Toward a One-interaction Data-driven Guide: Putting co-Speech Gesture Evidence to Work for Ambiguous Route Instructions

Nick DePalma
Facebook AI Research
Pittsburgh, PA

Jesse Smith
Facebook AI Research
Pittsburgh, PA

Sonia Chernova
Georgia Institute of Technology
Atlanta, GA

Jessica Hodgins
Facebook AI Research
Pittsburgh, PA

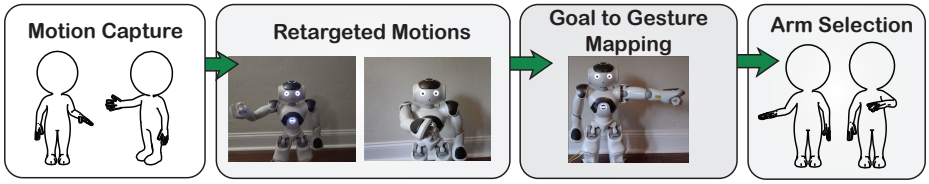

Figure 1: Gesture was retargeted to a robot for evaluation in a navigational instruction setting. We highlight our proposed pipeline of capturing gestures through human example (motion capture), retargeting them to a robot, selecting the appropriate gesture for instruction depending on goal, selecting the best arm and standing formation to maximize understandability. This data-driven framework may allow for usable, readable, direction providing robots that can leverage ambiguous utterances. We show that *each of these decisions have an impact on the usability of the synthesized instructional plan.*

ABSTRACT

While recent work on gesture synthesis in agent[9] and robot literature[24] has treated gesture as co-speech and thus dependent on verbal utterances, we present evidence that gesture may leverage model context (i.e. the navigational task) and is not solely dependent on verbal utterance. This effect is particularly evident within ambiguous verbal utterances. Decoupling this dependency may allow future systems to synthesize *clarifying gestures* that clarify the ambiguous verbal utterance while enabling research in better understanding the semantics of the gesture. We bring together evidence from our own experiences in this domain that allow us to see for the first time what kind of end-to-end concerns models need to be developed to synthesize gesture for one-shot interactions while still preserving user outcomes and allowing for ambiguous utterances by the robot. We discuss these issues within the context of *"cardinal direction gesture plans"* which represent instructions that refer to the actions the human must follow in the future.

HRI '21 Companion, March 8–11, 2021, Boulder, CO, USA.

ACM ISBN 978-1-4503-8290-8/21/03.
https://doi.org/10.1145/3434074.3447223

CCS CONCEPTS

• **General and reference** → **Experimentation**; • **Human-centered computing**; • **Computing methodologies** → **Intelligent agents**;

ACM Reference Format:
Nick DePalma, Jesse Smith, Sonia Chernova, and Jessica Hodgins. 2021. Toward a One-interaction Data-driven Guide: Putting co-Speech Gesture Evidence to Work for Ambiguous Route Instructions. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction (HRI '21 Companion), March 8–11, 2021, Boulder, CO, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3434074.3447223

1 INTRODUCTION

Navigational settings provide a fascinating domain for understanding one-shot co-speech (i.e. verbal and nonverbal) interactions. By one-shot we mean that following the dyadic interaction between the human and robot, the robot does not follow the human. The human then cannot revisit the robot or interaction for clarifications. We focus our efforts on understanding and synthesizing gesture to better understand this exchange. Recently, autonomous systems have begun to leverage the well known fact within the social robotics literature [3, 22] and machine learning literature[14] that referential gestures (sometimes called deixis or deictic referencing) can be used to reference spatial targets like visual objects and social targets. Proper use of gesture allows for reference resolution in the context of ambiguous pronominal usage (e.g. this, that, her, etc). However, much less is known about how conversational gesture may be synthesized and interpreted by a human participant within a navigational context[2, 18]. Navigation settings are particularly interesting to the wider robotics community due to the popularity

of path planning research in social settings[1, 16, 21] and research toward SLAM implementations that provide localization and maps for mobile robots. This *task context* (e.g. map and path plan) has value in gesture generation [7]. Some of these mobile robots that also have arms may also need to interact with people and provide route instructions to a human partner. If the robot has one opportunity to interact with the interlocutor, we do not understand whether a more data-driven co-speech gestural instruction is sufficient for the human observer to achieve their goal.

