

A Prosodic Feature that Invites Back-Channels in Egyptian Arabic

Nigel Ward and Yaffa Al Bayyari
University of Texas at El Paso

Abstract

One of the signs of listening attentively and supportively is occasional back-channel feedback, small utterances produced by the listener while the speaker continues his turn. To do this appropriately it is necessary to understand when back-channels are and are not welcome. In Egyptian Arabic, times when the listener is especially welcome to back-channel are indicated by various prosodic features produced by the speaker, including a steep pitch downslope. This particular feature contrasts with the downward pitch staircase (Kadenz) characteristic of turn-yields. This finding is based on qualitative and quantitative analysis of the contexts of occurrence of 660 back-channels in 168 minutes of Egyptian Arabic telephone dialogs from the Callhome corpus.

Note: reflecting new developments, we have improved the paper slightly; moving section 8.2 into section 9, and rewriting and adding to sections 9.2 and 9.5.

Corresponding Author:

Nigel Ward

nigelward@acm.org

phone: 915-747-6827

fax: 915-747-5030

<http://www.cs.utep.edu/nigel/>

Computer Science, University of Texas at El Paso

El Paso, TX 79968-0518

February 14, 2007

⁰Acknowledgements: We thank Thamar Solorio, W. Lewis Johnson, Jon Amastae and the participants of the 20th Arabic Linguistics Symposium for discussion. This work was supported in part by DARPA and in part by the National Science Foundation under Grant No. 0415150.

A Prosodic Feature that Invites Back-Channels in Egyptian Arabic

1 Back-Channeling as a Dialog Skill

To be a good listener you have to be able to show you're listening. In dialog this includes the active display of attention, interest, understanding and/or willingness to let the other person continue. This is accomplished in part with back-channels, also known as "minimal responses" and "continuers": the short utterances produced while the interlocutor has the turn. In English these are typically utterances such as *uh-huh*; in Egyptian Arabic the most common back-channels are *ah*, *mmm*, laughter, *tayeb*, and *aiwa*.

This raises the question of how a listener can know when it is appropriate to produce a back-channel. Work in other languages (Yngve 1970; Ward & Tsukahara 2000; Fujie *et al.* 2005) suggests this depends on both speaker-related factors and listener-related factors. That is, the listener is free to produce back-channels based on his own understanding and intentions, but these back-channels are especially welcome at certain times in the dialog, and these times are determined by what the speaker is saying and how he is saying it. Further, these times are indicated in part by prosody: in several languages there is a prosodic cue that a speaker can use to indicate when he welcomes a back-channel from the listener.

The present study was motivated by the desire to be able to teach Arabic back-channel skills to non-natives, specifically by extending an intelligent tutoring system, the Tactical Language Trainer (Johnson *et al.* 2005). The motivating problem is that a second language learner who lacks turn-taking skills, even if a master of the vocabulary and grammar, can easily appear uninterested, ill-informed, thoughtless, discourteous, passive, indecisive, untrusting, dull, pushy, or worse. Indeed, our earlier study of Japanese and English back-channel behavior showed that not only do Japanese back-channel twice as often as Americans (Maynard 1989), but that the interval between the prosodic cue from the speaker to the back-channel response by the listener was typically only half as long in Japanese (Ward & Tsukahara 2000). The potential for awkward intercultural interactions here is clear. Unfortunately the rules governing turn-taking are seldom taught to language learners, largely because they are not known. This has been the case for Arabic.

This paper describes the initial identification of a prosodic cue in Egyptian Arabic that indicates to the interlocutor when back-channel feedback is especially welcome, and which makes it statistically more likely that the listener will indeed produce a back-channel in response.

2 Prosody and Turn-Taking in Arabic

Research findings relating to our question are found in two areas: turn-taking and prosody.

