Beyond Words: Identification of Back-Channel Communication Rules in Arabic and Development of Training Methods

Quarterly Progress Report 3
January 27, 2006 – April 26, 2006
The University of Texas at El Paso
Nigel G. Ward, Principal Investigator
Computer Science, UTEP, 79968-0518
915 747-6827, fax 915 747-5030
nigel@cs.utep.edu

This report describes progress on the project entitled “Beyond Words: Identification of Back-Channel Communication Rules in Arabic and Development of Training Methods”, funded by DARPA’s Defense Sciences Office, via the Department of the Interior under Grant NBCH1050025 to the Information Sciences Institute of the University of Southern California under Subaward PO#105814.

It specifies what we have achieved in the third quarter\(^1\) and what we plan to achieve in the fourth.

1 Specific Tasks from the Project Charter

The following tasks were specified in the project charter.

1.1 Rule Discovery

Our first task was to “discover the basic rules governing back-channel behavior (production of uh-huh, etc. as a display of serious listening) in Eastern Arabic”.

This task is now essentially complete.

In the third quarter we 1. labeled the Iraqi data, 2. experimented with various predictive rules, and 3. discovered that the same rule previously identified works for Iraqi Arabic as well as Egyptian Arabic. This result was summarized in “A Case Study in the Identification of Prosodic Cues to Turn-Taking: Back-Channeling in Arabic”, Nigel Ward and Yaffa Al Bayyari, submitted to Interspeech 2006.

We also presented this work at the 20th Arabic Linguistics Symposium. The work was well received, making us confident that the result is both original and correct.

The Symposium presentation was revised into a full paper: “A Prosodic Feature that Invites Back-Channel in Egyptian Arabic”, Nigel Ward and Yaffa Al Bayyari, and submitted

\(^1\) considering the Project Start Date to be the day the Letter of Authorization was received at UTEP, July 27, 2005

We also did failure analysis, on a sample of cases where our model failed, and determined the limitations of the model. This also is reported in the Interspeech submission. This analysis did not reveal any counter-examples to the central claim.

We archived our collection of dialogs, The UTEP Corpus of Iraqi Arabic, in the Special Collections Department of the UTEP library.

In the next quarter we plan to undertake a few follow-on tasks, as time permits. One is to examine whether the cue discovered is also present in Levantine Arabic, using the newly available LDC corpus. Another is to quantify the pitch up-turn prosodic cue, which accounts for some 30% of the predictions missed by our model, although this is not top priority, since this cue is much less common than the downdash cue, and since this pattern is not unfamiliar to English speakers. A third is to see whether there is a simple way to categorize the kinds of cues and the kinds of feedback.

We may also do additional publishing and publicity related to this result, in conjunction with other work (described below).

### 1.2 Toolset Production

As stated in the project charter, we will “produce a toolset for the automatic discovery of new rules in new languages and cultures”.

In this quarter these activities are not on the critical path.

**Discovery Procedure Development and Documentation**  This quarter we revised our description of our typical workflow, and submitted it as the Interspeech paper noted above.

We also extended our list of desiderata for the toolset.

To test whether our ideas for how to discover prosodic cues are in fact generally useful, we are applying the same method to the problem of back-channel prediction in Spanish. The experience has added to our list of desired toolset features.

In the next quarter we hope to lay some more groundwork for the production of the toolset. One step for this will be to organize our list of desiderata into a more formal specification of the requirements for an improved toolset to support our cue discovery workflow. However, to have maximum utility and impact, toolset specification and production is not something we should do in isolation. First, we may seek partners to help us understand their current workflow and how the tool could be designed to support them. In particular, we may attempt to connect with the Conversation Analysis community, which is a large group of researchers who are very sensitive to dialog phenomena, but who have been poorly served by existing tools. This may require some delicacy, as this community is traditionally proudly non-quantitative. Second, we will survey existing toolsets and then decide whether to continue to extend our toolset or join forces with another team of developers. Incidentally, the work on Spanish may lead to a spin-off publication.
**Alternative Approach** In parallel, we have explored a radically different approach, relying on generic feature detection and machine learning techniques: in particular instance-based learning. This gave quantitative results as good as those obtained by the analysis above. This work was reported in “Prosodic Feature Generation for Back-Channel Prediction”, by Thamar Solorio, Olac Fuentes, Nigel Ward, and Yaffa Al Bayyari, submitted to Interspeech 2006.

In the fourth quarter we intend to examine whether the model resulting from this work is an independent discovery of the same facts noted above, or whether it contains new, complementary features. We also plan to examine whether instance-based learning is effective for generating speaker-independent models.

