Feedback and its role in Speech Recognition for Robots: An Architecture and an Experiment

Chris Winsor
Worcester Polytechnic Institute
Professor Sonia Chernova
29-April-2013

Prepared in partial fulfillment of the requirements of the special topic source on Human-Robot Interaction

Abstract

This paper investigates the role of feedback in speech for robots, specifically the role of proprioception in speech recognition. To give perspective to the investigation a set of five types of feedback are reviewed based on developmental studies, and an overall Proprioceptron Architecture is introduced. The Proprioceptron Architecture is compared to Brooks' Subsumption Architecture as well as traditional software approaches. Finally, a robot based on the Proprioceptron Architecture is developed, and experiments performed on what we believe is a new form of computational learning - external proprioception.

1. Introduction

It is expected that speech will be an essential element in human-robot interaction (HRI). The related field of human-computer interaction (HCI) has foreshadowed this with Siri and Windows 8 Speech Recognition as testament to the convenience and naturalness of speech as an interface between humans and machines. But despite extensive research even today's highly situated mobile and desktop applications have challenges with speech. We observe Siri has difficulty with Scottish accents, Windows 8 Speech Recognition has limited context and vocabulary, and text-to-speech recognition applications such as NaturallySpeaking require a headset, noise-free environment and user-specific training to achieve satisfactory recognition rates.

HRI will only further compound these challenges for speech because it is:

- Multi-model: Robots use a variety of modalities and need to integrate this variety of information into speech recognition. Examples include backchannel feedback, conversational context, and evidence from gesturing and other visual cues.

- Social Learning - Robots will want to learn interactively from humans and will be faced with context-specific words and pronunciations, as well as accents and nuances that reflect the local dialect and environment in which they are situated.

- Embedded - robots are expected to be deployed out in the real world which is noisy and ambiguous.

In this paper we explore the role of feedback in speech recognition. Section 2 is a review of prior research from developmental science which relates to feedback in humans, and identifies five types of feedback observed in infants and adults. Section 3 visits Brooks' Subsumption Architecture and evaluates where Subsumption does, and does not, explain the observations from developmental science. Sections 4 and 5 then establishes the Proprioceptron Architecture and demonstrates how this does explain the five feedback paths. In Section 6 we develop Proprioception Machine 1 and experiment with one of these feedback types - external proprioception.

2. Observations from Developmental Science

Imitation - Meltzoff, Decety 2003

Meltzoff and Decety [1] studies the subject of imitation in infants based on evidence from developmental science and neuroscience.

The first observation is that infants at the earliest stages of development (<72 hours of age) imitate the facial expressions of a caregiver. It is suggested there is an innate wiring to support this low level imitation - a "link" between perception and production" where perception is the "visual space" and the output is the "action space". They further suggest that imitation is the developmental "seed" by which the "fruit" of adult theory of mind will later come.

The second observation is that 14-month-old infants demonstrate a high level of interest in an adult that "mimes" the infant's behavior. It is observed the infant will interact significantly longer with a miming caregiver as compared to one performing unrelated behaviors. It is indicated that these infants appear to intentionally explore the mime as if trying to detect anomalies in the response.

The third observation involves 18-month-old infants and their ability to infer intent when they observed the erroneous action of an adult. In this experiment the infant produces a corrected version of the behavior of the adult - one that is inferred by the adult's action, not the erroneous action that the infant actually observes.

The research provides three examples of cross-modal neurological paths in infants - facial imitation, mime exploration and intent inference.

Speech impairment due to deafness

It is well established that deafness in an individual is associated with a loss of fidelity in speech. The most frequently cited study is [2] which identifies a number of phonological anomalies exhibited in speech of deaf individuals. A number of studies have investigated factors such as how the degree of hearing loss contributes to loss in speech fidelity strategies to compensate for an existing hearing condition, and the effect of assistive devices.

The ability to hear one's own speech is a classic example of proprioceptive feedback.

Mental Imagery, Mental Practice

Mental imagery and mental practice [3], also known as known as psychoneuromuscular theory, involve imagining a situation or an activity without actually being there or performing the activity. Athletes use mental practice to prepare for a performance by rehearsing the event in advance. It has been shown that mental practice can statistically improve performance as compared to the non-practice baseline.

