Integrating Incremental Speech Recognition and POMDP-based Dialogue Systems

Ethan O. Selfridge[†], **Iker Arizmendi**[‡], **Peter A. Heeman**[†], **and Jason D. Williams**[1]

[†] Center for Spoken Language Understanding, Oregon Health & Science University, Portland, OR, USA
[‡]AT&T Labs – Research, Shannon Laboratory, Florham Park, NJ, USA
[1]Microsoft Research, Redmond, WA, USA
{selfridg,heemanp}@ohsu.edu
iker@research.att.com jason.williams@microsoft.com

Abstract

The goal of this paper is to present a first step toward integrating Incremental Speech Recognition (ISR) and Partially-Observable Markov Decision Process (POMDP) based dialogue systems. The former provides support for advanced turn-taking behavior while the other increases the semantic accuracy of speech recognition results. We present an *Incremental Interaction Manager* that supports the use of ISR with strictly turn-based dialogue managers. We then show that using a POMDP-based dialogue manager with ISR substantially improves the semantic accuracy of the incremental results.

1 Introduction and Background

This paper builds toward integrating two distinct lines of research in spoken dialogue systems: incremental speech recognition (ISR) for input, and Partially Observable Markov Decision Processes (POMDPs) for dialogue management.

On the one hand, ISR improves on whole-utterance speech recognition by streaming results to the dialogue manager (DM) in real time (Baumann et al., 2009; Skantze and Schlangen, 2009). ISR is attractive because it enables sophisticated system behavior such as interruption and back-channeling. However, ISR output is particularly error-prone, and often requires a specialized dialogue manager to be written (Buß and Schlangen, 2011; Schlangen and Skantze, 2009).

[1]Work done while at AT&T Labs - Research

On the other hand, POMDP-based dialogue managers improve on traditional approaches by (in part) tracking a distribution over many possible dialogue states, rather than just one, thereby improving robustness to speech recognition errors (Williams and Young, 2007; Thomson and Young, 2010; Young et al., 2010). The overall aim of combining these two lines of research is to improve the robustness of error-prone ISR output.

To our knowledge only one study to date has combined ISR and POMDPs. Lu et al. (2011) show how 1-best ISR hypotheses can be used within a single dialogue turn. This work is different than the present paper, where we use N-Best lists of ISR results across multiple turns of a dialogue.

Specifically, this paper makes two contributions. First, as a foundation, we introduce an *Incremental Interaction Manager* (IIM) that enables ISR to be used within the traditional turn-based dialogue management framework. The IIM confers many, but not all, of the benefits of ISR without requiring modification to a traditional dialogue manager. Thus, in theory, any existing dialogue system architecture could use ISR with the addition of an IIM. Second, we show that pairing our IIM with a POMDP-based dialogue manager yields a substantial improvement in accuracy for incremental recognition results at the dialogue level.

The paper is organized as follows. Section 2 describes the IIM, section 3 describes the POMDP integration, sections 4 and 5 describe experiments and results, and section 6 concludes.

Table 1: Example IIM operation. P = partial ISR result; A = dialogue action.

ISR	IIM	Original DM state	Copied DM state	DM Action
Prompt: "Where are you leaving from?"				
yew	Rej. P	0	0	-
ridge	Acc. P / Rej. A	0	0	"I'm sorry..."
mckee	Acc. P / Acc. A	0	1	"Ok, Mckee..."
mckeesport	Acc. P / Acc. A	0	2	"Ok, Mckeesport.."
mckeesport center	Acc. P / Rej. A	0	2	"Ok, Mckeesport.."
Prompt: "Ok, Mckeesport. Where are you going to?"				
pitt	Acc. P / Rej. A	2	4	"I'm sorry..."
pittsburgh	Acc. P / Acc. A	2	5	"Ok, Pittsburgh..."

2 Incremental Interaction manager

The Incremental Interaction Manager (IIM) mediates communication between the incremental speech recognizer and the DM. The key idea is that the IIM evaluates potential dialogue moves by applying ISR results to temporary instances of the DM. The IIM *copies* the current state of the DM, provides the copied DM with a recognition result, and inspects the action that the copied DM would take.[2] If the action does not sufficiently advance the dialogue (such as re-asking the same question), the action is rejected and the copied DM is discarded. If the action advances the dialogue (such as asking for or providing new information), then that action is immediately executed.

