American and Arab perceptions of an Arabic turn-taking cue

Nigel G. Ward, Yaffa Al Bayyari

University of Texas at El Paso, 79968 USA

Abstract

Languages differ in the way that speakers coordinate their interaction moment-by-moment, and this can cause intercultural misunderstandings. We explore this in the domain of listening behavior. One way that listeners show interest and attention is by producing back-channel feedback (short utterances such as *okay* and *hmm*) at appropriate times, and these times are determined, in part, by the interlocutor, who signals when such feedback is welcome with various cues. In Arabic these cues include a prosodic feature in the form of a steep continuous drop in pitch. This paper shows that English speakers can misinterpret this, perceiving it as an expression of negative affect, and that this tendency can be substantially alleviated by training.

Keywords: cross-cultural interaction, dialog, listener behavior, prosody, back-channel feedback

Some social behaviors can carry an emotional punch, whereas others seem mundane. In intercultural interaction, if the participants don't know which are which, then small misinterpretations can lead to deep misunderstandings. On the one hand, the mechanics of interaction in dialog seem at first glance to fall in the mundane category, and indeed "interaction appears to have detailed universal properties ... for a wide range of features, from turn taking [to] ... greetings ... the languages and cultural systems that have been studied reflect very similar ... subsystems" (Levinson, 2006). On the other hand, it also appears that "the speakers' mutual judgments of abilities and intentions are profoundly influenced, if not determined, by automatic uses of the nuts and bolts of language — pitch and amplitude; intonational patterns; pacing and pausing ..." (Tannen, 2005). Although

in this Brief Report we cannot properly survey work on this topic, there is a long history of work identifying interactional behaviors that have been or can be misinterpreted. Indeed, the details of conversation management and non-verbal behavior — including voice quality, pitch, loudness and timing — have been a frequent topic of study in the fields of cross-cultural and intercultural communication (Street, Jr. & Hopper, 1982), especially in the tradition of the ethnography of speaking and communication. In addition, a good deal of work has applied the methodology of conversation analysis, in which turn-taking cues loom large, to intercultural or intergroup encounters. In this paper we apply experimental methods to the study of cross-cultural interpretations of the "nuts and bolts."

One salient aspect of turn-taking is back-channeling. Back-channels, also sometimes known as "response tokens," "reactive tokens," "minimal responses," and "continuers," are short utterances produced while the interlocutor has the turn. In English, common back-channels include *uh-huh*, *yeah*, *hm*, *right* and *okay*. As a common way in which listeners shows that they are listening, back-channels are pervasive in many languages, occurring on average once every 12 seconds or so in casual English conversation. Although back-channeling is seldom salient — indeed it seems to happen largely below the level of conscious awareness — it is important both for the successful communication of content and for the development of rapport (Gratch, Wang, Gerten, Fast & Duffy, 2007). Although the production or non-production of a back-channel is ultimately up to the listener, there are places in dialog where back-channels are especially welcome, and these can be marked by a prosodic cue from the speaker, as shown by statistical studies of dialog corpora and by experiments in which subjects back-channel to a pre-recorded track; references appear in (Ward & Al Bayarri, 2008).

Arabic and English are two languages which are genetically unrelated but which are increasingly in contact. High-level cultural differences between the two speaker populations are well known, and these can explain some differences in discourse patterns and account for

many misunderstandings, but the possible effects of differences in the low-level mechanics of interaction have not previously been explored. Semantically and pragmatically, back-channels in Arabic appear to bear the same basic functions as in English (Havez, 1991), and the overall frequency of back-channeling is similar between the two languages. In both languages the production of certain prosodic cues by the speaker is strongly associated with the subsequent production of a back-channel response by the listener. although the specific cues involved are different. In English (at least in parts of North America) the main cue is a region of low pitch (Ward & Tsukahara, 2000). In Arabic (at least in parts of Egypt and Iraq) a common cue is a prosodic feature complex which includes a steep pitch downslope; we will refer to this as a "downdash" (Ward & Al Bayyari, 2007, 2008).

We therefore hypothesized that the pitch downdash as a back-channel cue would be perceived differently by Arabs and naive Americans. A further question arose from the first author's impression, as an English native speaker ignorant of Arabic, that the downdash often sounded like an accusation or expression of resignation. Wondering whether this misinterpretation would be common, we hypothesized that this prosodic pattern would be perceived negatively by Americans but not by Arabs.

