks, we observe that 3 acoustic features are higher during subtask presence for KT/narration: maximum pitch ($F(1,8) = 8.25, p < 0.05$), maximum energy ($F(1,8) = 5.94, p < 0.05$), maximum loudness ($F(1,8) = 4.13, p = 0.07$). The other Audio I and Audio II acoustic features are higher during subtask absence. Similarly, we find that the following features have higher values *during* subtasks— KT/Natural: maximum pitch ($F(1,8) = 2.27, p = 0.17$), maximum energy ($F(1,8) = 1.68, p = 0.23$), maximum loudness ($F(1,8) = 2.03, p = 0.19$), word rate ($F(1,8) = 0.58, p = 0.47$), and normal annotated speech ($F(1,8) = 1.52, p = 0.25$); Video/Narration: mean pitch ($F(1,8) = 1.86, p = 0.21$), maximum pitch ($F(1,8) = 1.12, p = 0.32$), mean energy ($F(1,8) = 0.46, p = 0.52$), maximum energy ($F(1,8) = 7.07, p < 0.05$), mean loudness ($F(1,8) = 1.38, p = 0.27$), maximum



(a) Video Demonstration (Box Opening Task) (b) Kinesthetic Demonstration (Box Opening Task)

Fig. 4: PCA projections of GloVe word embeddings for speech accompanying demonstrations of the box opening task from all participants. Human teachers used semantically similar words for both kinesthetic and video demonstrations, though a larger vocabulary of words was used during the longer and more challenging kinesthetic demonstrations.

loudness ($F(1,8) = 1.71, p = 0.23$), word rate ($F(1,8) = 0.23, p = 0.65$); Video/Natural: maximum pitch ($F(1,8) = 1.07, p = 0.33$), mean energy ($F(1,8) = 0.31, p = 0.59$), maximum energy ($F(1,8) = 0.62, p = 0.45$), mean loudness ($F(1,8) = 0.90, p = 0.37$), maximum loudness ($F(1,8) = 0.87, p = 0.38$). As a few of these results are statistically significant (more so during the narration condition), they provide partial support for H3. Even if the amount of speech used is higher during the absence of subtask executions, some emphasis in speech with specific sounds (excitement, frustration) or words (such as 'oops!', 'oh no!', 'yes!') is conveyed *during* subtask executions.

For analysis of errors, we observe that in the KT/Narration subset, maximum pitch ($F(1,8) = 8.72, p < 0.05$) and maximum loudness ($F(1,8) = 1.19, p = 0.31$) are higher during errors, whereas in the KT/Natural subset, maximum energy ($F(1,8) = 0.29, p = 0.60$) and maximum loudness ($F(1,8) = 0.02, p = 0.89$) are higher. While only one of these features is significantly higher *during* errors versus in their absence, a general trend of 'maximum' features being higher *during* errors indicates that teachers convey their reactions with emphasis in speech when a mistake happens. This finding provides partial support for H4.

3) Information Conveyed via Spoken Words: We do not observe a significant difference in the variance of GloVe features under any of the data subsets for subtask analysis (KT/Narration: $F(1,8) = 2.79, p = 0.13$; KT/Natural: $F(1,8) = 0.17, p = 0.69$; Video/Narration: $F(1,8) = 1.45, p = 0.26$; Video/Natural: $F(1,8) = 0.58, p = 0.47$) or error analysis (KT/Narration: $F(1,8) = 4.17, p = 0.07$, KT/Natural $F(1,8) = 0.10, p = 0.76$). This implies that the concepts being communicated in the presence or absence of subtasks and errors are quite similar, while a few words are specifically spoken under certain conditions (such as the use of 'oops!', 'oh oh' etc. during errors).

TABLE VII: Avg. F1 scores of Subtask Detection and Error Detection for users in the 'Narration' instruction condition.

Features	Subtask Detection				Error Detection	
	Box Video	Box KT	Cutting Video	Cutting KT	Box KT	Cutting KT
Random	0.44	0.4	0.44	0.43	0.43	0.47
Constant	0.46	0.47	0.46	0.46	0.46	0.42
Audio I	0.86	0.84	0.76	0.78	0.89	0.8
Audio II	0.71	0.6	0.72	0.7	0.56	0.55
Audio (I+II)	**0.91**	**0.87**	**0.79**	**0.89**	**0.89**	**0.82**
PASE	0.59	0.49	0.71	0.5	0.46	0.46
Video	0.82	0.8	0.66	0.74	0.72	0.67
Audio I+Video	0.78	0.84	0.77	0.8	**0.9**	**0.84**
Audio II+Video	0.8	0.82	0.53	0.77	0.69	0.66
Audio (I+II)+Video	0.82	0.83	0.76	0.78	**0.9**	**0.84**