The system should gracefully handle revisions following a premature action execution, and a copying procedure is a viable solution for any DM. When a revision is received, a *second* copy of the original DM is made and the new ISR result is passed to that second copy; if that second copy takes an action that advances the dialogue *and is different* from the action generated by the first copy, then the first action is terminated, the first copy of the DM is discarded, the second action is initiated, and the second copy assumes the position of the first copy. Additional revisions can be handled by following the same procedure. Terminating a speech action and immediately starting another can be jarring ("Say a city / Ok, Boston..."), which can be mitigated by preceding actions with either a sound or simple silence (at the expense of some response delay). Once recognition is complete, the copied DM is installed as the new original DM.

Many ISR results can be discarded before passing them to the DM. First, only incremental results that could correspond to complete user utterance are considered: incomplete results are discarded and never passed to the DM. In addition, ISR results are often unstable, and it is undesirable to proceed with an ISR result if it will very likely be revised. Thus each candidate ISR result is scored for stability (Selfridge et al., 2011) and results with scores below a manually-set threshold are discarded.

Table 1 shows an example of the recognizer, the IIM, and the DM. For sake of clarity, stability scores are not shown. The system asks "Where are you leaving from?" and the user answers "Mckeesport Center." The IIM receives five ISR results (called *partials*), rejecting the first, *yew*, because its stability score is too low (not shown). With the second, *ridge*, it copies the DM, passes *ridge* to the copy, and discards the action of the copied DM (also discarded) because it does not advance the dialogue. It accepts and begins to execute the action generated by the third partial, *mckee*. The fourth partial revises the action, and the fifth action is rejected since it is the same. The original DM is then discarded and the copied DM state is installed in its place.

Overall, the IIM enables a turn-based DM to enjoy many of the benefits of ISR – in particular, the ability to make turn-taking decisions with a complete account of the dialogue history.

[2] If the DM design does *not* force a state transition following a result then the DM supplies the the action without copying.

3 Integrating ISR with a POMDP-based dialogue manager

A (traditional) dialogue manager based on a partially observable Markov decision process (POMDP DM) tracks a probability distribution over multiple hidden dialogue states called a *belief state* (Williams and Young, 2007).[3] As such, POMDP DMs readily make use of the entire ASR N-Best list, even for low-confidence results — the confidence level of each N-Best list item contributes proportionally to the probability of its corresponding hidden state.

It is straightforward to integrate ISR and a POMDP DM using the IIM. Each item on the N-Best list of an incremental result is assigned a confidence score (Williams and Balakrishnan, 2009) and passed to the POMDP DM as if it were a complete result, triggering a belief state update. Note that this approach is not *predicting* future user speech from partial results (DeVault et al., 2009; Lu et al., 2011), but rather (tentatively) assuming that partial results are complete.

The key benefit is that a belief state generated from an incremental result incorporates all of the contextual information available to the system *from the start of the dialogue until the moment of that incremental result*. By comparison, an isolated incremental result includes only information from the current utterance. If the probability models in the POMDP are estimated properly, belief states should be more accurate than isolated incremental results.

4 Experimental design

For our experiments we used a corpus of 1037 calls from real users to a single dialogue system that provides bus timetable information for Pittsburgh, PA (a subsequent version of Williams (2011)). This dialogue system opened by asking the caller to say a bus route number or "I don't know"; if the system had insufficient confidence following recognition, it repeated the question. We extracted the first 3 responses to the system's bus route question. Often the system did not need to ask 3 times; our experimental set contained 1037 calls with one or more attempts, 586 calls with two or more attempts, and 356 calls with three or more attempts. These utterances were all transcribed, and tagged for the bus route they contained, if any: 25% contained neither a route nor "I don't know".

