Continuous Listening for Unconstrained Spoken Dialog

Tim Paek, Eric Horvitz, and Eric Ringger

Microsoft Research
One Microsoft Way
Redmond, WA 98052 USA
{t-tpaek, horvitz, ringger}@microsoft.com

ABSTRACT

A major hindrance to rendering spoken dialog systems capable of ongoing, continuous listening without requiring a push-to-talk device is the problem of distinguishing speech which is intended for the system from that which is overheard. We present a decision-theoretic approach to this problem that exploits Bayesian models of spoken dialog at four levels of analysis within a domain-independent, multi-modal computational architecture called *Quartet*. We applied Quartet to the task of navigating PowerPoint slide shows during a spoken presentation in a prototype system called *Presenter*. We describe the runtime behavior of Presenter as well as the results of an experimental study comparing the performance of Presenter to human subjects in discriminating arbitrarily formed spoken requests for slide navigation during a recorded lecture.

1. INTRODUCTION

Ideally, a spoken dialog system should be hands-free, or capable of ongoing, continuous listening, irrespective of dialog context. A major hindrance to reaching this goal is the challenge of distinguishing speech that is intended for the system from that which is overheard. This problem prevails not only when the system is listening, but also when the system is speaking, since not all barge-in utterances are actually meant for the system.

In the absence of methods for discriminating when utterances are actually intended for the system, developers interested in building continuous listening systems are confronted with a delicate tradeoff. Although sensitivity is generally desirable in a speech recognizer, sensitivity at the wrong times can impede dialog and frustrate the user. For example, if the system is responding to a request and the user coughs or some other background noise is heard, the system may assume it took the wrong action, or that the user is altering the original utterance. Whatever the case, the natural flow of dialog is disrupted.

While people utilize multiple sources of information to reason about what they are hearing, such as whether the focus of attention is on them, and how well they have understood the goal associated with an utterance, spoken dialog systems have focused primarily on analyzing the phonetic structure of the speech input. In stark contrast to the rich reasoning skills about dialog context characteristic of human listening, these systems have relied on manual controls, such as push-to-talk devices, to filter out utterances that are unrelated to the domain.

We present a decision-theoretic approach to unconstrained, continuous listening that exploits Bayesian models of multiple sources of information at four levels of analysis within a domain-independent, multi-modal computational architecture called Quartet. Beginning with an overview of research in psychology and linguistics motivating the architecture, we describe representations and decision strategies that are relevant to distinguishing speech that is intended for the system from that which is overheard. We highlight the approach by applying Quartet to the task of navigating PowerPoint slide shows during a spoken presentation in a prototype system called Presenter. We describe the runtime behavior of Presenter as well as the results of an experimental study comparing the performance of Presenter to human subjects in discriminating arbitrarily formed spoken requests for slide navigation during a recorded lecture.

2. THEORETICAL BACKGROUND

When people engage in dialog, they typically do so with the intent of making themselves understood. To do this they need to make sure, as they speak, that the other participants are at the same time attending to, hearing, and understanding what they are saying. Since unresolved uncertainties often result in communication failures, people collaborate to establish and maintain the mutual belief that their utterances have been understood well enough for current purposes [2].

Researchers in psychology, linguistics, and artificial intelligence have argued that given so much coordination, conversation should be construed more as a collaborative effort or joint activity than as simply a structured sequence of utterances [2][4]. The process by which participants elegantly coordinate the presentation and acceptance of their utterances to establish, maintain, and confirm mutual understanding has been called *grounding* [2][3]. Grounding involves not only the consideration of how key uncertainties depend on each other and influence mutual understanding, but also what decisions to make in light of these uncertainties.

The approach we take in the Quartet architecture is to treat the process of grounding as a type of decision making under uncertainty [6][7][9][10]. We explicitly represent key uncertainties with Bayesian networks, and use local expected utility and value-of-information analyses to identify actions that maximize mutual understanding and bolster grounding. Since the networks and decision rules focus on the basic process of grounding, they generalize across task domains as well as multi-modal interactions, including visual and desktop information.