Psychoneuromuscular theory suggests the mental pathways used in imagining physical actions are the same as those in actually conducting the activity physically, and that there is a link from production to perception that is used to exercise these paths.

Summary Of Observations from Developmental Science

The five feedback paths are summarized as:

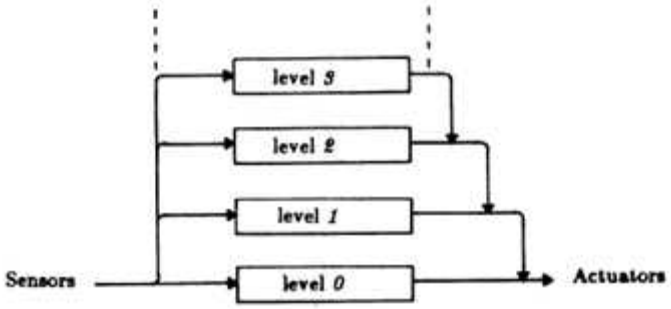
- Facial Imitation in infants
- Mime exploration in infants
- Intent inference in infants
- Proprioceptive feedback in speech (as evidenced by impairment due to deafness)
- Mental practice in adults

3. Brooks Subsumption

In his landmark 1991 paper "Intelligence Without Representation" [4], Rodney Brooks takes a reflective look at AI research and the factors contributing to where it is today. Brooks suggests the current goal of most AI research is to develop a model of the world, then employ that model as part of planning to determine the best course of action. This approach is characterized as sense-model-plan-act (SMPA). But Brooks argues the real world is too complex and unpredictable to be modeled. He believes the majority of what we attribute to "intelligence" is actually creatures residing within, and interacting closely with, the real world. He believes there is no top-level model, rather learning occurs from a somewhat haphazard sequences of events, learned corrections, and prioritization among independent response tasks, and not the result of a top-level plan or model. He introduces the ideas of situatedness, embodiment, intelligence and emergence.

Situatedness states the robot will be situated in its environment. It can expect the world to be around it. The world is available at the robot's disposal to query or interact with. Situatedness significantly lessens (Brooks would say eliminates) the need for modeling. Embodiment means that the robot is physical - grounded with the realities of noise and uncertainty of the real world. Intelligence is measured with respect to the robot's actual demonstrated behavior, not on presupposed software metrics as in traditional approaches. Emergence suggests that software does not need to have a cleanly pre-thought "top-down" structure such as a control loop. Rather a patchwork of separate modules, each independently arrived at, can build themselves into an intelligent system.

An example of what will later be referred to as the "Subsumption Architecture" is shown in the figure at right which is taken from Brooks [5]. Sensor data is received and processed by competence levels working independently and in parallel. The lowest numbered competences are 'subsumed' by the higher numbers (higher numbered competencies assume lower numbers have taken care of the highest priority tasks).



Evaluating Subsumption

We now review the five proprioceptive/feedback loops discussed earlier to see if they are explained by the Subsumption Architecture.

Facial imitation in infant: This form of feedback is supported by Subsumption. In fact Subsumption does a good job of explaining facial imitation by an infant, since within an infant the sensor/stimulus would be received at a very low level (Level 0) and Subsumption would expect a rudimentary response which is what is observed.

Mime exploratory behavior in infant: Subsumption does not explain this behavior. Subsumption does not embrace the concept of internal intent, rather Subsumption is described in reactive terms - the infant is the subject of external stimulus and responds to it. Furthermore Subsumption would suggest that only the

most rudimentary levels are innately present and the more sophisticated exploratory behavior described here would not exist innately so this behavior would be unlikely in a 14 month old infant.

Intent inference in infant: In this case the infant is the recipient of stimulus and infers the intent of the external agent by observing the agent's mistaken action, then demonstrates the corrected action. Subsumption would certainly explain the perception-to-production path. But since Subsumption assumes only the most rudimentary competence levels it is questionable whether Subsumption would suggest the highest level "intent" would have developed in an 18 month old infant.

Speech impairment due to deafness: Subsumption would certainly include proprioceptive feedback as an essential element in its architecture. The situatedness expectation of Subsumption would specifically say that impaired feedback resulting from deafness would significantly affect speech.