We ran incremental speech recognition on each utterance using Lattice-Aware Incremental Speech Recognition (Selfridge et al., 2011) on the AT&T WATSON^{SM} speech recognizer (Goffin et al., 2005) with the same rule-based language models used in the production system. On average, there were 5.78, 5.44, and 5.11 incremental results per utterance (plus an utterance-final result) for the first, second, and third attempts. For each incremental result, we noted its time stamp and interpretation: *correct*, if the interpretation was present and correct, otherwise *incorrect*. Each incremental result included an N-Best list, from which we determined oracle accuracy: *correct* if the correct interpretation was present anywhere on the most recent ISR N-Best list, otherwise *incorrect*.

Each incremental result was then passed to the IIM and POMDP DM. The models in the POMDP DM were estimated using data collected from a different (earlier) time period. When an incremental result updated the belief state, the top hypothesis for the route was extracted from the belief state and scored for correctness. For utterances in the first attempt, the belief state was initialized to its prior; for subsequent attempts, it incorporated all of the prior (whole-turn) utterances. In other words, each attempt was begun assuming the belief state had been running up to that point.

5 Results and Discussion

We present results by showing instantaneous semantic accuracy for the raw incremental result (baseline), the top belief state, and oracle. Instantaneous semantic accuracy is shown with respect to the *percent* of the total recognition time the partial is recognized at. An utterance is incorrect if it has no incremental result before a certain percentage.

We show 2 sets of plots. Figure 1 shows only incremental recognition results and excludes the end-of-utterance (*phrase*) results; Figure 2 shows incremental recognition results and includes phrase results. It is useful to view these separately since the phrase result, having access to all the speech, is sub-

[3] It also uses reinforcement learning to choose actions, although in this paper we are not concerned with this aspect.

Figure 1: Instantaneous semantic accuracy of incremental results, excluding phrase-final results

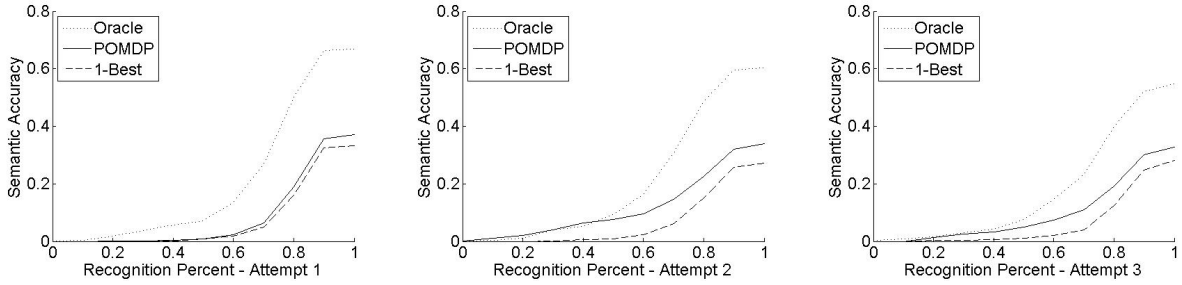
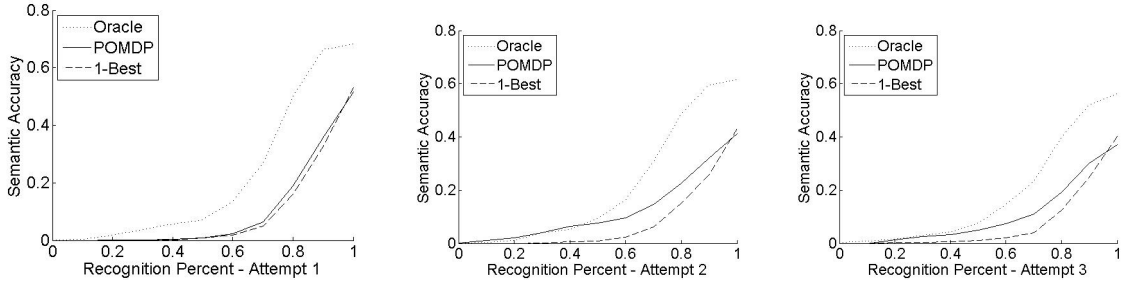


Figure 2: Instantaneous semantic accuracy of incremental and phrase-final results



stantially more accurate than the incremental results.