C. Learning Experiments

From results of Sec. IV-B.1 and Sec. IV-B.2, we observe that the narration condition has more significant results for acoustic feature differences in relation to subtasks and errors. Thus, we use acoustic features from the narration condition to evaluate performance on two binary classification tasks: subtask detection and error detection. Random $80\% - 20\%$ splits are used to respectively train and test random forest classifiers. We use a combination of acoustic and video features as input to the model. Two baseline models (random prediction and a constant class prediction) are also evaluated. We train three random forests per experiment, and report the average F1 score for best of three in Table VII. Our results show that acoustic features perform better than random and constant class prediction baselines. Combining both annotated (Audio I) and hand-crafted (Audio II) acoustic features is better than using either alone for both subtask detection and error detection. Pretrained PASE features (no finetuning) are unable to match the performance of Audio I and Audio II features, performing only slightly better than the baselines. In addition, pretrained video features (no finetuning) are also not as effective alone as the acoustic features for either detection task. However, combining acoustic and video features performs better at error detection than using acoustic features alone. For subtask detection, we find that concatenating the annotated and hand-crafted acoustic features gives the best performance. These results highlight that acoustic features contain rich information about demonstration quality and task segments, with potential to augment other LfD approaches.

V. Conclusion

Our work highlights several characteristics of audio signals exhibited by human teachers providing demonstrations for multi-step manipulation tasks. Our findings indicate that human audio cues carry rich information, potentially beneficial for further technological advancement in robot learning. While several human demonstration datasets leverage environment and object sounds [29], [30] often human audio of demonstrators is not recorded. Since speech data can be recorded easily with light-weight and cheap sensors, we propose that collection of human audio data, as part of future demonstration datasets, can be beneficial for algorithm development. Integrating the information from environmental and object sounds with human audio is a topic for future work. Furthermore, several recent LfD approaches use suboptimal demonstrations [31], [32] as input. Real-world demonstration data can contain partial errors and gaps in between execution steps of a task. Instead of discarding suboptimal data, leveraging an additional modality like audio can provide additional information about errors and subtasks to aid learning. We take the key first step in leveraging human audio for robot learning—understanding the information present in speech and highlighting that it is possible for automated methods to extract it.

Acknowledgments

The authors would like to thank Dr. Greg Trafton for helpful discussions about data analysis, Prof. Maya Cakmak for helpful discussions about prior work, and Prof. Elaine Schaertl Short for providing valuable feedback on the draft. This work has partly taken place in the Personal Autonomous Robotics Lab (PeARL) at The University of Texas at Austin. PeARL research is supported in part by the NSF (IIS-1724157, IIS-1749204, IIS-1925082), AFOSR (FA9550-20-1-0077), ARO (78372-CS), DFG (448648559). This research was also sponsored by the Army Research Office under Cooperative Agreement Number W911NF-19-2-0333. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

[1] S. Oviatt, R. Coulston, and R. Lunsford, "When do we interact multimodally? Cognitive load and multimodal communication patterns," in *Proceedings of the 6th International Conference on Multimodal Interfaces (ICMI)*, 2004, pp. 129–136.

[2] B. D. Argall, S. Chernova, M. Veloso, and B. Browning, "A survey of robot learning from demonstration," *Robotics and Autonomous Systems*, vol. 57, no. 5, pp. 469–483, 2009.

[3] D. Yu and L. Deng, *AUTOMATIC SPEECH RECOGNITION.* Springer, 2016.

[4] H. H. Clark and J. E. F. Tree, "Using uh and um in spontaneous speaking," *Cognition*, vol. 84, no. 1, pp. 73–111, 2002.

[5] B. Scassellati, "Affective prosody recognition for human-robot interaction," in *Microsoft Research's External Research Symposium. Redmond, WA, USA*. Citeseer, 2009.

[6] E. S. Kim, D. Leyzberg, K. M. Tsui, and B. Scassellati, "How people talk when teaching a robot," in *Proceedings of the 4th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2009, pp. 23–30.

[7] E. S. Short, M. L. Chang, and A. Thomaz, "Detecting contingency for hri in open-world environments," in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 2018, pp. 425–433.

[8] Y. Nagai, C. Muhl, and K. J. Rohlfing, "Toward designing a robot that learns actions from parental demonstrations," in *2008 IEEE international conference on robotics and automation*. IEEE, 2008, pp. 3545–3550.