Imagine walking through an unfamiliar public space, like a mall, seeking a bathroom. You see a robot interacting with people and you ask it for directions to a bathroom. However, the bathroom is upstairs and the robot will not be able to take you up because it is on a wheeled mobile base. The robot may then need to synthesize instructions to get to the location and convey this information in just a single interaction. We define this interaction as a *navigational instruction task*, one in which a participant receives a plan of action from a robot through verbal and nonverbal channels, followed by an attempt by the human interlocutor to achieve the specified goal by executing the plan. Navigational strokes are those gestures that provide clarity of direction during future path planning by the human observer. We define a successful outcome of this interaction as one in which a participant interprets from the robot a plan of action to execute and successfully finds the room of interest.

Present day research in machine learning has focused primarily on synthesizing conversational gesture from speech or utterance patterns. This method of synthesizing gesture is then compared toward a predicted gesture given a speech utterance[5, 6, 17, 18] for success. This approach is in contrast to user based evaluations to determine whether the agent is communicating appropriately. As we will describe within the context of specific tasks like navigational instruction, these gestures will not suffice for interactive robots, but *neither will strictly referential action systems previously explored in face-to-face studies of deictic action.* This literature gap is substantial. We demonstrate that gesture can operate separately alongside an ambiguous utterance to provide strong enough semantic meaning for the interaction partner to understand and leverage toward their own goals.

2 PREVIOUS WORK

We take inspiration from previous work in gesture modeling from psychology, human-robot interaction, and computer animation. Research in gesture modeling can be traced for decades in psychology but has only received attention by computational scientists in the last 20-30 years. However, studies of gesture within the context of navigational instruction are limited to a small subset of relevant literature.

Gesture analysis. David McNeill[15], Adam Kendon[11] and Susan Goldin Meadow[10] provide a comprehensive understanding of gestural use in humans. In particular, they note that gesture is not just a movement of the arm but has functional use within communication. But building a cohesive taxonomy for gesture is still challenging and a significant objective going forward. The topic itself is frequently recategorized and reunderstood within the psychological and cognitive science community. We leverage the vocabulary of psychological literature of gesture analysis to discuss the issues in this domain.

Rule based gesture synthesis. One popular method of synthesizing gesture is to use a rule based system. Our experiments take inspiration from Ravenet[20] and Lhommet's[12] perspective when synthesizing gesture. That is, we synthesize gesture from task rather than utterance which has no precedent in gestural instruction synthesis and evaluation. However, this minority view of synthesizing gesture from task semantics can be contrasted by a more popular approach in the machine learning community: to synthesize gesture from speech utterance. Rule-based pipelines have been shown to properly handle ambiguous speech utterances while providing clarifying gestures. However, these systems are highly designed and are not data driven which makes scaling these speech complementary gestures a challenge given current known approaches to gesture synthesis.

Gesture from speech. Meanwhile, some researchers skip pose extraction and focus only on gesture synthesis from higher quality motion data. Ferstl et. al [8, 9] in particular synthesizes gesture from a speech signal using a generative adversarial network. This is similar to Chiu and Marsella's work that synthesizes gesture from speech signal [6] as well as Yoon et. al.'s work with co-speech synthesis on a robot [24]. However, much of this approach can be traced back to the early work of Cassell that focuses primarily on verbal utterance. MACK [4, 5] was an early implementation of a kiosk that synthesized gesture from text utterance and provided navigational instruction but whose gestures were hand-designed. All of these implementations focus not on readability of the gesture's meaning but on motion prediction when they specify the learning objective and we ask the question whether the speech itself is the right input to synthesize *usable* gesture for the participant.