The Quartet architecture provides a framework for maintaining a dialog without the luxury of precise component technologies. Just as people rely on grounding techniques to compensate for extra uncertainties from impaired language skills such as

imperfect hearing, the Quartet architecture enables a dialog system to adapt its strategies based on its beliefs or representations of mutual understanding and evaluation of the costs and benefits of taking various grounding measures [10].

2.1. Four Levels of Analysis

Taking inspiration from Clark [2], we evaluate grounding in dialog at four levels of analysis, as displayed in Figure 1.

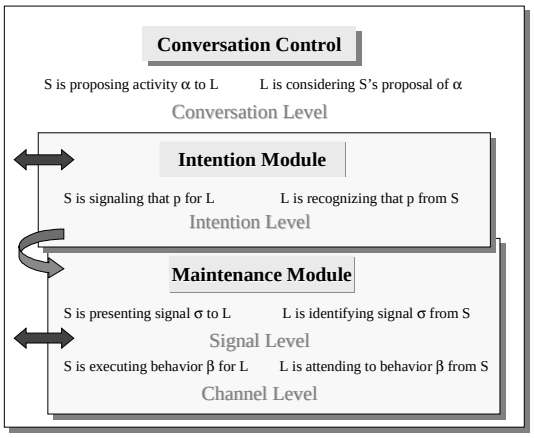

Figure 1: Four levels of analysis for reasoning about dialog context. The Quartet architecture utilizes information at these levels to discriminate overheard speech in continuous listening.

At the most basic level, which we denote as the *channel level*, a speaker *S* attempts to open a channel of communication by executing behavior β, such as an utterance or action, for listener *L*. However, *S* cannot get *L* to perceive β without coordination; *L* must be attending to and perceiving β *precisely* as *S* is executing it. Likewise, at the *signal level*, *S* presents β as a signal σ to *L*. Not all behaviors are meant to be signals, as for example, a cough. Hence, *S* and *L* must coordinate what *S* presents with what *L* identifies.

The *intention level* is where the task of understanding the *semantic content* of signals occurs, and where, to date, dialog systems focus almost entirely. Here, *S* signals some proposition *p* for *L*. What *L* recognizes to be the goal of *S* in signaling σ is *how L* will arrive at *p*. This again takes coordination.

Finally, at the *conversation level*, *S* proposes some joint activity α which *L* considers and takes up by providing a conditionally relevant response defined by α. *S* cannot get *L* to engage in the proposed activity without the coordinated participation and cooperation of *L*.

In summary, all four levels require coordination and collaboration to establish mutual understanding. For spoken dialog systems that integrate diverse component technologies, uncertainties like overhearing typically span multiple levels. For example, the speech recognizer at the signal level may pick up utterances that create problems for understanding at the intention level. Checking the channel level, however, may reveal that the user was actually attending to someone else.

3. QUARTET ARCHITECTURE

The Quartet architecture builds upon the four levels using two modules within a larger control subsystem, as also shown in Figure 1. The *Maintenance Module* handles uncertainties about the channel and signal levels, and the *Intention Module* about the intention level. Surrounding both modules is the *Conversation Control*, which keeps track of the grounding status by continually exchanging information with both modules, as depicted by the arrows. The Conversation Control operates at the meta-level by assessing the status of key variables in all of the modules; it decides where to focus on resolving uncertainties, and what grounding actions to take in light of their likely costs and benefits which fluctuate continuously as the dialog progresses [6][9]. Both Modules and the Conversation Control communicate within a distributed agent architecture called *InConcert* [1]. The design choice of maintaining distinct modules is discussed elsewhere [10].

Every component of Quartet employs temporal Bayesian networks [5] to model significant probabilistic dependencies at that level of analysis (model are shown in [6][7][9][10]). In the Maintenance Module, beliefs about channel fidelity are captured in a probability distribution over the "User's Focus of Attention," which keeps track of whether the user is attending to the system, another person, or to anything else. Relevant variables for diagnosing attention include eye gaze (face-pose tracking [12]), desktop activity, focus of attention at the previous time slice, and all kinds of timing information, such as the pause following an utterance. For the signal level, the distribution "Signal Identified" integrates uncertainty information from the speech recognizer (Microsoft Whisper system [8]) and natural language parser (Microsoft NLPWin system [11]). An overall "Maintenance Status" distribution summarizes both the attention and signal distributions, along with their probabilities in the previous time slice, into four grounding states: CHANNEL SIGNAL, CHANNEL NO SIGNAL, SIGNAL NO CHANNEL, and finally, NO CHANNEL NO SIGNAL. Note that in overhearing, the dominant state is SIGNAL NO CHANNEL; that is, the system receives a signal reasonably well, but the focus of attention is elsewhere.