Mental Imagery / Mental Practice: Subsumption does not explain Mental Imagery or Mental Practice. First - Subsumption has no concept of intent-driven behavior like that upon which Mental Practice is based. Second - there is no abbreviated path from production to perception which would bypass the final stage of production and perception. Third - the situatedness assumption would suggest that production and perception are closely tied to the real world (the world is the model). The internal-only nature of Mental Imagery and Mental Practice is a strong violation of the situated expectation inherent in Subsumption.

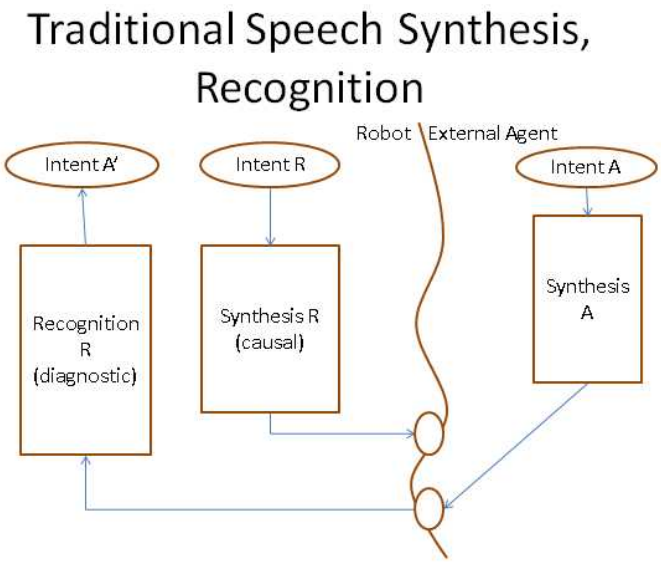
In summary, Subsumption does a good job of explaining feedback loops where external stimulus results in a low level observed response such as the facial imitation or speech impairment due to deafness. It has a difficult time explaining feedback loops where the human's internal intent is the initiator, such as miming. And it specifically does not explain feedback loops where external production is excluded, namely mental imagery.

4. Development of the Proprioceptron Architecture

We now develop, in a step-wise manner, the Proprioceptron Architecture. We start with a baseline review of speech recognition and speech synthesis as it is traditionally deployed.

Baseline: Traditional Speech Synthesis, Recognition

The figure at right illustrates the key elements involved in speech as deployed by traditional software. On the left is the robot and on the right is the external world. The robot system includes synthesis and recognition. Synthesis starts with robot's intent and generates phonetic sound as an output to the world. Recognition receives phonetic sound from the external world and attempts to recover the intent of the agent.



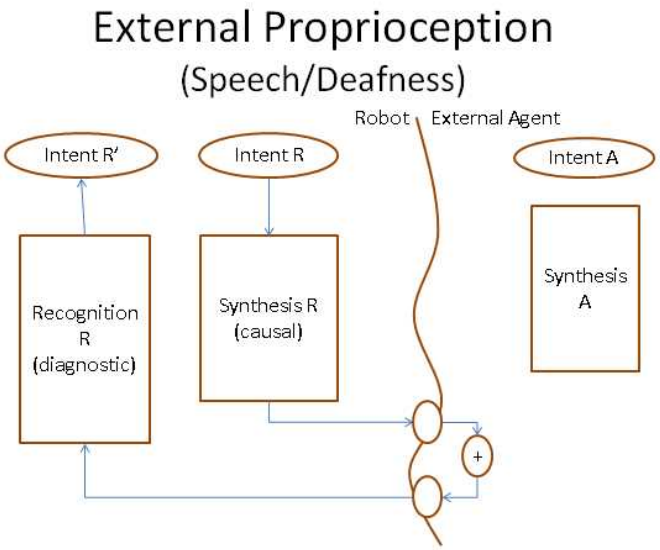
We suggest that synthesis processes (R and A) are causal (deterministic) systems. Given an intent and synthesis there is exactly one phone stream that will result. Recognition is a diagnostic process, necessarily in-deterministic and probabilistic. Given a phone stream there are any number of intents that may have caused it.

Synthesis R performs the same function as Synthesis A in that they each map Intent to phone stream. We believe that for purposes of recognition (recognizing the output of Synthesis A) there is significant utility in the Synthesis R causal model. We know that techniques involving Bayes networks use a causal model - these techniques iteratively guess at possible causes and apply a causal model to measure the likelihood the observed data would have resulted from that cause. So we are intrigued by the possibility that the causal model associated with a robot's speech synthesis might be applied to the task of recognition. And that a proprioceptive feedback loop may be involved in this process.