Figure 1 shows that the POMDP is more accurate than the raw incremental result (excluding end-of-phrase results). Its performance gain is minimal in attempt 1 because the belief is informed only by the prior. In attempt 2 and 3, the gain is larger since the belief also benefits from the previous attempts. Since the top POMDP result in subsequent attempts is sometimes already correct (because it incorporates past recognitions), the POMDP sometimes meets and occasionally exceeds the oracle during the early portions of attempts 2 and 3.

Figure 2 shows that when end-of-phrase recognition results are included, the benefit of the belief state is limited to the initial portions of the second and third turns. This is because the POMDP models are not fit well to the data: the models were estimated from an earlier version of the system, with a different user base and different functionality. Identifying and eliminating this type of mismatch is an important issue and has been studied before (Williams, 2011).

Taken as a whole, we find that using belief tracking increases the accuracy of partials by over 8% (absolute) in some cases. Even though the final phrase results of the 1-best list are more accurate than the belief state, the POMDP shows better accuracy on the volatile incremental results. As compared to the whole utterance results, incremental results have lower 1-best accuracy, yet high oracle accuracy. This combination is a natural fit with the POMDPs belief state, which considers the whole N-Best list, effectively re-ranking it by synthesizing information from dialogue history priors.

6 Conclusion

This paper has taken a step toward integrating ISR and POMDP-based dialogue systems. The Incremental Interaction Manager (IIM) enables a traditional turn-based DM to make use of incremental results and enjoy many their benefits. When this IIM is paired with a POMDP DM, the interpretation accuracy of incremental results improves substantially. In the future we hope to build on this work by incorporating Reinforcement Learning into turn-taking and dialogue action decisions.

Acknowledgments

Thanks to Vincent Goffin for help with this work, and to the anonymous reviewers for their comments and critique. We acknowledge funding from the NSF under grant IIS-0713698.

References

T. Baumann, M. Atterer, and D. Schlangen. 2009. Assessing and improving the performance of speech recognition for incremental systems. In *Proc. NAACL: HLT*, pages 380–388. Association for Computational Linguistics.

O. Buß and D. Schlangen. 2011. Dium–an incremental dialogue manager that can produce self-corrections. *Proceedings of semdial*.

David DeVault, Kenji Sagae, and David Traum. 2009. Can i finish? learning when to respond to incremental interpretation results in interactive dialogue. In *Proceedings of the SIGDIAL 2009 Conference*, pages 11–20, London, UK, September. Association for Computational Linguistics.

V. Goffin, C. Allauzen, E. Bocchieri, D. Hakkani-Tur, A. Ljolje, S. Parthasarathy, M. Rahim, G. Riccardi, and M. Saraclar. 2005. The AT&T WATSON speech recognizer. In *Proceedings of ICASSP*, pages 1033–1036.

D. Lu, T. Nishimoto, and N. Minematsu. 2011. Decision of response timing for incremental speech recognition with reinforcement learning. In *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*, pages 467–472. IEEE.

D. Schlangen and G. Skantze. 2009. A general, abstract model of incremental dialogue processing. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 710–718. Association for Computational Linguistics.

E.O. Selfridge, I. Arizmendi, P.A. Heeman, and J.D. Williams. 2011. Stability and accuracy in incremental speech recognition. In *Proceedings of the SIGdial 2011*.

G. Skantze and D. Schlangen. 2009. Incremental dialogue processing in a micro-domain. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 745–753. Association for Computational Linguistics.

B. Thomson and S. Young. 2010. Bayesian update of dialogue state: A pomdp framework for spoken dialogue systems. *Computer Speech & Language*, 24(4):562–588.

J.D. Williams and S. Balakrishnan. 2009. Estimating probability of correctness for asr n-best lists. In *Proc SIGDIAL, London, United Kingdom*.

J.D. Williams and S. Young. 2007. Partially observable markov decision processes for spoken dialog systems. *Computer Speech & Language*, 21(2):393–422.

J.D. Williams. 2011. An empirical evaluation of a statistical dialog system in public use. In *Proceedings of the SIGDIAL 2011 Conference*, pages 130–141. Association for Computational Linguistics.

S. Young, M. Gasic, S. Keizer, F. Mairesse, J. Schatzmann, B. Thomson, and K. Yu. 2010. The hidden information state model: A practical framework for pomdp-based spoken dialogue management. *Computer Speech & Language*, 24(2):150–174.