Methods

There were 3 groups of 18 participants each. The Arab subjects were native Arabic speakers, speaking a variety of dialects, but all familiar with Egyptian Arabic speech patterns. Seven were living in the United States and 11 in Qatar. All were recruited by word of mouth and compensated with $20. The other two groups of subjects were recruited from Introduction to Computer Science classes, and were compensated either with class credit or $10. Subjects were recruited without regard to linguistic background, however we kept data only from those who had been using English since at least age 16 and were judged by the experimenter to have good English dialog skills. Five datasets were accordingly discarded and five additional subjects recruited. Even so, most of the subjects were Spanish-English

bilinguals; it is therefore worth noting that the back-channel cuing feature-complex in Spanish is different again from those in Arabic or English, but the overall frequency, pragmatic contexts, and typical time from cue to response are similar across all three languages (Rivera & Ward, 2007). One group had earlier been trained in the Arabic listening skill of responding to pitch downslope cues with back-channels, as part of an evaluation of the effectiveness of a 25-minute software-based training experience suite (Ward, Escalante, Al Bayyari & Solorio, 2007). These subjects were tested two to three months after this training experience. The other group had no such training. All subjects gave informed consent.

Stimuli

To isolate the effects of the downdash, the stimuli were built up by splicing a single lead-in phrase with one of three endings: the downdash and two controls. The lead-in was chosen by the the second author, a native speaker of Arabic, as a fragment without pragmatically or emotionally salient pitch movements and sounding natural when spliced with each of the prosodic cues. The endings were chosen to be clear yet typical examples of the pitch downdash and of the two control patterns: the first a downward staircase of three flattish pitch regions, that is, a "cadence" pattern, tentatively identified as an indication of finality and turn yield; the second a pitch pattern used when giving one item in a list while indicating that there are more to come. The stimuli were derived from audio fragments extracted from dialog AR_4023_1_pt1 in the LDC corpus of Egyptian Arabic telephone conversations, with the exception of the pitch downdash itself, which was from our Iraqi Arabic corpus. To ensure that listener judgments were based on the pitch patterns alone, uninfluenced by lexical information, all fragments were resynthesized using Praat to discard the segmental and volume information. The pitch points used to specify the contours for resynthesis were extracted automatically. The resulting stimuli sounded, for example, like *babaa-ba ba-bababa*, but were recognizably human.

The ending fragments were scaled up in pitch by factors sufficient to make the splicing

unnoticeable. The pieces were assembled using Reaper. The quality of the resulting stimuli was evaluated by a second native speaker: although all were somewhat unnatural due to the resynthesis, none was perceived to be particularly bad. The stimuli were also evaluated by a question asking the Arab subjects to rate how much the speaker sounded "like a native Arabic speaker" on a 5-point scale. The stimulus ending in an upturn was rated 4.1, in downturn 3.5, and in cadence 3.3, but these differences do not correlate with the judgments reported below, so it is unlikely that the results were artifacts of synthesis problems.

The first experiment aimed to probe impressions of the discourse function of the downdash, so the 3 stimuli were presented followed by one of 3 types of response: a back-channel, a full turn, or silence, the judgments of interest being those regarding the stimuli in which a back-channel response came after an utterance ending in a downdash. The pause between the end of the cue and the start of the response was 500ms, a typical value in the corpus both for the time-gap between a pause onset and a back-channel response, and for the time-gap between utterances at a turn hand-off. Since pilot studies had shown that repeated listening to the full 16-second clips was tiring, subjects were also provided with abbreviated 6-second versions containing just the cue and the response parts. These short fragments were intended to make it easier for subjects to focus on the transition between the turns when making judgments. The second experiment aimed to examine affective impressions, so the stimuli were just the lead-in spliced with one of the three ending patterns, presented without the response. All stimuli are available at our website, http://www.cs.utep.edu/nigel/abc/.

Procedure

The subjects were told that "we are interested in patterns of dialog in various languages, especially Arabic and English. This study is about which patterns of interaction are preferred or disliked by various people." After the demographic survey, subjects were told they would be "listening to dialog fragments which have been modified so that you can't identify the speakers or their words." For familiarization with this kind of stimulus, they initially listened

to two English dialog fragments and then their resynthesized versions, to give them a sense of how such sounds related to unmodified speech. All samples were provided on a laptop computer screen and played through headphones in stereo. Subjects were able to adjust the volume and were allowed to proceed at their own pace, being able to listen to the stimuli any number of times and in any order. We encouraged them to listen repeatedly to stimuli until they were comfortable making their judgments. Afterwards we asked various questions, which for the American subjects included questions about exposure to and opinions about Arabs and Arabic. The order in which samples appeared was balanced across subjects. These orders were fixed beforehand and each group of participants saw the same orders — Arab subject number n was given the stimuli in the same order as naive subject number n, and so on — to enable matched-pairs analysis.