In summary - the structure provided by traditional software provides no feedback from synthesis to recognition. It provides no accommodation for any of the five feedback loops we have discussed. It is essentially stove-piped with no internal cross-modal data flow.

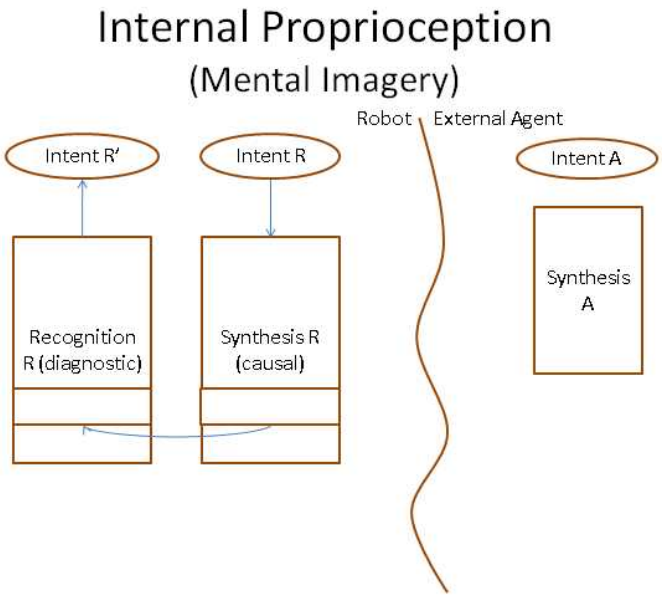
External Proprioception (Speech/Deafness)

The most straightforward type of feedback in speech is classic proprioception - hearing ones-self speak. We will refer to this as External Proprioception. As diagrammed below this involves the entire synthesis path - intent to vocalization of phones. The external world serves as the loopback connection – the robot hears its own vocalizations and applies the recognition process to it. The recovered intent is R' is the robot's perception of its own vocalization. This form of proprioception (or lack thereof) would account for speech impairment due to deafness. An individual with compromised hearing would not receive feedback and would not have access to recovered intent for correction.



Internal Proprioception (Mental Imagery)

The second form of proprioception is Internal Proprioception. Mental imagery used by athletes is an example of Internal Proprioception. As with External Proprioception the loop starts with the robot's Intent R. The majority of Synthesis and Recognition infrastructure is used but the lowest levels of Synthesis and Recognition are bypassed through internal feedback path. Internal proprioception would explain why mental imagery would tune or prepare the path by exercising it. It should be noted the Internal Proprioception loop could also be used as part of higher level planning, such as performing "what-if" scenarios, like "what if I were pulling into my garage and stepped on the accelerator instead of the brake."

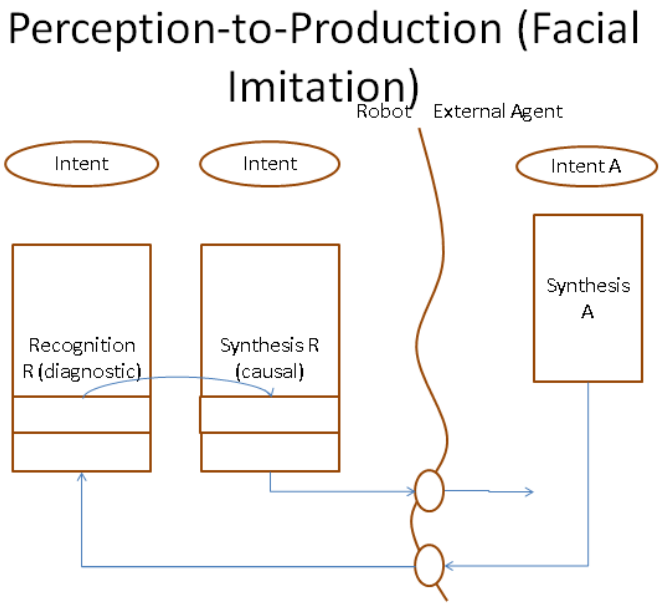


Internal Proprioception presumes the existence of an internal cross-modal path from synthesis to recognition.