3.1. Decision Strategy for Overhearing

Checking the state SIGNAL BUT NO CHANNEL in "Maintenance Status" gives one clue about overhearing, but another comes from the Intention Module. Using keywords, syntactic, and semantic features obtained from the natural language parser, a distribution is inferred over the "Intention Status," which conveys how likely the system understood the "meaning" or goal of an utterance. Intuitively, speech that is not intended for the system is not well understood in a given dialog domain.

The Conversation Control modifies the "Maintenance Status" and "Intention Status" distributions to reflect priors on their past performance or historical accuracy [9][10]. The two distributions are then used to infer the "Activity Goal," which diagnoses whether the user is participating in a joint activity with the system, another person, or doing something else. This is critical for detecting overhearing. To summarize progress, an overall "Grounding Status" distribution is also evaluated with respect to fives states: OKAY, CHANNEL FAILURE, SIGNAL

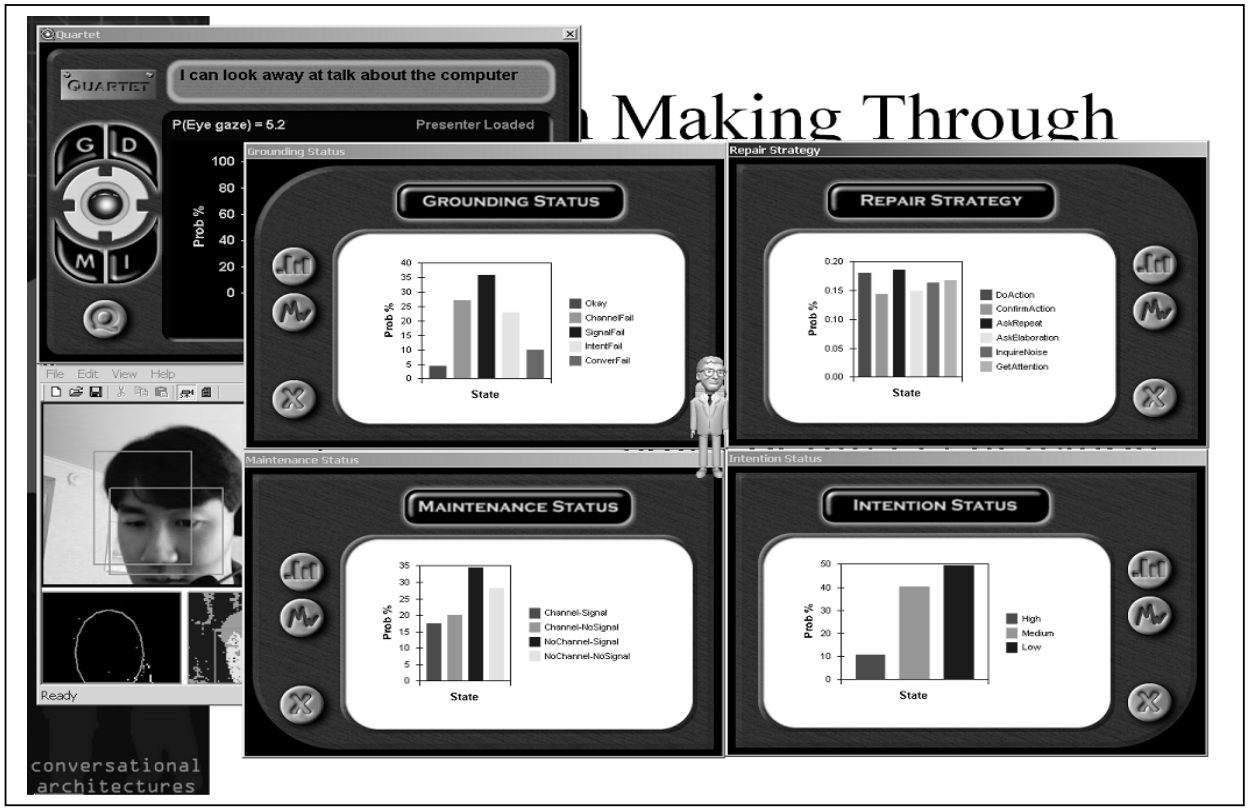

Figure 2: Presenter ignoring overheard speech by reasoning about dialog context at multiple levels of analysis and information.

