Incremental Speech Understanding in a Multi-Party Virtual Human Dialogue System

David DeVault and **David Traum**
Institute for Creative Technologies
University of Southern California
12015 Waterfront Drive, Playa Vista, CA 90094
{devault,traum}@ict.usc.edu

1 Extended Abstract

This demonstration highlights some emerging capabilities for incremental speech understanding and processing in virtual human dialogue systems. This work is part of an ongoing effort that aims to enable realistic spoken dialogue with virtual humans in multi-party negotiation scenarios (Plüss et al., 2011; Traum et al., 2008b). These scenarios are designed to allow trainees to practice their negotiation skills by engaging in face-to-face spoken negotiation with one or more virtual humans.

An important component in achieving naturalistic behavior in these negotiation scenarios, which ideally should have the virtual humans demonstrating fluid turn-taking, complex reasoning, and responding to factors like trust and emotions, is for the virtual humans to begin to understand and in some cases respond in real time to users' speech, as the users are speaking (DeVault et al., 2011b). These responses could range from relatively straightforward turn management behaviors, like having a virtual human recognize when it is being addressed by a user utterance, and possibly turn to look at the user who has started speaking, to more complex responses such as emotional reactions to the content of what users are saying.

The current demonstration extends our previous demonstration of incremental processing (Sagae et al., 2010) in several important respects. First, it includes additional indicators, as described in (DeVault et al., 2011a). Second, it is applied to a new domain, an extension of that presented in (Plüss et al., 2011). Finally, it is integrated with the dialogue models (Traum et al., 2008a), such that each partial interpretation is given a full pragmatic interpretation by each virtual character, which can be used to generate real-time incremental non-verbal feedback (Wang et al., 2011).

Figure 1: SASO negotiation in the saloon: Utah (left) looking at Harmony (right).

Our demonstration is set in an implemented multi-party negotiation domain (Plüss et al., 2011) in which two virtual humans, Utah and Harmony (pictured in Figure 1), talk with two human negotiation trainees, who play the roles of Ranger and Deputy. The dialogue takes place inside a saloon in an American town in the Old West. In this negotiation scenario, the goal of the two human role players is to convince Utah and Harmony that Utah, who is currently employed as the local bartender, should take on the job of town sheriff.

One of the research aims for this work is to support natural dialogue interaction, an example of which is the excerpt of human role play dialogue shown in Figure 2. One of the key features of immersive role plays is that people often react in multiple ways to the utterances of others as they are speaking. For example, in this excerpt, the beginning of the

Ranger	We can't leave this place and have it overrun by outlaws. Uh there's no way that's gonna happen so we're gonna make sure we've got a properly deputized and equipped sheriff ready to maintain order in this area. *00:03:56.660 - 00:04:08.830*
Deputy	Yeah and you know and and we're willing to *00:04:06.370 - 00:04:09.850*
Utah	And I don't have to leave the bar completely. I can still uh be here part time and I can um we can hire someone to do the like day to day work and I'll do the I'll supervise them and I'll teach them. *00:04:09.090 - 00:04:22.880*

Figure 2: Dialogue excerpt from one of the role plays. Timestamps indicate the start and end of each utterance.

Deputy's utterance overlaps the end of the Ranger's, and then Utah interrupts the Deputy and takes the floor a few seconds later.

Our prediction approach to incremental speech understanding utilizes a corpus of in-domain spoken utterances, including both paraphrases selected and spoken by system developers, as well as spoken utterances from user testing sessions (DeVault et al., 2011b). An example of a corpus element is shown in Figure 3. In previous negotiation domains, we have found a fairly high word error rate in automatic speech recognition results for such spontaneous multi-party dialogue data; for example, our average word error rate was 0.39 in the SASO-EN negotiation domain (Traum et al., 2008b) with many (15%) out of domain utterances. Our speech understanding framework is robust to these kinds of problems (DeVault et al., 2011b), partly through approximating the meaning of utterances. Utterance meanings are represented using an attribute-value matrix (AVM), where the attributes and values represent semantic information that is linked to a domain-specific ontology and task model (Traum, 2003; Hartholt et al., 2008; Plüss et al., 2011). The AVMs are linearized, using a path-value notation, as seen in Figure 3. In our framework, we use this data to train two data-driven models, one for incremental natural language understanding, and a second for incremental confidence modeling.

The first step is to train a predictive incremental understanding model. This model is based on maximum entropy classification, and treats entire individual frames as output classes, with input features extracted from partial ASR results, calculated in increments of 200 milliseconds (DeVault et al., 2011b).

- Utterance (speech): *i've come here today to talk to you about whether you'd like to become the sheriff of this town*
- ASR (NLU input): *have come here today to talk to you about would the like to become the sheriff of this town*
- Frame (NLU output):

```
<S>.mood interrogative
<S>.sem.modal.desire want
<S>.sem.prop.agent utah
<S>.sem.prop.event providePublicServices
<S>.sem.prop.location town
<S>.sem.prop.theme sheriff-job
<S>.sem.prop.type event
<S>.sem.q-slot polarity
<S>.sem.speechact.type info-req
<S>.sem.type question
```

Figure 3: Example of a corpus training example.