Evaluating the impact of gesture. Within human-robot interaction, we are similar to Rakita, Mutlu, and Gleicher [19] in that we retarget motion from a motion capture environment and to Sauppé and Multu [22] in that we are interested in understanding interaction effects of the gesture. However, taken together, it's unclear whether Rakita's motion retargeting applied to deictic gestures will result in the same interaction effects that you see in interaction studies. Further, it's also unclear whether subsequent spatial gesturing will allow users to appropriately find the goal room on a map or in person. We argue that previous implementations of co-speech gesture synthesis is not sufficient because they are not specifically grounded to the task context, (e.g. a map and plan of action). One potential related work is that of Bohus, Saw, and Horvitz [2]. However, due to the *in the wild* nature of their experiment, they neither document how the gestures were synthesized nor do they measure whether participants arrived at their goal after instruction. Evaluating the partner's outcome after an interaction is critical because the user's success using these systems is of utmost importance. Finally, Okuno et. al. [18] has obvious similarities to our work because they synthesize the gesture from a underlying path plans. However our experiments differ with their work by testing different hypotheses as well as making the case that gesture can clarify an ambiguous verbal utterance.

Cond.	Has gesture	Verbal Ambig.	Verbal clarity
C1L / C1R	X		
C2		X	
C3	X	X	
C4	X		X

Table 1: Our study focuses on comparing the effects of gestures in isolation and with ambiguous and precise utterances. Condition 1 uses motion for both the left (C1L) and right arm (C1R). We use these conditions to compare the effect of arm choice.

3 STUDY PARAMETERS AND PROCEDURE

To better understand the role of gesture in the presence of various navigation instruction speech utterances and how it impacts user outcomes, we designed a study to explore the role of this type of nonverbal behavior. Through these experiments, we refined our hypotheses to focus on the following: **H1)** navigational gestures clarify ambiguous verbal utterances, **H2)** retargeted natural human motion onto the robot's kinematics for both arms is enough communicate a reference to a room and **H3)** that determining which communicatory intention in route instructions is based on the path plan rather than a reference "through walls".

Existing co-speech synthesis systems (i.e. [24]) leverage verbal utterances to synthesize gesture. **H1** is important because these models lack the ability to synthesize gesture from ambiguous utterances. Past research[7] failed to synthesize gestures that were meaningful enough for use in interaction. **H2** was selected to better understand the impact of arm choice on the success of the interaction partners. We believe this was a significant factor in the success of previous attempts at this problem. Finally, **H3** posits that the understanding semantics of what is being communicated in route instructions is path plans rather than strict deictic referencing.

To further explore our hypothesis, we compare four conditions that we overview in Table 1. Conditions *C1-2* represent the isolated factor of an ambiguous instruction and a gesture. Conditions *C3* and *C4* represent specific co-speech cases. By looking specifically at the effect of gesture and the effect of an ambiguous utterance, we can better characterize the impact of gesture's role in the co-speech utterance.

Procedure. To synthesize the gesture for conditions C1-4, we used CMU data that was collected in [7]. We used this human motion and retargeted the joint rotations into configuration space of a NAO robot. For retargeting, we used the same approach as detailed in [13]. This motion was used for the NAO for playback and we recorded a video for each condition. For the audio track, we aligned the audio by hand and used the NAO TTS APIs.

To evaluate our hypotheses, we recorded a robot performing particular gestures and asked online participants from Mechanical Turk to watch the robotic instruction as a video and respond to a prompt about which room the robot was directing them towards. Participants were allowed to watch the video up to three times. Once they clicked *Next*, a map appeared (pictured in Figure 2) and the user was prompted to select one of the possible rooms. Mechanical Turk was set to only select participants who agreed to our consent form, who are located in the United States (and thus use the same gestural protocol that guided our collection of motions), and who are 18 years of age or older. All participants were paid $5 and there were 25 participants per condition measured.

After the participants observed the motion and selected the goal room, we tested the results for significance using a t-test to determine the effect of gesture in the context of ambiguous utterances, unambiguous utterances, and other effects we detailed above. We present the results in Section 4.

4 RESULTS

Our investigation into factors that impact the success of interpreting the nonverbal plan of action resulted in a number of interesting findings. We found that the arm choice for the communicatory act (Section 4.1), that the underlying path plan may provide the best input data to a gesture synthesis system meant for route instructions (Section 4.2), and whether or not it is also in the presence of an ambiguous utterance have different effects on the interpretation of the co-speech utterance (Section 4.3). Properly deciding an optimal configuration to best communicate the plan is still a significant challenge for interactive systems. For all results, mean value refers to data visualized in Figure 2 which is measured as a fraction of condition participants that correctly chose the goal room intended. For confusion matrices, the values are fraction of participants who answered for a particular room.