FAILURE, INTENTION FAILURE, and CONVERSATION FAILURE.

With these distributions, to decide what action to take, the Conversation Control simply calculates the expected utility, or weighted sum of all possible actions given "Grounding Status" and "Activity Goal." These actions can be broken down into two decisions: first, whether or not to ignore an utterance, and second, what grounding strategy to use if the system heeds the utterance. Grounding strategies include fulfillment of requests, as well as repairs, such as asking for clarification [7].

4. PRESENTER

To assess the performance of the architecture in discriminating speech that is intended for the system from overheard speech, we applied Quartet to the task of navigating PowerPoint slide shows for a user during a spoken presentation in a system called Presenter. This involved adding a new Intention Module for the PowerPoint domain, setting parameters for the interface, and adjusting the utilities of various dialog outcomes [10].

Figure 2 displays Presenter embodied as an animated character. Here the user states, "***I can look away and talk about the computer***" which is heard by the recognizer as "***I can look away at talk about the computer***" using only a dictation grammar. Even though "***computer***" is a keyword for getting attention, since the probability of eye gaze on the system is so low, it infers that the most likely Maintenance Status is NO CHANNEL SIGNAL, as shown in the bottom left panel. Since the recognized utterance is not pertinent to the task of navigating PowerPoint, the Intention Status in the bottom right panel shows low understanding. With these two distributions, the system decides that the user is most likely speaking to someone else. Overall, it infers that if the user were actually speaking to the system it would be experiencing a grounding problem at the signal level, as shown in the upper left panel. The upper right panel displays the expected utilities of repair actions it would have taken if the system first decided to respond to the user.

5. EVALUATION EXPERIMENT

To get a quantitative measure of how well Presenter identifies arbitrarily formed spoken requests for slide navigation in a real-time lecture, we conducted an experimental study.

5.1 Method

Stimuli: A native English speaker was asked to record a 20 minute lecture while giving requests to a hypothetical person to move slides forward or backward. The speaker was naïve to the information sources used by Quartet. The speaker made 16 requests, only 3 of which were for moving backward.

Procedure: 9 naïve subjects were asked to listen to the recorded lecture and to press two buttons to move forward or backward when requested. The experiment did not include any visual

information, such as eye gaze of the speaker or the content of the slides. For Presenter, the audio recording was patched through the speech recognizer in real-time. Using multiple sources of information at the four levels described earlier, Presenter decided whether to ignore utterances detected by the recognizer. If it choose to heed an utterance, the grounding strategy with the highest expected utility was logged, and if that strategy was to do the action, the most likely action was logged.

5.2 Results

	Hit Rate	False Alarm Rate	A′
Human	0.92	0	0.98
Presenter	0.38	0.007	0.84

Table 1: Results of an evaluation experiment on Presenter.

The hit rate of Presenter failed to match that of human subjects, as shown in Table 1. Under closer analysis of the data, however, we found that, of the 16 requests, the speech engine was incorrectly recognizing 10. The most common misrecognition was mistaking "**Next slide**" for "**Excellent**." Of the remaining 6 requests, Presenter correctly selected the proper navigation action. In terms of false alarms, human subjects were perfect. For Presenter, the speech recognizer segmented utterances that were intended for an audience listening to the presentation into 135 final phrases. Presenter mistook only 1 of these phrases for a request. Since assumptions about normality do not hold in this signal detection task, we calculated the non-parametric measure A′ to compare sensitivities, where A′ = 0.5 is chance level (equal to a d′ of 0). Clearly, Presenter is performing well above chance, but is still *significantly* lower than human subjects ($t = 36.6$, $p < 0.0001$)