Perception-to-Production Feedback (Facial Imitation)

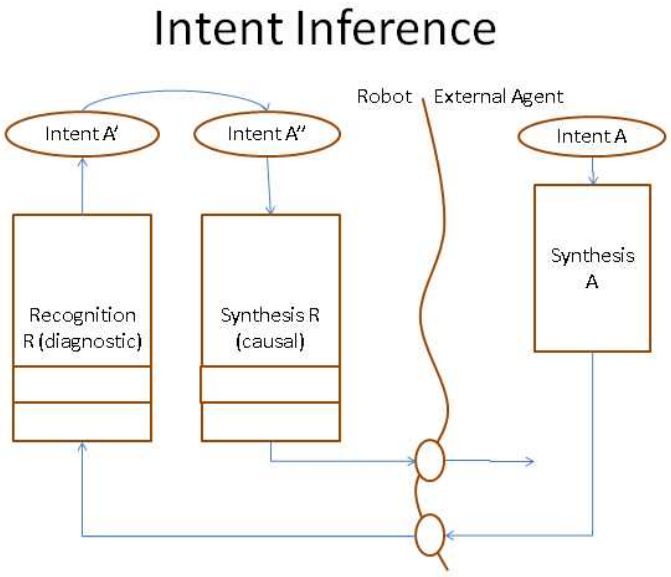
The third type of feedback path is that observed with facial imitation in infants. In this case Intent A is that of an external caregiver. The infant, or in our case, robot with only innate capabilities, has a low-level feedback path from perception to production. This feedback path would explain the low level facial imitation that we presume to be innate in the infants less than 24 hours old.

Perception-to-Production Feedback presumes the existence of an internal cross-modal path from recognition to synthesis.



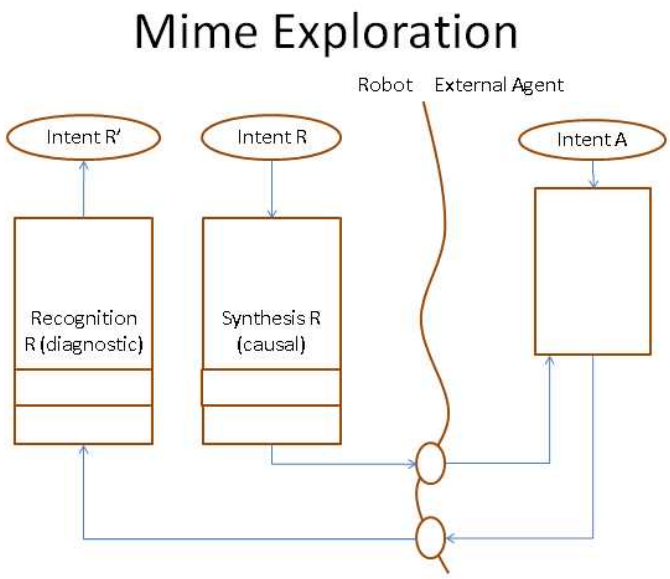
Intent Inference and Exploration

The next case is that of the intent inference and exploration. In this case the infant observes the failed behavior of a caregiver, infers the original intent, and produces the corrected intent. Our model explains this as the utilization of the full diagnostic path to recover an intent A' - the failed intent, which is translated into A`` the corrected intent, and output through the production system.