Each partial ASR result then serves as an incremental input to NLU, which is specially trained for partial input as discussed in (Sagae et al., 2009). NLU is predictive in the sense that, for each partial ASR result, the NLU module produces as output the *complete* frame that has been associated by a human annotator with the user's *complete* utterance, even if that utterance has not yet been fully processed by the ASR. For a detailed analysis of the performance of the predictive NLU, see (DeVault et al., 2011b).

The second step in our framework is to train a set of incremental confidence models (DeVault et al., 2011a), which allow the agents to assess in real time, while a user is speaking, how well the understanding process is proceeding. The incremental confidence models build on the notion of NLU F-score, which we use to quantify the quality of a predicted NLU frame in relation to the hand-annotated correct frame. The NLU F-score is the harmonic mean of the precision and recall of the attribute-value pairs (or *frame elements*) that compose the predicted and correct frames for each partial ASR result. By using precision and recall of frame elements, rather than simply looking at frame accuracy, we take into account that certain frames are more similar than others, and allow for cases when the correct frame is not in the training set.

Each of our incremental confidence models makes a binary prediction for each partial NLU result as an utterance proceeds. At each time t dur-

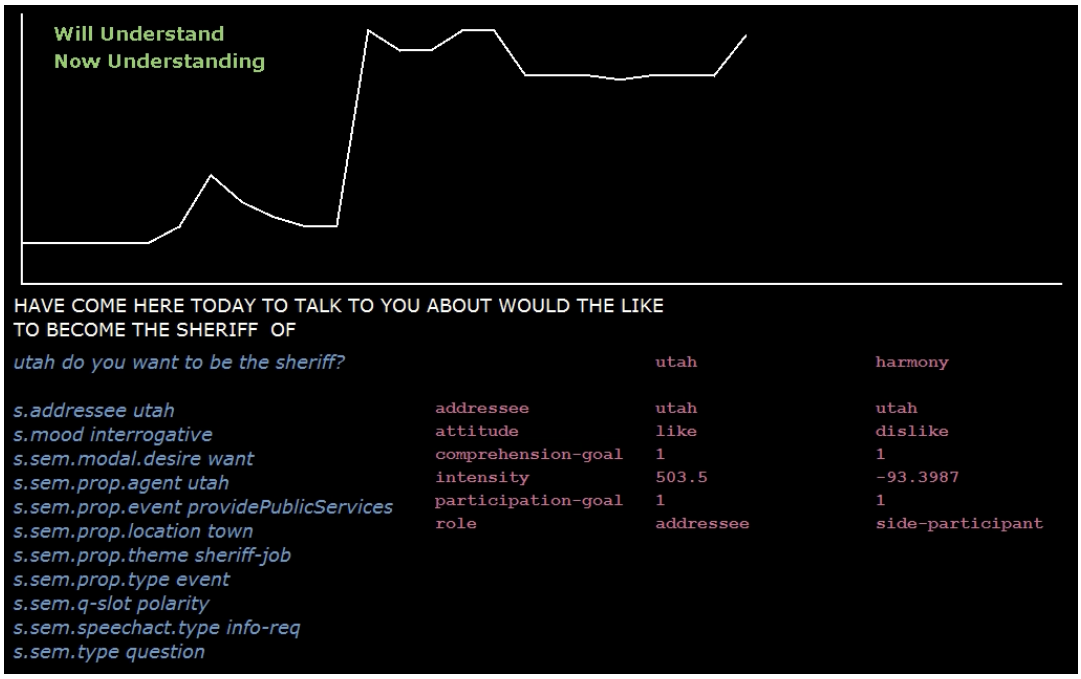

Figure 4: Visualization of Incremental Speech Processing.

ing an utterance, we consider the current NLU F-Score F_t as well as the final NLU F-Score F_{final} that will be achieved at the conclusion of the utterance. In (DeVault et al., 2009) and (DeVault et al., 2011a), we explored the use of data-driven decision tree classifiers to make predictions about these values, for example whether $F_t \geq \frac{1}{2}$ (current level of understanding is "high"), $F_t \geq F_{\text{final}}$ (current level of understanding will not improve), or $F_{\text{final}} \geq \frac{1}{2}$ (final level of understanding will be "high"). In this demonstration, we focus on the first and third of these incremental confidence metrics, which we summarize as "Now Understanding" and "Will Understand", respectively. In an evaluation over all partial ASR results for 990 utterances in this new scenario, we found the Now Understanding model to have precision/recall/F-Score of .92/.75/.82, and the Will Understand model to have precision/recall/F-Score of .93/.85/.89. These incremental confidence models therefore provide potentially useful real-time information to Utah and Harmony about whether they are currently understanding a user utterance, and whether they will ever understand a user utterance.