Back-channel behavior is an aspect of turn-taking, that is, the way that speakers in dialog manage their interactions to allow smooth exchanges and minimize awkward silences and interruptions. In Arabic, the only work on this is that of Hafez (1991), which provides a useful taxonomy of the ways in which speakers manage turn-taking in Egyptian Arabic. Hafez further identifies lexical discourse markers which often accompany turn taking and turn yielding, however unfortunately not for back-channeling. In his brief discussion of back-channels, Hafez provides examples of their semantic and pragmatic functions; these appear to be similar to those seen in other languages (Ward & Tsukahara 2000). In particular, some back-channels do not display understanding, but merely attention. Hafez also notes that back-channels can occur not only in "slots" (places where the interlocutor is momentarily silent) but also "in overlap" with the other speaker's turn; again this is also seen in other languages.

The second relevant body of work is that on the prosody of phrases, sentences and utterances. This reveals prosodic features which express pragmatic functions intimately related to turn-taking, including utterance type distinctions and expressions of completion and finality.

Regarding the prosodic correlates of different sentence types, the basic facts are that statements typically have a falling final pitch contour (Kulk *et al.* 2005; Eldin & Rajouani 1999; Rifaat 2005), as do wh-questions (Kulk *et al.* 2005; El-Hassan 1988). Yes-no questions generally exhibit a pitch rise (Eldin & Rajouani 1999; El-Hassan 1988).

Also of interest are the prosodic correlates of finality and completion. Rifaat (2005) observes that a pitch rise (a “[LH#]” tone) occurring “tune-medially” at a “phrase boundary” can indicate “non-finality” in Cairean Arabic. His example shows this occurring in the enumeration of a list, but a similar pattern can appear at phrase boundaries; this perhaps serves the same function as the “comma intonation” contour of English. Similarly both Kulk *et al.* (2005) and Corvetto (1982) observe that in Damascene Arabic non-final utterances have a level or rising contour, in contrast to the falling contour characteristic of final utterances.

A related phenomenon is that described by Bergsträsser as a “Kadenz” (cadence), indicating semantic completion in Damascene Arabic (Bergsträsser 1968). Cadence is a musical term, referring to a sequence of chords at the close of a piece of music, where the chords are typically sustained for a whole note, with a drop of two semitones seen between the two chords. This stands in contrast to the general pattern of pitch progressions in Arabic: where pitch glides are far more common than pitch jumps, and sustained level pitches are rare. Bergsträsser further observes that vowel lengthening can co-occur as part of a final Kadenz, and Kulk *et al.* also observe vowel lengthening pre-pausally, although de Jong and Zawaydeh observe that vowel lengthening may be less in Jordanian Arabic (de Jong & Zawaydeh 1999).

At this level of description there seems to be general agreement, perhaps surprisingly so, given that these findings reflect study of a wide variety of discourse types, dialects, speaking styles, elicitation methods, analysis methods, and theoretical frameworks.

3 Corpus Preparation

There are many ways to approach the prosody of turn-taking. For this project a corpus-based analysis was adopted, for two reasons. First, back-channels are intrinsically a dialog phenomenon and so they can only be observed in dialog. Second, cause-effect relations at this time scale are not introspectable and so must be studied empirically.

3.1 The Corpus

The corpus used was the CallHome Corpus of Egyptian Arabic Speech (Canavan *et al.* 1997). This is a collection of “unscripted telephone conversations between native speakers of Egyptian Colloquial Arabic . . . calls . . . originated in North America and were placed to locations overseas (typically Egypt). Most participants called family members or close friends”.

As a manageable-sized subset of the dialogs, we chose, pseudo-randomly, the first five dialogs from each of the three CDs. Specifically, this subset consists of the first 32 minutes of dialog AR_4023, the first 7 minutes of AR_4367, and the first 10 minutes of each of AR_4150, AR_4194, AR_4213, AR_4283, AR_4297, AR_4299, AR_4392, AR_4419, AR_4931, AR_4949, AR_4950, AR_4981, and AR_4985. This gave us 168 minutes. Later we realized that one of dialogs had severe line noise, but we decided against retroactively excluding it.