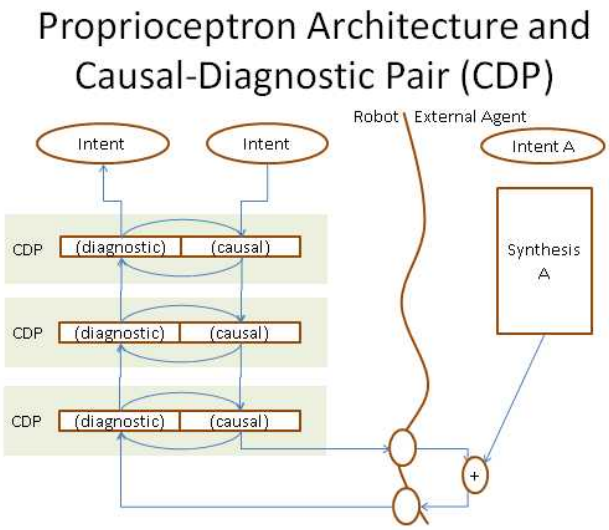
Mime Exploration

The final feedback path is that of mime exploration. In this case the infant, or in our case robot, is the initiator with intent R. The production system is used to generate output to the world. The external caregiver internalizes this and mimes back to the infant, or robot, which uses the perception system to recover its own intent R' as presented by the external caregiver. This feedback loop is very similar to Internal Proprioception and External Proprioception in that it starts with the robot's intent. But in this case the feedback loop is external, and the feedback is at a very high conceptual or semantic level. This might explain why infants find this type of proprioceptive feedback to be interesting - there is a level of correlation between R and R` involving higher level abstractions and manipulations.



5. Proprioceptron Architecture and CDPs

The final Proprioceptron Architecture is in the figure below. Here we have added a handful of hierarchical levels of increasing abstraction, each with an Internal Proprioception and Perception-To-Production feedback path. Each hierarchical level consists of a diagnostic element and a causal element, and this we refer to as a Causal-Diagnostic Pair (CDP). Thus the Proprioceptron Architecture is structured as a set of horizontally oriented CDPs, with each CDP consisting of the causal and diagnostic capabilities for production and perception at that level. Each CDP receives context/expectation from higher-level CDPs and input from the external world through lower level CDPs. Each CDP includes a path for Internal Proprioception and Perception-to-Production feedback. This contrasts with traditional software which is vertically integrated (a stovepipe for production and a stovepipe for perception), and Subsumption which is perception-to-production originating from outside the robot but does not include the concept of production-to-perception feedback for stimulus originating from the robot's internal intent.



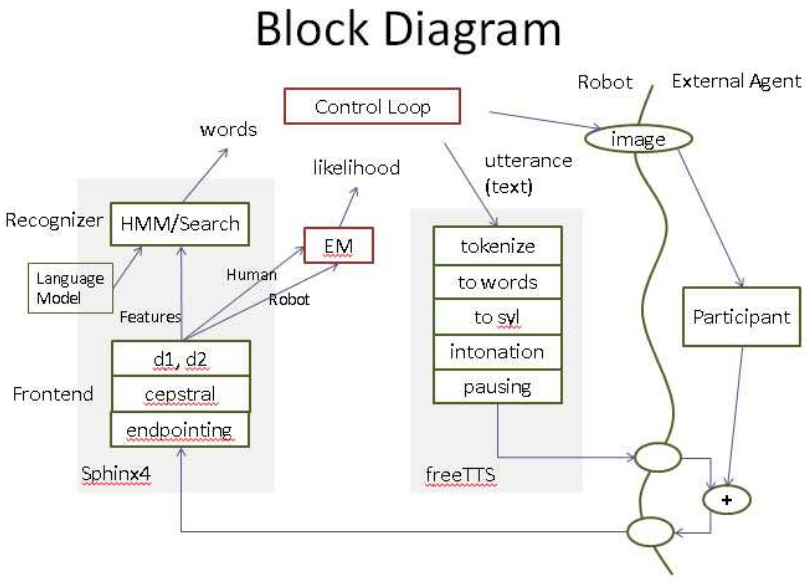
6. Proprioceptron Machine 1 - Experimental Infrastructure

Our experimental infrastructure - Proprioception Machine 1 is a rudimentary robot that utilizes speech recognition and speech synthesis to demonstrate the effect of External Proprioception in speech recognition. To situate the robot and create a social learning environment the robot is given a "head" which allows it to gesture, establish shared focus and provide feedback to the participant. The software components used for Proprioceptron Machine 1 are:

- CMUSphinx4 (recognition)
- freeTTS (synthesis)
- Weka (analysis)
- Mindstorms/NXT w/Lejos Java API (feedback/gesturing)
- Java Imagex library (image presentation)
- Eclipse IDE

Block Diagram

The block diagram for Proprioception Machine 1 is shown in the illustration below and follows that of the earlier architecture. The primary speech-related components are Sphinx4 and freeTTS. The Lejos API is not shown. The presentation of an image to the user is shown in the top right corner. A top-level control loop establishes robot intent and directs interaction with the participant.