4.1 Impact of arm choice (H2)

One significant challenge in using a data driven approach is that even retargeted motion may not be feasible to retarget**??**. In Figure 2, we compared two different conditions (C1R and C1L), one in which the retargeted motion from the right arm gestures a "go left" action($\mu = 0.14$, $\sigma = 0.35$), and one in which the left arm gestures a "go left" with a separate motion from a separate demonstration on the left arm($\mu = 0.66$, $\sigma = 0.48$). We measured success as binary success (1) if user chose room 3 on the normal map and 0 otherwise. We found that these two gestures differ significantly ($p \ll 0.01$) in their outcomes with participants measured by those who picked the correct room after observing the gesture. We also found this to be true with both the normally oriented map($p \approx 0.006$, right arm $\mu = 0.16, \sigma = 0.374$, left arm $\mu = 0.52, \sigma = 0.5$) than with the flipped orientation map($p \ll 0.01$, right arm $\mu = 0.12, \sigma = 0.33$, left arm $\mu = 0.8, \sigma = 0.41$).

We see that the left arm seems to communicate "go left" better than the right by a significant margin which means we must reject the hypothesis. We believe this is due to the feasibility of referencing left with the opposite arm, or a cross-the-chest gesture. *While it may still be possible to retarget a "go left" motion to the right arm with limited feasibility space, we find that this to be an interesting proposal to solve the problem.* For this reason, we will only report the rest of the results using the left arm for the left reference.

4.2 As the crow flies? (H3)

While piloting this study, we asked to direct participants to a target goal room that the robot randomly chooses. In Figure 2, we selected two different options. Do we direct them towards the doorway or

