Recognition of Affective Communicative Intent in Robot-Directed Speech

Cynthia Breazeal & Lijin Aryananda

Artificial Intelligence Laboratory
Massachusetts Institue Of Technology
Cambridge, Massachusetts 02139

http://www.ai.mit.edu

The Problem: Human speech provides a natural and intuitive interface for both communicating with humanoid robots as well as for teaching them. In general, the acoustic pattern of speech contains three kinds of information: who the speaker is, what the speaker says, and how the speaker says it. This work focuses on the question of recognizing affective communicative intent in robot-directed speech.

Motivation: Over the past three years, we have been building infant-level social competencies into our robot, Kismet, so that we may explore social development and socially situated learning between a robot and its human caregiver [1]. This work is aimed towards implementing similar learning mechanisms on Kismet but with an added twist: the ability of the human caregiver to directly modulate the robot's affective state through verbal communication. This provides the human caregiver with a natural and intuitive means for shaping the robot's behavior and for influencing what the robot learns.

Previous Work: There have been a number of vocal emotion recognition systems developed in the past few years [2, 3]. These systems use various acoustic features and different types of learning algorithm to identify the speaker's affective state. However, for the purposes of training a robot, the raw emotional content of the speaker's voice is only part of the message. A few researchers have developed recognition systems that can recognize speaker approval versus speaker dissaproval from child-directed speech [4], or recognize praise, prohibition, and attentional bids from infant-directed speech [5].

Developmental psycholinguists can tell us quite a lot about how preverbal infants recognize affective speech and how caregivers exploit it to regulate the infant's behavior. Infant-directed speech is typically quite exaggerated in the pitch and intensity (often called motherese). Based on a series of cross-linguistic analyses, there appear to be at least four different pitch contours (approval, prohibition, comfort, and attentional bids), each associated with a different communicative intention [6]. Maternal exaggeration in infant-directed speech seems to be particularly well matched to the innate affective responses of human infants.

Approach: Inspired by this work, we have implemented a five-way recognizer that can distinguish Fernald's four prototypical prosodic contours as well as neutral speech. As shown in figure 1, the affective speech recognizer receives robot directed speech as input. The speech signal is analyzed by the low level speech processing system. The next module performs filtering and pre-processing to reduce the amount of noise. The resulting pitch and energy data are then passed through the feature extractor, which calculates a set of features (F_1 to F_n) selected using a sequential forward selection process. The classifier is trained using a set of recordings of two female caregivers. Each class of data is modeled using the Gaussian mixture model, updated with the EM algorithm and a Kurtosis-based approach for dynamically deciding the appropriate number of kernels [7]. Finally, based on the trained model, the classifier determines whether the computed features are derived from an approval, an attentional bid, a prohibition, a soothing, or a neutral utterance.

The output of the vocal affective intent classifier is interfaced with Kismet's emotion subsystem where the information is appraised at an affective level and then used to directly modulate the robot's own affective state. In this way, the affective meaning of the utterance is communicated to the robot through a mechanism similar to the one Fernald suggests. The robot's current "emotive" state is reflected by its facial expression and body posture. This affective response provides critical feedback to the human as to whether or not the robot properly understood their intent.

Difficulty: Naturally occuring robot-directed speech doesn't come in nicely packaged sound bites. Often there is clipping, multiple prosodic contours of different types in long utterances, and other background noise. Again, targetting infant-caregiver interactions goes some ways in alleviating these issues, as infant-directed speech is slower, shorter, and more exaggerated. However, our collection of robot-directed utterances demonstrates a need to address these issues carefully.

Impact: By integrating this perceptual ability into our robot's "emotion" system, we allow humans to directly manipulate the robot's affective state. This has a powerful organizing influence on the robot's behavior, and will ultimately be used to socially communicative affective reinforcement.

Future Work: The recognizer in its current implementation is specific to female speakers, and it is particularly tuned to women who can use motherese effectively. Granted not all people will want to use motherese to instruct their robots. However, at this early state of research we are willing to exploit *naturally occurring* simplifications of robot-directed speech to explore human-style socially situated learning scenarios. Future improvements include either training a male adult model, or making the current model more gender neutral.

To provide the human instructor with greater precision in issuing vocal feedback, we will need to look beyond *how* something is said to *what* is said. It is also a fascinating question of how the robot could *learn* the valence and arousal associated with particular utterances by bootstrapping from the correlation between those phonemic sequences that show particular persistence during each of the four classes of affective intents. Over time, Kismet could associate the utterance "Good robot!" with positive valence and "No, stop that!" with negative valence by grounding it in an affective context and Kismet's emotional system. Developmental psycholinguists posit that human infants learn their first meanings through this kind of affectively-grounded social interaction with caregivers.

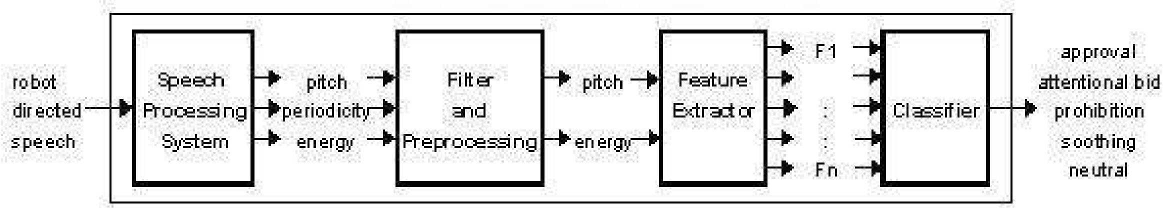

Figure 1: The affective speech recognition system.

Research Support: This research is funded by the DARPA MARS grant under contract number DABT 63-99-1-0012.

References:

[1] C. Breazeal. A Motivational System for Regulation Human-Robot Interaction. *Proceedings of AAAI98*, pp 54-61, 1998.

[2] F. Dellaert, F. Polzin, A. Waibel. Recognizing Emotion in Speech *Proceedings of the ICSLP*, 1996.

[3] R. Nakatsu, J. Nicholson, N. Tosa. Emotion Recognition and Its Application to Computer Agents with Spontaneous Interactive Capabilities. *ICMCS*, Vol 2: pp 804-808, 1999.

[4] D. Roy, A. Pentland. Automatic Spoken Affect Classification and Analysis. *Proceedings of the 2nd International Conference on Automatic Face and Gesture Recognition*, pp 363-367, 1996.

[5] M. Slaney, G. McRoberts. Baby Ears: A Recognition System for Affective Vocalization. *Proceedings of the 1998 International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Seattle, WA, 1998.

[6] G. McRoberts, A. Fernald, L. Moses. An Acoustic Study of Prosodic Form-function Relationships in Infant-directed Speech: Cross Language Similarities. *In press.*

[7] N. Vlassis, A. Likas. A Kurtosis-Based Dynamic Approach to Gaussian Mixture Modelling. *IEEE Trans. on Systems, Man, and Cybernetics (Part A: Systems and Humans)*, Vol. 29: No. 4, 1999.