### 1.3 Training Methods

The charter also calls for us to develop methods for training American speakers to understand and emulate the rules in Arabic, and to apply these training methods in ISI’s Tactical Language Trainer (TLT).

**Skill Builder Module Core Code** We delivered the first version of the basic software infrastructure to USC. This code plays back a recorded conversation to a trainee and scores the user depending on the accuracy of the productions. The original conversation is a two-track recording of two subjects talking on the telephone; to use the original recording in the trainer, the track containing the productions to be emulated by the trainee is removed from the original conversation and the remaining track is used as a cue track in the Module. The productions to be emulated contained in the original conversation are manually labeled and a file containing the backchannel points is created; this file, along with the single-track conversation are fed to the Module. The Module plays back the conversation to the trainee by means of a headset and assumes that a production has been made by the user when the microphone input level exceeds a certain threshold. Each trainee production time is compared to the manually labeled times contained in the backchannel points file. If a production time is near a labeled time, the production is counted as a “hit”. Otherwise, it counted as a “miss”. The final score, computed from the numbers of hits and misses, measures the ability to be a good listener in Arabic: specifically, it is higher to the extent a student produces more back-channels at appropriate places and/or fewer in inappropriate places.

**Pedagogy Design** In this quarter we documented the pedagogical features probably needed and the user interface functionality to provide in a Mission Skill Builder Module to teach back-channel behavior, “Back-Channel Skill Building Module, Version 1.0” (internal document, January 31, 2006).

**First Set of Experiments** We excerpted a few minutes of dialog from our Iraqi data, containing 20 opportunities to BC, and imported this into the module.

We ran 3 subjects, students from the freshman CS course, and quickly discovered that a simple explanation of the rule was not adequate to improve their score over two trials. After training they did tend to produce more back-channels, but not where they should have.
Second Set of Experiments  We then made two changes to the code. First, we loosened the window of opportunity from 700 milliseconds to 2000 milliseconds. We want to count learners as correct if they successfully produce back-channels in about the right places, and not penalize their score if they are a little slower than a native speaker would be. The second change was a refinement to the scoring calculation. Users having a high ratio of hits to misses were given a low total score because the scoring procedure caused a miss to severely reduce the score. As a result, we adopted a looser scoring calculation, based on the F-measure.

We also added a native Arabic tutor into the procedure, contributing in two ways. First, she enhanced the training phase: rather than just telling them the rule, she explained it and illustrated with her own voice. Second, she emulated the behavior of the training module, by telling a story in Arabic and encouraging the learners to follow along and back-channel. For consistency, this story was done by reading aloud part of a transcript from the corpus.

Thus the procedure was as follows:

1. In the first step, the student is instructed to use the tool and back-channel using his/her common sense.
2. Then the student talks to the tutor and is told to listen to her and back-channel when appropriate. This conversation is also recorded for future analysis.
3. Next we explain to the student the prosodic cue we have found for Arabic: the downward slope of the pitch. We provide different examples to the student of how a downward pitch sounds and instruct him/her to focus on listening to the intonation of the utterance.
4. Once the student is confident that he/she can detect the aforementioned prosodic cues, the student uses the tool again.
5. Finally, the student interacts again with the tutor and tries to back-channel with her.

We then ran 6 more subjects, 5 females and 1 male. Judging by the tool-based scores, 3 improved, 2 showed no significant change, and 1 did worse. Judging by their performance with the human tutor, again 3 improved, 2 showed no significant change, and 1 did worse. However there was no correlation between their improvements as measured by the techniques: some did better with the tutor, some with the tool. We see the evidence of learning as promising. Even if only half the learners manage to acquire this skill, that is of value. Whether the skill is retained remains to be determined.

We are concerned, however, that learners who didn’t improve may feel perplexed and demotivated, which is a potential problem in the TLT context.

While 3 students were fairly willing to produce backchannels, 3 were fairly reluctant, never taking up more than a third of the opportunities. This subpopulation was consistent across both sessions with the tutor and both sessions with the tool. This may be a personality issue. However by the last two sessions all subjects were responding to at least a quarter of the opportunities, which may be enough in practice to give the impression of being an attentive listener.

We had expected learners to do better with the human tutor than with the tool, but that was not seen.
We observed a tendency for students to produce few misses (few inappropriate back-channels), especially when working with the human tutor. This was not seen in the first experiment, when some students ‘went wild’ and back-channeled in many wrong places. This suggests that when foreigners trying to back-channel in Arabic either produce them in their correct places (with some delay) or they just miss them, at least when in the presence of a native speaker.

