Characterizing and Processing Robot-Directed Speech

Paulina Varchavskaia and Paul Fitzpatrick

The Problem: Speech directed at infants and pets has properties that distinguish it from speech among adults. Some of those properties are potentially useful for language learning. By careful design of form and behavior, robots can hope to evoke a similar speech register and take advantage of these properties. This work is directed at examining this claim based on experiments carried out with the infant-like robot Kismet [2], and to develop appropriate procedures for processing robot-directed speech.

Motivation: A natural-language interface is a desirable component of a humanoid robot. In the ideal, it allows for natural hands-free communication with the robot without necessitating any special skills on the human user's part. In practice, we must trade off flexibility of the interface with its robustness. Contemporary speech understanding systems rely on strong domain constraints to achieve high recognition accuracy [6]. We hope to identify constraints available to a suitably engineered robotic "personality", and to make an initial exploration of how speech recognition techniques may be applied to the domain of robot-directed speech with flexibility that matches the expectations raised by the robot's humanoid form.

Previous Work: This work builds on that of Breazeal [2], who constructed Kismet, an "infant-like" robot. We draw on recordings of children interacting with Kismet that were collected by the MIT Initiative on Technology and Self. The work also relies heavily on speech recognition tools and corpora developed by the SLS group at MIT.

Approach: When interacting with a youthful-appearing robot such as Kismet, we can expect that the speech input may have specialized characteristics similar to those of infant-directed speech (IDS). We examined the following two questions regarding the nature of IDS:

- Does it include a substantial proportion of single-word utterances? Presenting words in isolation side-steps the problematic issue of word segmentation.
- How often, if at all, is it clearly enunciated and slowed down compared to normal speech? Over-articulated speech may be helpful to infants, but has important consequences for artificial speech recognizers.

Some preliminary results on this are presented in [5].

For processing robot-directed speech, we currently use a vocal mimicry system triggered by simple keywords which is sufficient to extract vocabulary items from "cooperative" speech. We then build a language model around this initial vocabulary, and bootstrap from it to identify further candidate vocabulary items drawn from arbitrary speech in an unsupervised manner. We cast this process in a form that can be largely implemented using a conventional speech recognition system [3], even though such systems are designed with very different applications in mind. This is advantageous since, after decades of research, such systems are expert at making acoustic judgments in a probabilistically sound way from acoustic, phonological, and language models. Figure 1 shows the overall structure of this process.

Impact: This work attacks the question of language for a humanoid robot from several directions. One direction is concerned with characterizing and influencing the speech register that people use when addressing the robot. Another addresses how to extract vocabulary items from such speech, be it cooperative or otherwise. Other work, not described here, is addressing the crucial issue of binding vocabulary to meaning. One line of research under way is to use transient, task-dependent vocabularies to communicate the temporal structure of processes. Another line of research looks more generally at how a robot can establish a shared basis for communication with humans by learning expressive verbal behaviors as well as acquiring the humans' existing linguistic labels.

Future Work: Parents tend to interpret their children's first utterances very generously and often attribute meaning and intent where there may be none [1]. It has been shown, however, that such a

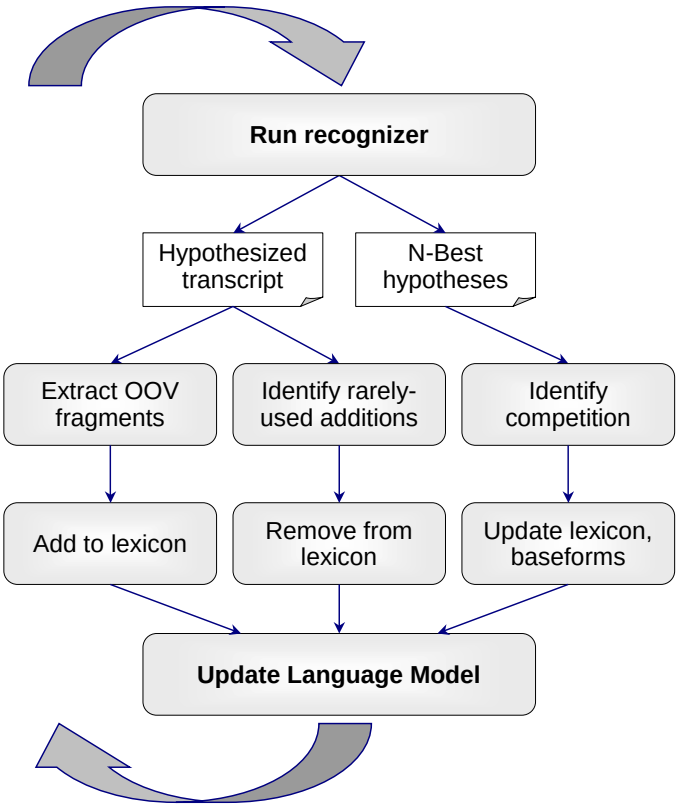

Figure 1: An iterative clustering procedure for identifying candidates for vocabulary extension.

strategy may indeed help infants coordinate meaning and sound and learn to express themselves verbally. Pepperberg [4] formalized the concept into a teaching technique called referential mapping. The strategy is for the teacher to treat the pupil's spontaneous utterances as meaningful, and act upon them. This, it is shown, will encourage the pupil to associate the utterance with the meaning that the teacher originally gave it, so the student will use the same vocalization again in the future to make a similar request or statement. The technique was successfully used in aiding the development of children with special needs. In future work, we hope to apply this technique to build a shared basis for meaningful communication between the human and the robot.

Research Support: Funds for this project were provided by DARPA as part of the "Natural Tasking of Robots Based on Human Interaction Cues" project under contract number DABT 63-00-C-10102, and by the Nippon Telegraph and Telephone Corporation as part of the NTT / MIT Collaboration Agreement.

References:

[1] P. Bloom. *How Children Learn the Meaning of Words.* Cambridge: MIT Press, 2000.

[2] C. Breazeal. *Sociable Machines: Expressive Social Exchange Between Humans and Robots.* PhD thesis, MIT Department of Electrical Engineering and Computer Science, 2000.

[3] J. Glass, J. Chang, and M. McCandless. A probabilistic framework for feature-based speech recognition. In *Proc. International Conference on Spoken Language Processing*, pages 2277–2280, 1996.

[4] I. Pepperberg. Referential mapping: A technique for attaching functional significance to the innovative utterances of an african grey parrot. *Applied Psycholinguistics*, 11:23–44, 1990.

[5] Paulina Varchavskaia and Paul Fitzpatrick. Characterizing and processing robot-directed speech. Submitted to Humanoids 2001, Tokyo, Japan.

[6] V. Zue and J.R. Glass. Conversational interfaces: Advances and challenges. *Proceedings of the IEEE, Special Issue on Spoken Language Processing*, Vol. 88, August 2000.