The Sphinx4 system includes two major sub-blocks - the Frontend and the Recognizer. Frontend provides classic speech-oriented signal processing including endpointing, cepstral frequency-based feature definition, and calculation of first and second order derivatives The output of the frontend is 13 frequency-based features plus first and second order derivatives for a total of 39 features. The Recognizer sub-block of Sphinx4 provides recognition using a Hidden-Markov Model. The key elements of this Hidden-Markov search are a lattice of candidate events,

a search algorithm, and a language model to constrain the search. freeTTS provides text-to-speech synthesis converting utterance from the control loop into tokens, then to words and phonetic syllables with intonation and pausing.

An interface has been added to Sphinx4 so further analysis can be done on the features from the frontend. Specifically an Expectation-Maximization model can be created on selected words.

Game-driven Dialogue and Language Model

A game-driven dialog is used to give structure to interaction with the user. The first phase is introductory allowing the participant to become familiar with the robot and its capacities such as its ability to speak, recognize speech, ability to gesture, ability to handle basic turn-taking, and the robot's goal of social learning.

The second phase is game-play where the robot uses the full Sphinx4 recognition path to attempt to recognize the word. The data flow starts with the Control Loop which chooses one of three images to the user - a cat, a rat or a car. The user will state the word and the robot will use Sphinx4 to make an attempt to recognize the word. It will then articulate its understanding of the word back to the user asking the user if it had recognized the word correctly. It the word is incorrect the robot will then proceed to the third (proprioceptive exploration) portion of the experiment.

In the third (proprioceptive exploration) portion of the experiment the original raw frontend data captured from the user is used to create an Expectation-Maximization model. The robot then sequentially utters the nine candidate words which are the combinations ([c,r,t] a [c,r,t]) and proprioceptively receive these through the Sphinx4 frontend. The EM model is then used to compare the features from the robot's articulation of the candidate words with those of the participant's utterance and the word with the highest log-likelihood is chosen as the most likely candidate.

The language model used is a trigram model consisting of the words "cat", "rat", and "car", plus about 600 words which sound like these. Within the proprioceptive guessing portion of the game the robot will guess combinations ([c,r,t] a [c,r,t]) of which there are a total of nine.

7. Results

The Expectation-Maximization model used four clusters based on 13 of the frontend features, namely the first derivative of the cepstral frequency.

Both Sphinx4 and Proprioception Machine 1 had single-word recognition below 20%. It is believed the reasons for the low recognition rates are due to:

1. The single-word nature of recognition and resulting flat distribution of our language model. For example - the language model offers no guidance with respect to which is more likely - cat (a very common word) or cac (a word that is almost nonexistent).

2. The choice of short (single-syllable) words as our target. Single-syllable words typically consist of less than 200 timeslice samples from the frontend whereas a longer word or phrase would provide more features for the recognition and EM systems.

3. We did not use a headset microphone preferring instead a more situated environment for our robot

8. Summary

In Summary - we first reviewed five types of feedback observed in infants and adults. We reviewed Brooks' Subsumption Architecture and traditional software approaches and found neither satisfactorily explained these five types of feedback. We then presented the Proprioceptron Architecture which does explain the five types of feedback, and developed Proprioceptron Machine 1 which implements one type of feedback, external proprioception, to demonstrate the architecture's feasibility.

There are a number of opportunities for future experimentation based on this work. In particular, if the cornerstone of the Proprioception Architecture is the CDP - is possible to implement a CDP? For example, the final phase in freeTTS synthesis is "pausing" and the first phase in Sphinx4 recognition is "endpointing". These are obviously complementary functions, one causal and one diagnostic. Is it possible to implement pausing/endpointing as a CDP?

9. References

[1] Meltzoff, A. N., & Decety, J. (2003). What imitation tells us about social cognition: a rapprochement between developmental psychology and cognitive neuroscience. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *358*(1431), 491-500.

[2] Hudgins, C. V., & Numbers, F. C. (1942). *An investigation of the intelligibility of the speech of the deaf*. The Journal press.

[3] Richardson, A. (1967). MENTAL PRACTICE: A REVIEW AND DISCUSSION: II.*Research quarterly*.

[4] Brooks, R. A. (1991). Intelligence without representation. *Artificial intelligence*,*47*(1), 139-159.

[5] Brooks, R. (1986). A robust layered control system for a mobile robot. *Robotics and Automation, IEEE Journal of*, *2*(1), 14-23.