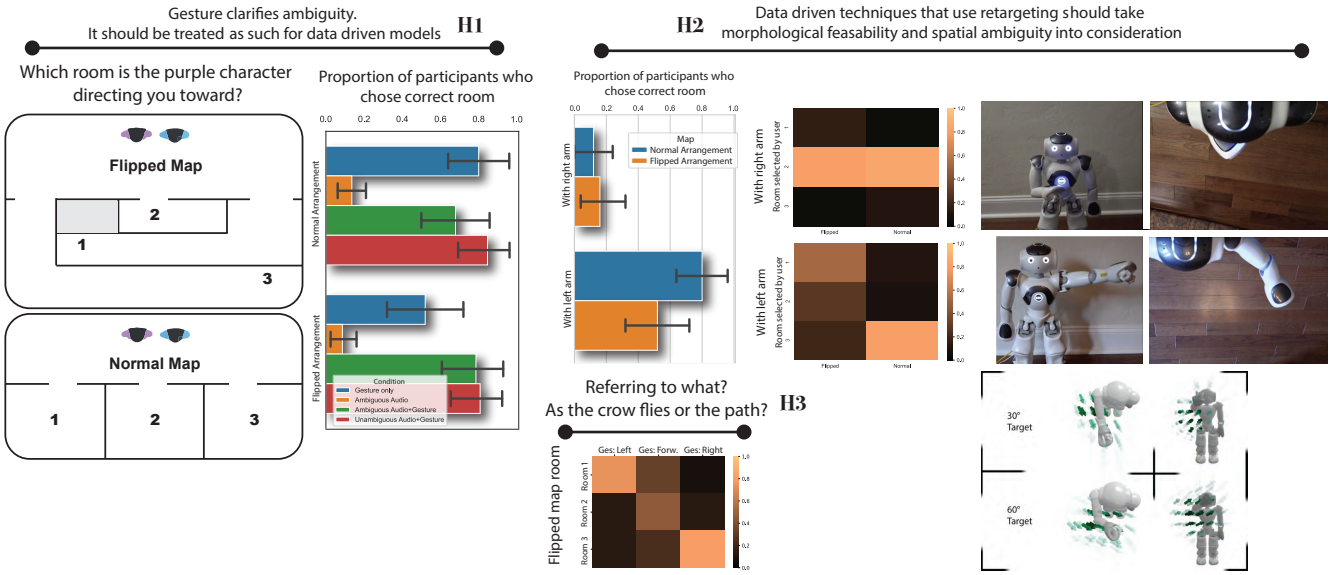

Figure 2: Results of our study from left to right: Left, *H1*) Instructional gesture can supplement ambiguous verbal utterance, Top Middle and Right, *H2*) when using the data from a dataset to retarget onto the robot, feasibility issues present themselves and one potential solution is to understand the communicated action and select a different arm to make the instruction usable, Bottom Middle, *H3*) the most helpful directions seem to be provided ego-centric and according to the path plan rather than general direction of the goal.

the goal location without respect to obstacles, the so called "as the crow flies" direction?

Looking at the confusion matrix in Figure 2, we found that people tended to read the gesture as "go left" as go to door left rather than goal left and we used this result in future experiments.

4.3 Gesture offers something similar to utterance (H1)

For this hypothesis, we wanted to understand if some types of gesture contain usable content that indicates a particular action when presented alongside a verbal utterance. Finding this evidence may enable complimentary gestures during co-speech acts with ambiguous verbal utterances. While some types of gestures like beats may not contain any specific "route instruction," we were interested in understanding whether movements of the robot can communicate a single action if it's isolated from the verbal utterance. Figure 2 explores this question.

From Figure 2, we argue that the stroke information in the gesture should also leverage the task and be *grounded* similar to the challenges in verbal grounding. In the case of navigational gesture, we ground the gesture to the path plan [7, 23], albeit naively for these experiments. Looking more specifically at Figure 2,

We found that our selection of gesture both without the presence of a verbal utterance (*C1L*) as well as in the presence of an ambiguous utterance (*C3*, "Just go this way"), and an unambiguous utterance (*C4* "Just go left," "Just go right") still works sufficiently well in all cases of verbal utterance content. Most interestingly, the same experiment with gesture and the same ambiguous utterance(*C3*) performs significantly better than *C2*, ambiguous utterance alone ($p \ll 0.01$, $\mu_{AA} = 0.11$, $\sigma_{AA} = 0.32$, $\mu_{AA+G} = 0.73$, $\sigma_{AA+G} = 0.45$).

We found that **H1** seems to hold true. Indeed, because gesture can contain some communicatory referencing, it can be used alongside verbal utterances both ambiguous, and clearly stated. This effect holds across both maps (Normal and Flipped). While other research has found similar results [23], we argue that taken with the rest of our results, it's clear that the participants are interpreting the referential action as referring to which path plan to take to a given room. Furthermore, synthesizing these plans would not be possible if ambiguous utterances are to be used. In the case of navigational gesture, it can contain instructions to the user through arm movements, in comparison to other forms of social manipulation like verbal instruction.

5 DISCUSSION

Through our experiments, we found that communicatory gestures can be challenging to interpret. Our contribution is in characterizing a new approach to improving gesture interpretation that suggests 1) robots may need dynamic choice of the gesture arm to perform the gesture, and 2) co-speech gesture interpretation demands gesture synthesis that utilizes task context rather than synthesizing them from speech alone. This suggestion is significantly different from previous approaches to gesture synthesis [5, 17].

Going forward, we intend to extend this work by both verifying these results in an in-person human robot interaction as well as leveraging these simple insights to explore a fully synthesized route instruction system for human understanding in a data driven manner.

REFERENCES

[1] Santosh Balajee Banisetty, Scott Forer, Logan Yliniemi, Monica Nicolescu, and David Feil-Seifer. 2019. Socially-aware navigation: A non-linear multi-objective optimization approach. *arXiv preprint arXiv:1911.04037* (2019).

[2] Dan Bohus, Chit W Saw, and Eric Horvitz. 2014. Directions robot: in-the-wild experiences and lessons learned. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*. 637–644.

[3] Andrew G Brooks and Cynthia Breazeal. 2006. Working with robots and objects: Revisiting deictic reference for achieving spatial common ground. In *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction*. 297–304.