The incremental ASR, NLU, and confidence models are passed to the dialogue managers for each of the agents, Harmony and Utah. These agents then relate these inputs to their own models of dialogue context, plans, and emotions, to calculate pragmatic interpretations, including speech acts, reference resolution, participant status, and how they feel about what is being discussed. A subset of this information is passed to the non-verbal behavior generation module to produce incremental non-verbal listening behaviors (Wang et al., 2011).

In support of this demonstration, we have extended the implementation to include a real-time visualization of incremental speech processing results, which will allow attendees to track the virtual humans' understanding as an utterance progresses. An example of this visualization is shown in Figure 4.

2 Demo script

The demonstration begins with the demo operator providing a brief overview of the system design, negotiation scenario, and incremental processing capabilities. The virtual humans Utah and Harmony (see Figure 1) are running and ready to begin a dialogue with the user, who will play the role of the Ranger. As the user speaks to Utah or Harmony, attendees can observe the real time visualization of speech

processing to observe changes in the incremental processing results as the utterance progresses. Further, the visualization interface enables the demo operator to "rewind" an utterance and step through the incremental processing results that arrived each 200 milliseconds, highlighting how specific partial ASR results can change the virtual humans' understanding or confidence.

For example, Figure 4 shows the incremental speech processing state at a moment 4.8 seconds into a user's 7.4 second long utterance, *i've come here today to talk to you about whether you'd like to become the sheriff of this town*. At this point in time, the visualization shows (at top left) that the virtual humans are confident that they are both Now Understanding and Will Understand this utterance. Next, the graph (in white) shows the history of the agents' expected NLU F-Score for this utterance (ranging from 0 to 1). Beneath the graph, the partial ASR result (`HAVE COME HERE TODAY TO TALK TO YOU ABOUT...`) is displayed (in white), along with the currently predicted NLU frame (in blue). For ease of comprehension, an English gloss (*utah do you want to be the sheriff?*) for the NLU frame is also shown (in blue) above the frame.

To the right, in pink, we show some of Utah and Harmony's agent state that is based on the current incremental NLU results. The display shows that both of the virtual humans believe that Utah is being addressed by this utterance, that utah has a positive attitude toward the content of the utterance while harmony does not, and that both have comprehension and participation goals. Further, Harmony believes she is a side participant at this moment. The demo operator will explain and discuss this agent state information, including possible uses for this information in response policies.

Acknowledgments

We thank all the members of the ICT Virtual Humans team. The project or effort described here has been sponsored by the U.S. Army Research, Development, and Engineering Command (RDECOM). Statements and opinions expressed do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

References

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 SIGDIAL*.

David DeVault, Kenji Sagae, and David Traum. 2011a. Detecting the status of a predictive incremental speech understanding model for real-time decision-making in a spoken dialogue system. In *Proceedings of InterSpeech*.

David DeVault, Kenji Sagae, and David Traum. 2011b. Incremental interpretation and prediction of utterance meaning for interactive dialogue. *Dialogue & Discourse*, 2(1).

Arno Hartholt, Thomas Russ, David Traum, Eduard Hovy, and Susan Robinson. 2008. A common ground for virtual humans: Using an ontology in a natural language oriented virtual human architecture. In *Proceedings of LREC*, Marrakech, Morocco, may.

Brian Plüss, David DeVault, and David Traum. 2011. Toward rapid development of multi-party virtual human negotiation scenarios. In *Proceedings of SemDial 2011, the 15th Workshop on the Semantics and Pragmatics of Dialogue*.

Kenji Sagae, Gwen Christian, David DeVault, and David R. Traum. 2009. Towards natural language understanding of partial speech recognition results in dialogue systems. In *Short Paper Proceedings of NAACL HLT*.

Kenji Sagae, David DeVault, and David R. Traum. 2010. Interpretation of partial utterances in virtual human dialogue systems. In *Demonstration Proceedings of NAACL-HLT*.

D. Traum, W. Swartout, J. Gratch, and S. Marsella. 2008a. A virtual human dialogue model for non-team interaction. In L. Dybkjaer and W. Minker, editors, *Recent Trends in Discourse and Dialogue*. Springer.

David Traum, Stacy Marsella, Jonathan Gratch, Jina Lee, and Arno Hartholt. 2008b. Multi-party, multi-issue, multi-strategy negotiation for multi-modal virtual agents. In *Proceedings of IVA*.

David Traum. 2003. Semantics and pragmatics of questions and answers for dialogue agents. In *Proc. of the International Workshop on Computational Semantics*, pages 380–394, January.

Zhiyang Wang, Jina Lee, and Stacy Marsella. 2011. Towards more comprehensive listening behavior: Beyond the bobble head. In Hannes Vilhjlmsson, Stefan Kopp, Stacy Marsella, and Kristinn Thrisson, editors, *Intelligent Virtual Agents*, volume 6895 of *Lecture Notes in Computer Science*, pages 216–227. Springer Berlin / Heidelberg.