The tutor also noticed that most of the correct back-channels were produced after cues which co-occurred with pauses. This suggests that clear pauses help non Arabic speakers recognize back-channel opportunities more easily.

**Next Quarter**  Improving the training code and methods is now our highest priority.

1. We plan to improve the code in two ways. First, we suspect that some learners may benefit from a visual cue to draw attention to the timepoint when the prosodic pattern appears; we will add such a visual cue. Second, we will improve the back-channel detection code, which is currently sometimes confused by rattling of the earphones of the Plantronics microphones when the user moves his head. We plan to deliver this improved version to ISI by mid-May.

2. We plan to discover what the human tutor is adding, and provide that in the tool. This may be more time-on-task, better explanations, better audio examples, exaggerated prosodic cues, longer pauses, facial or other gestures, a sense of seriousness, or immediate feedback to the learner. We plan to make the necessary changes to the tool and deliver it to ISI by the end of May.

3. We plan then to do another round of formative experiments, and to revise the tool as needed. Hopefully we will also have feedback from ISI.

4. Then we will do controlled experiments, to formally evaluate learnability using this tool. The value of this tool as compared to other teaching methods will also be measured.

5. In the fifth quarter we expect to write-up the results, probably for some journal focusing on CALL (Computer-Aided Language Learning).

6. As soon as ISI is ready to work with us, we will assist them with the integration of the tool into TLT.

**1.4 Other Tasks**

Regarding the other Year 1 tasks listed in the project charter:

In the fourth quarter we will design an experiment that will quantify the importance of following the right rules, with the experiment itself to be done in Year 2. This is our third priority. It may be possible to combine this with the experiment on misinterpretations discussed below.

In the fourth quarter we may build a system able to emulate the back-channel behavior of a native Arabic speaker. This is probably not of major importance for TLT, as noted in the second quarterly report, and may be anyway be hard to port on Windows, but it should make an interesting demo, and may be useful for various experiments.
2 Additional Tasks

Five additional tasks were noted in previous quarterly reports.

2.1 Dialog Pacing in Direction-Giving

As noted in the first quarterly report, we want to extend TLT to make learners able to actively listen and control the pace of information delivery by Iraqi conversants. In the next quarter we plan to label this data and begin analysis (not critical path).

2.2 Endpointing

As noted in the first quarterly report, an understanding of how turn-ends are signaled in Iraqi Arabic would make the Trainer more compelling and also allow the teaching of more turn-taking skills to users. In the next quarter we may determine exactly what is needed here, and begin analysis.

2.3 Back-Channel Detection

As noted in the first quarterly report, TLT could benefit from a swift low-level detector for user-produced back-channels. We may build this in the fourth quarter, although this is not high priority.

2.4 Gesture Analysis

For purposes of better understanding the cues to back-channels, and for the sake of preparing video clips to use in our module and/or the Mission Rehearsal exercise, we plan to look at the video records to determine if there are gestures or postures which are synergistic or antagonistic to the prosody-based back-channel cues. This is our fourth priority for this quarter.

Incidentally, in the third quarter, we donated a set of DVDs to the Institute for Creative Technologies, where David Traum and Bilyana Martinovski are also interested in modeling Iraqi gestures.

2.5 Inter-Cultural Misinterpretations

As noted in the second quarterly report, the fact that Americans can find Arabic grating may be due in part to the prosodic features we have uncovered.

In the fourth quarter, design an experiment to verify that this is true and determine the magnitude of the effect. We will probably create some audio clips and ask both native Arab speakers and Americans to describe both the pragmatic and emotional impressions they get from these utterances. Hopefully we will be able to design the experiment in such a way as to simultaneously server two other purposes: to provide additional confirmation of the
back-channel cue-ness of downdash; and to quantify the value of following these rules for Americans listening to Arabs.

Alternatively, an experiment on this topic as part of evaluation of our Mission Skill Builder Module, that is, to determine if that experience has the side-effect of making learners understand the speech patterns relating to back-channels and turn-ends in Arabic do not actually bear the emotional load they appear to.

These studies could be high impact and may be easy to do, so they are our second priority.

3 Other Activities

We made friends with two Arabic language researchers who expressed a willingness to help with experiments if necessary: Professor Samira Farweneh of the University of Arizona in the Department of Near Eastern Studies and Professor Fred Hoyt of the UT Austin Linguistics Department.

4 References

Project documents and software are now available at a private URL: http://www.cs.utep.edu/nigel/dso/.