[4] Justine Cassell, Tom Stocky, Tim Bickmore, Yang Gao, Yukiko Nakano, Kimiko Ryokai, Dona Tversky, Catherine Vaucelle, and Hannes Vilhjálmsson. 2002. Mack: Media lab autonomous conversational kiosk. In *Proc. of Imagina*, Vol. 2. 12–15.

[5] Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. 2004. Beat: the behavior expression animation toolkit. In *Life-Like Characters*. Springer, 163–185.

[6] Chung-Cheng Chiu and Stacy Marsella. 2014. Gesture generation with low-dimensional embeddings. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*. 781–788.

[7] Nick Depalma and Jessica Hodgins. 2020. Leveraging knowledge asymmetries to evaluate synthesized gesture based communication in human-robot interaction. In *Workshop in AI & Its Alternatives in Assistive & Collaborative Robotics: Decoding Intent at Robotics: Science and Systems*.

[8] Ylva Ferstl, Michael Neff, and Rachel McDonnell. 2019. Multi-objective adversarial gesture generation. In *Motion, Interaction and Games*. 1–10.

[9] Ylva Ferstl, Michael Neff, and Rachel McDonnell. 2020. Adversarial gesture generation with realistic gesture phasing. *Computers & Graphics* (2020).

[10] Susan Goldin-Meadow. 2005. *Hearing gesture: How our hands help us think*. Harvard University Press.

[11] Adam Kendon. 2004. *Gesture: Visible action as utterance*. Cambridge University Press.

[12] Margot Lhommet and Stacy Marsella. 2014. Metaphoric gestures: towards grounded mental spaces. In *International Conference on Intelligent Virtual Agents*. Springer, 264–274.

[13] Yuwei Liang, Weijie Li, Yue Wang, and Rong Xiong. 2020. Dynamic Movement Primitive based Motion Retargeting for Dual-Arm Sign Language Motions. *arXiv preprint arXiv:2011.03914* (2020).

[14] Cynthia Matuszek, Liefeng Bo, Luke Zettlemoyer, and Dieter Fox. 2014. Learning from Unscripted Deictic Gesture and Language for Human-Robot Interactions. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence* (Québec City, Québec, Canada) *(AAAI'14)*. AAAI Press, 2556–2563.

[15] David McNeill. 1992. *Hand and mind: What gestures reveal about thought*. University of Chicago press.

[16] Ross Mead and Maja J Matarić. 2017. Autonomous human–robot proxemics: socially aware navigation based on interaction potential. *Autonomous Robots* 41, 5 (2017), 1189–1201.

[17] Michael Neff, Michael Kipp, Irene Albrecht, and Hans-Peter Seidel. 2008. Gesture modeling and animation based on a probabilistic re-creation of speaker style. *ACM Transactions on Graphics (TOG)* 27, 1 (2008), 1–24.

[18] Yusuke Okuno, Takayuki Kanda, Michita Imai, Hiroshi Ishiguro, and Norihiro Hagita. 2009. Providing route directions: design of robot's utterance, gesture, and timing. In *2009 4th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 53–60.

[19] Daniel Rakita, Bilge Mutlu, and Michael Gleicher. 2017. A motion retargeting method for effective mimicry-based teleoperation of robot arms. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. 361–370.

[20] Brian Ravenet, Catherine Pelachaud, Chloé Clavel, and Stacy Marsella. 2018. Automating the production of communicative gestures in embodied characters. *Frontiers in psychology* 9 (2018), 1144.

[21] Jorge Rios-Martinez, Anne Spalanzani, and Christian Laugier. 2015. From proxemics theory to socially-aware navigation: A survey. *International Journal of Social Robotics* 7, 2 (2015), 137–153.

[22] Allison Sauppé and Bilge Mutlu. 2014. Robot deictics: How gesture and context shape referential communication. In *2014 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 342–349.

[23] Christopher David Wallbridge, Séverin Lemaignan, Emmanuel Senft, and Tony Belpaeme. 2019. Generating Spatial Referring Expressions in a Social Robot: Dynamic vs Non-Ambiguous. *Frontiers in Robotics and AI* 6 (2019), 67.

[24] Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2019. Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 4303–4309.