A Demonstration of Incremental Speech Understanding and Confidence Estimation in a 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 Overview

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., 2008). In these negotiation scenarios, ideally the virtual humans should demonstrate fluid turn-taking, complex reasoning, and appropriate responses based on factors like trust and emotions. An important component in achieving this naturalistic behavior 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 include relatively straightforward turn management behaviors, like having a virtual human recognize when it is being addressed and turn to look at the user. They could also include more complex responses such as emotional reactions to what users are saying.

Our demonstration is set in an implemented 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 scenario, the goal of the two human role players is to convince Utah and Harmony that Utah, who is currently the local bartender, should take on the job of town sheriff. We presented a substantially similar demonstration of this scenario in (DeVault and Traum, 2012).

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

To support more natural behavior in such negotiation scenarios, we have developed an approach to incremental speech understanding. The understanding models are trained using 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). Every utterance in the corpus is annotated with an utterance meaning, which is represented using a frame. Each frame is 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 at the lower left in Figure 2. Our framework uses this corpus to train two data-driven models, one for incremental natural language understanding, and a second for incremental confidence modeling. We briefly summarize these two models here; for additional details and motivation for this framework, and discussion of alternative approaches, see (DeVault et al., 2011b; DeVault et al., 2011a).

The first step is to train a predictive incremental understanding model. This model is based on maxi-

Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), pages 131–133,
Seoul, South Korea, 5-6 July 2012. ©2012 Association for Computational Linguistics

mum 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). Each partial ASR result 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 tries to output the *complete* frame that a human annotator would associate with the user's *complete* utterance, even if that utterance has not yet been fully processed by the ASR.

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.

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

The incremental ASR, NLU, and confidence outputs 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).

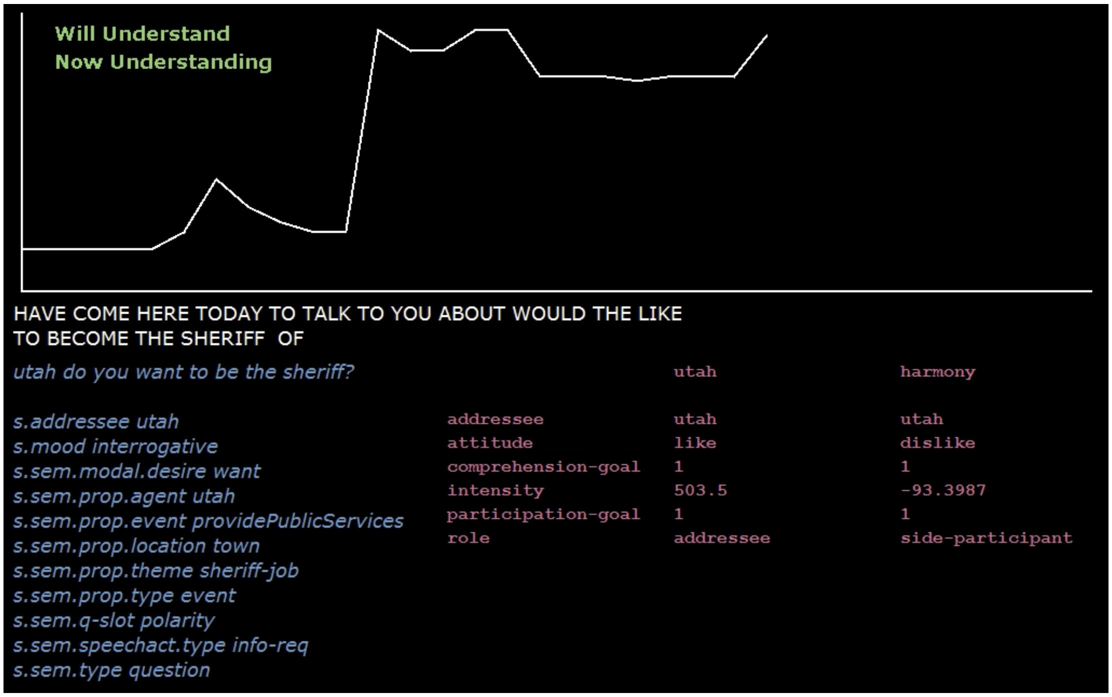
2 Demo script

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. The demonstration includes 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 2.

As the user speaks to Utah or Harmony, attendees can observe the real time visualization of incremental speech processing. Further, the visualization interface enables the demo operator to "rewind" an utterance and step through the incremental processing results that arrived each 200 milliseconds.

For example, Figure 2 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 Now Understanding and also 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



```
HAVE COME HERE TODAY TO TALK TO YOU ABOUT WOULD THE LIKE
TO BECOME THE SHERIFF  OF
utah do you want to be the sheriff?                          utah          harmony

s.addressee utah                   addressee                 utah          utah
s.mood interrogative               attitude                  like          dislike
s.sem.modal.desire want            comprehension-goal        1             1
s.sem.prop.agent utah              intensity                 503.5         -93.3987
s.sem.prop.event providePublicServices participation-goal    1             1
s.sem.prop.location town           role                      addressee     side-participant
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 2: Visualization of Incremental Speech Processing.

and participation goals. Further, Harmony believes she is a side participant at this moment.

Acknowledgments

We thank the entire 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 and David R. Traum. 2012. Incremental speech understanding in a multi-party virtual human dialogue system. In *Demonstration Proceedings of NAACL-HLT*.

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*.

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*.

David Traum, Stacy Marsella, Jonathan Gratch, Jina Lee, and Arno Hartholt. 2008. 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*.

Zhiyang Wang, Jina Lee, and Stacy Marsella. 2011. Towards more comprehensive listening behavior: Beyond the bobble head. In *Proceedings of IVA*.