The first experiment asked for judgments of "a positive or a negative feeling about the second speaker," on a scale from 1, "very negative," to 7, "very positive," (eschewing more sophisticated measures (Bradac, 1990) because of the number of judgments to be made and the nature of the stimuli.) Before the 9 stimuli of interest they were presented with two English-derived examples, one with a normal pattern of back-channeling and one with a badly delayed back-channel. The value of the rating for the first sample was used for normalization, to compensate for any response bias due to tendencies for some groups to be overall more positive than others. Both the Arab and the Arabic-naive groups rated the benchmark stimulus 5.06 on average and the Arabic-exposed group rated it 5.28 on average, so the results below are those given by subtracting .22 from all ratings by the exposed group.

Asked whether making the judgments was difficult, most subjects said that it was, to varying degrees, and mostly because the words were masked. Three subjects indicated that the downslope and cadence patterns sounded the same to them. No subject reported difficulty perceiving the stimuli as modified versions of two people talking.

The second experiment was done immediately after the first. Subjects were asked to

"listen to three audio fragments from one speaker and try to infer their emotional state" for each. Ratings were again on a 7-point scale. Before the actual stimuli, subjects were given two emotional American English utterances, one happy and one angry, both in original and resynthesized form, to provide a benchmark. For each sample, subjects were asked to "write two or three adjectives describing the feeling (sad, angry, happy, surprised, scared, disgusted, etc.)." Considering only those who labeled the benchmark as "happy" and included no other term, the average for the naive Americans was 6.50, for the Arabs 6.67, and for the Arabic-exposed Americans 6.33, and the results given below have been normalized accordingly.

Results

Both hypotheses were supported. In the first experiment the back-channel response to a downdash ending was ranked relatively highly by the Arabic speakers, averaging 4.7, but lower by the naive Americans, averaging 3.7, and this difference was significant, $t = 2.36, p < .02$, with an effect size of 0.9 standard deviations (matched-pairs, one-tailed t-test with 17 degrees of freedom; effect size in terms of Cohen's d, computed using the pooled standard deviation of all judgments of the stimulus; here and throughout). In the second experiment Arabs overall ranked the downdash pattern as neutral (average 4.0) and the naive Americans as slightly negative (mean of 3.2, significantly less than neutral by the z-test, $z = -2.55$, $p < .05$), and the difference was significant, $t = 2.02, p < .05$; effect size 0.7.

The judgments of the exposed Americans were closer to those of the Arabs. They rated the downslope+back-channel pairing 4.5 on average, close to the rating by the Arabs and significantly different from that of the naive Americans, $t = 2.54, p < .02$; effect size 0.7. They did not perceive a negative affect in the downslope cue; rating it 4.0 on average, neutral and close to the Arab ratings, but significantly different from the ratings of the naive Americans, $t = 2.12, p < .05$; effect size 0.7. However the exposed Americans rated all samples higher than did the naive Americans, so this effect may be partly due to a general tendency for exposure to Arabic to reduce negative perceptions.

One might ask whether a general prejudice against Arabs could explain the results. This is unlikely, as the responses to the survey questions indicated that the Americans had little knowledge of or opinions regarding Arabs or Arabic, and because the tendency for the Arabic-naive group to rank the stimuli lower overall was not an undifferentiated dislike – in fact, in the first experiment they rated two of the nine pairings more positively than did the Arabs, and in the second experiment they rated one of the controls more positively than did the Arabs — so it seems that this prosodic feature specifically is being felt to convey a negative affect. More details on this and the other results appear in (Ward & Al Bayyari, 2008).

Discussion

The generalizability of the findings is clearly limited, pending *in vivo* experiments and experiments with other subject populations, including monolinguals. However, the results indicate that the pitch downdash is perceived as a cue to back-channels in Arabic, but that it is not perceived as such by naive Americans (Experiment 1). Thus the interpretation of this cue is indeed culture-dependent rather than universal. The results also show that this cue can be misinterpreted by naive Americans as expressing negative affect, and that, perhaps surprisingly, a small amount of training suffices to prevent this misperception (Experiment 2). The mechanics of turn-taking may seem harmless and mundane: unlike gesture or emotional expressions, people in intercultural encounters may not expect these to be a source of misunderstandings. However these results show that misperceptions of turn-taking cues can occur.