5.3 Discussion

While Presenter seemed to be quite sensitive in distinguishing speech that was intended for the audience, the robustness of the architecture was not fully explored in this experiment. The Quartet architecture can be employed to work collaboratively with users to resolve uncertainties about intentions (i.e., grounding). For example, if Presenter is unsure of whether an utterance is directed for the system, it may simply ask the user. Such clarification dialog can enhance the performance of the system at the cost of additional interaction [7]. Evidence from interaction is critical for the system, and for communication in general. In a real life situation, if a speaker asks someone to move to the next slide and the person does not respond, the speaker will wait for a certain amount of time and ask again. Quartet considers all such information. The more information Quartet gathers, the better the performance. Indeed, eye gaze plays a significant role in discriminating overheard speech, though this experiment did not allow for visual information.

For future studies, we plan to evaluate the appropriateness of interactive responses, such as repairs, as well as to conduct lesion experiments, where parts of the architecture are removed to examines its effect on behavior. Furthermore, we plan to record more lectures as testing data to find common misrecognitions and patterns of behavior that can be learned automatically and incorporated into the Bayesian models. The model tested in this experiment was originally hand-crafted.

6. CONCLUSION

In pursuit of spoken dialog systems that are capable of unconstrained, continuous listening without the need for a push-to-talk device, we have presented a domain-independent, multi-modal computational architecture, Quartet. The architecture analyzes multiple sources of information at four levels of dialog context to infer key probability distributions that are useful for discriminating when an utterance is directed toward the system or elsewhere. In particular, overheard speech is often characterized by low uncertainty at the signal level but high uncertainty at the channel level (i.e., SIGNAL NO CHANNEL in the Maintenance Module), as well as high uncertainty at the intention level (i.e., low understanding in the Intention Module). To assess the performance of the architecture, we described the runtime behavior of Presenter, which applies Quartet to the task of navigating PowerPoint slide shows, as well as the results of a preliminary experiment that compares Presenter to human subjects in discriminating arbitrarily formed spoken requests for slide navigation during a recorded lecture

6. ACKNOWLEDGMENTS

We are deeply indebted to Steve Austin, Martin Calsyn, and Matt Rhoten for developing and supporting InConcert, the underlying messaging infrastructure of Quartet.

REFERENCES

1. Calsyn, M. & Meyers, B. 2000. InConcert. Unpublished *MSR Technical Report*.
2. Clark, H.H. 1996. *Using Language*. Cambridge University Press.
3. Clark, H.H. & Brennan, S.A.. 1991. Grounding in communication. In *Perspectives on Socially Shared Cognition*, 127-149. APA Books
4. Cohen, P.R. & Levesque, H.J. 1991. Teamwork. *Nous*, 25(4): 487-512.
5. Dagum, P., Galper, A., & Horvitz, E. 1992. Dynamic network models for forecasting. In *Proc. of the Eighth Workshop on UAI*, 41-48.
6. Horvitz, E. & Paek, T. 1999. A computational architecture for conversation. *Proc. of the Seventh International Conference on User Modeling*, 201-210.
7. Horvitz, E. & Paek, T. 2000. DeepListener: Harnessing Expected Utility to Guide Clarification Dialog in Spoken Language Systems. ICSLP 2000, Beijing.
8. Huang, X., Acero, A., Alleva, F., Hwang, M., Jiang, L., & Mahajan, M. (1995). Microsoft Windows highly intelligent speech recognizer: WHISPER. In *Proc. of ICASSP*. IEEE.
9. Paek, T. & Horvitz, E. 1999. Uncertainty, utility, and misunderstanding. *AAAI Fall Symposium on Psychological Models of Communication*, 85-92.
10. Paek, T. & Horvitz, E. 2000. Conversation as action under uncertainty. *Proc. of the Sixteenth Conference UAI*, 455-464. Morgan Kaufmann.
11. Richardson, S. 1994. Bootstrapping statistical processing into a rule-based natural language parser. In *ACL Workshop The Balancing Act*.
12. Toyama, K. 1998. Prolegomena for robust face tracking. *MSR Technical Report*, MSR-TR-98-65.