[9] R. Nakamura, K. Miyazawa, H. Ishihara, K. Nishikawa, H. Kikuchi, M. Asada, and R. Mazuka, "Constructing the corpus of infant-directed speech and infant-like robot-directed speech," in *Proceedings of the 3rd International Conference on Human-Agent Interaction*, 2015, pp. 167–169.

[10] M. N. Nicolescu and M. J. Mataric, "Natural methods for robot task learning: Instructive demonstrations, generalization and practice," in *Proceedings of the second International Joint Conference on Autonomous Agents and Multiagent Systems*, 2003, pp. 241–248.

[11] M. Pardowitz, S. Knoop, R. Dillmann, and R. D. Zollner, "Incremental learning of tasks from user demonstrations, past experiences, and vocal comments," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 37, no. 2, pp. 322–332, 2007.

[12] A. C. Tenorio-Gonzalez, E. F. Morales, and L. Villaseñor-Pineda, "Dynamic reward shaping: training a robot by voice," in *Ibero-American Conference on Artificial Intelligence*. Springer, 2010, pp. 483–492.

[13] E. S. Kim and B. Scassellati, "Learning to refine behavior using prosodic feedback," in *2007 IEEE 6th International Conference on Development and Learning*. IEEE, 2007, pp. 205–210.

[14] O. Kroemer, S. Niekum, and G. Konidaris, "A review of robot learning for manipulation: Challenges, representations, and algorithms," *arXiv preprint arXiv:1907.03146*, 2019.

[15] T. Osa, J. Pajarinen, G. Neumann, J. A. Bagnell, P. Abbeel, and J. Peters, "An algorithmic perspective on imitation learning," *arXiv preprint arXiv:1811.06711*, 2018.

[16] M. Quigley, K. Conley, B. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, A. Y. Ng, *et al.*, "Ros: an open-source robot operating system," in *ICRA workshop on open source software*, vol. 3, no. 3.2. Kobe, Japan, 2009, p. 5.

[17] B. Akgun, M. Cakmak, J. W. Yoo, and A. L. Thomaz, "Trajectories and keyframes for kinesthetic teaching: A human-robot interaction perspective," in *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, 2012, pp. 391–398.

[18] K. Fischer, F. Kirstein, L. C. Jensen, N. Krüger, K. Kukliński, M. V. aus der Wieschen, and T. R. Savarimuthu, "A comparison of types of robot control for programming by demonstration," in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2016, pp. 213–220.

[19] L. Schillingmann, B. Wrede, and K. J. Rohlfing, "A computational model of acoustic packaging," *IEEE Transactions on Autonomous Mental Development*, vol. 1, no. 4, pp. 226–237, 2009.

[20] J. Hirschberg, D. Litman, and M. Swerts, "Prosodic and other cues to speech recognition failures," *Speech communication*, vol. 43, no. 1-2, pp. 155–175, 2004.

[21] Google, "Google Cloud Speech-to-Tex," https://cloud.google.com/speech-to-text, 2021.

[22] K. Charmaz and L. L. Belgrave, "Grounded theory," *The Blackwell encyclopedia of sociology*, 2007.

[23] A. Defossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," *arXiv preprint arXiv:2006.12847*, 2020.

[24] T. Tran, S. Toshniwal, M. Bansal, K. Gimpel, K. Livescu, and M. Ostendorf, "Parsing speech: a neural approach to integrating lexical

and acoustic-prosodic information," *arXiv preprint arXiv:1704.07287*, 2017.

[25] S. Pascual, M. Ravanelli, J. Serra, A. Bonafonte, and Y. Bengio, "Learning problem-agnostic speech representations from multiple self-supervised tasks," *arXiv preprint arXiv:1904.03416*, 2019.

[26] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[27] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[28] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[29] K. Zhang, M. Sharma, M. Veloso, and O. Kroemer, "Leveraging multimodal haptic sensory data for robust cutting," in *2019 IEEE-RAS 19th International Conference on Humanoid Robots (Humanoids)*. IEEE, 2019, pp. 409–416.

[30] D. Gandhi, A. Gupta, and L. Pinto, "Swoosh! rattle! thump!-actions that sound," in *Robotics: Science and Systems (RSS)*, 2019.

[31] D. S. Brown, W. Goo, P. Nagarajan, and S. Niekum, "Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations," *arXiv preprint arXiv:1904.06387*, 2019.

[32] Y. Gao, H. Xu, J. Lin, F. Yu, S. Levine, and T. Darrell, "Reinforcement learning from imperfect demonstrations," *arXiv preprint arXiv:1802.05313*, 2018.