This finding resembles those of studies showing that the ability to identify emotions from the prosody of a speech sample is weaker for speakers of other languages and members of other cultures, despite universal tendencies (Elfenbein & Ambady, 2002). The novelty here is in studying a turn-taking use of prosody. It would be convenient, both scientifically and practically, if there were a clear distinction between the emotional and linguistic uses of prosody, that is, between the interactionally significant and the mundane (and within a

single language such a distinction may exist, maybe even down to the neural level (Grandjean, Banziger & Scherer, 2006)). However these results indicate that, cross-culturally, the distinction is blurred.

This finding has at least two practical implications. First, Americans needing to interact with Arabs probably ought to be taught the meaning of this prosodic pattern, both to give a good impression as a good listener (Ward & Al Bayyari, 2008) and to avoid misunderstanding. This is probably equally true whether the words used to communicate are English or Arabic, as speakers of foreign languages often retain their native turn-taking patterns. Second, the function of this pitch pattern probably ought to be known even by those who merely overhear Arabs speaking. Many radio interviews with speakers of foreign languages start the translation voiceover a few seconds after the speaker has started, leaving the audience to gauge the speaker's personality and emotional state from those few seconds of speech. For interviews with Arabic speakers, a common place to start the voiceover is after the first pause, which often happens to be a place where the speaker is welcoming a back-channel by means of a pitch downslope. Along with other properties of the Arabic language, the pitch downslope could be leading to systematically mistaken impressions of Arabic speakers.

References

Bradac, J. J. (1990). Language attitudes and impression formation. In H. Giles and W. P. Robinson (Eds.), *Handbook of language and social psychology* (pp. 387–412). Hoboken, NJ: John Wiley.

Elfenbein, H. A. & Ambady, N. (2002). On the universality and cultural specificity of emotion recognition. *Psychological Bulletin, 128*, 203-235.

Grandjean, D., Banziger T., & Scherer K. R. (2006). Intonation as an interface between language and affect. *Progress in Brain Research, 156*, 235-268.

Gratch, J., Wang N., Gerten J., Fast E., & Duffy R. (2007). Creating rapport with virtual agents. In *IVA 2007, LNAI 4722* (pp. 125-138). Berlin: Springer.

Havez, O. M. (1991). Turn-taking in Egyptian Arabic: Spontaneous speech vs drama dialogue. *Journal of Pragmatics, 15*, 59–81.

Levinson, S. C. (2006). *On the human 'interaction engine.'* In N. J. Enfield & S. C. Levinson (Eds.), *Roots of human sociality* (pp 39-69). Oxford: Berg.

Rivera, A. G. & Ward, N. (2008). Prosodic cues that lead to back-channel feedback in Northern Mexican Spanish. In *Proceedings of the seventh annual high desert linguistics society conference* (pp. 19-26). Albuquerque, NM: University of New Mexico.

Street, Jr., R. L. & Hopper R. (1982). A model of speech style evaluation. In E. B. Ryan & H. Giles (Eds.), *Attitudes towards language variation* (pp. 175-188). London: Edward Arnold.

Tannen, D. (2005). Interactional sociolinguistics as a resource for intercultural pragmatics. *Intercultural Pragmatics, 2*, 205–208.

Ward, N. G. & Al Bayyari, Y. (2007). A prosodic feature that invites back-channels in Egyptian Arabic. In M. Mughazy (Ed.), *Perspectives in Arabic linguistics XX* (pp. 186–206). Amsterdam: John Benjamins.

Ward, N. G. & Al Bayyari, Y. (2008). Additional information about perceptions of an Arabic turn-taking cue (Tech. Rep. No. UTEP-CS-08-34). El Paso, TX: University of Texas at El Paso, Department of Computer Science.

Ward, N. G., Escalante, R., Al Bayyari Y. & Solorio T. (2007). Learning to show you're listening. *Computer Assisted Language Learning, 20*, 385–407.

Ward, N. G. & Tsukahara W. (2000). Prosodic features which cue back-channel respones in English and Japanese. *Journal of Pragmatics, 32*, 1177–1207.

Author Note

We thank Bill Lucker, David Novick, Maissa Khatib, Rafael Escalante, Lewis Johnson, Ralph Chatham, Marisa Flecha-Garcia, and Cindy Gallois. This work was supported by the Department of Defense and its Advanced Projects Research Agency.