These conversations were mostly between two people but a few contain parts where 3 or 4 people are active. Most speakers were adults, both women and men. Common topics were school, business matters, family health issues, financial transactions, babies, and general gossip and chit-chat. Dialog activities included openings and closings, telling family news, talking about work and school problems,

Token	Gloss	Occurrences
<i>ah</i>	yeah	234
<i>mmm</i>		147
laughter		76
<i>tayeb</i>	alright	28
<i>aiwa</i>	uh-huh	18
<i>aah</i>	yeah	16
<i>yeah</i>		16
<i>ha-</i>		14
<i>ah ah</i>	yeah yeah	12
<i>mmm-hum</i>		11
<i>mashi</i>	okay	11
<i>yah</i>		9
<i>kowayes</i>	good	7
<i>aih</i>	yeah	7
other		54
Total		660

Table 1: Number of Occurrences of Various Tokens as Back-Channels. Non-lexical and English-derived tokens are not glossed.

invitations, persuasion, planning and making commitments, defending past actions, and joking, among others. Two speakers appeared to be not truly fluent in Arabic, and some English back-channels and back-channeling patterns were observed.

3.2 Identifying Back-Channels

A native Arabic speaker (the second author) labeled all back-channels in this subset, giving 660 occurrences.

Back-channels were labeled according to a standard definition: To count as an instance of back-channel feedback, an utterance had to meet three criteria: (D1) respond directly to the content of an utterance of the other, (D2) be optional, and (D3) not require acknowledgement by the other. Other definitions of back-channel exist, but in practice these all delimit roughly the same set of phenomena (Ward & Tsukahara 2000). Although designed for two-party dialogs, this definition also worked for the multi-speaker segments of the dialogs: for example when one speaker made a statement and two listeners gave him back-channel feedback.

Note that in many cases the back-channel overlapped or interrupted the utterance of the main speaker; in the examples below this was salient in Examples 10, 12 and 13.

Back-channels were identified based on the audio only; transcriptions were added later, if at all.

Table 1 shows the most frequent tokens used as back-channels. There was great variety: 115 different types. Note that this is a rough, broad categorization, which obscures subtle phonetic variations and prosodic variations, although in fact such differences are likely to be pragmatically significant (Ward 2006; Ward 2004).

Back-channels were fairly frequent, about 4 per minute. Although direct comparisons are not possible, this seems to be roughly comparable to the frequencies reported for English and many other languages, and less frequent than in Japanese (Maynard 1989; Ward & Tsukahara 2000). Back-channels were not, however, evenly distributed. In particular, they were rare at the starts of the dialogs. This is probably for two reasons: first, most calls started with questions and answers (regarding e.g. family and health), and of course back-channels do not appear in such contexts, and second at the beginning most speakers were somewhat uncomfortable, in part because they knew they were being recorded.

There also seemed to be gender differences in back-channel use: the women generally seemed to use fewer back-channels and their back-channels seemed to be shorter in duration than the men's. Large individual differences in interaction styles were also observed; for example one dialog was completely lacking in back-channel behavior,

3.3 Examples

This subsection presents some examples of clear cases of back-channels (in bold), showing how they appear in context. The context is given in English and the back-channel itself in Arabic with a gloss in parenthesis, unless the back-channel was an English word or a non-lexical utterance.

speaker1:	<i>Last week I called your mom and she told me about your wedding</i>	1	
speaker2:	mmm	2	(1)
speaker1:	<i>but I told her if it was any country other than Russia I would go</i>	3	

speaker1:	<i>Dr. Malek gave me yesterday details of a cheap hotel for the reservation</i>	1	
speaker2:	mmm-hum	2	
speaker1:	<i>so I called and made the reservation yesterday</i>	3	(2)
Speaker2:	kowayes awi (<i>very good</i>)	4	
Speaker1:	<i>and he welcomed me as he'll be staying in the same hotel</i>	5	

speaker1:	<i>Yehya got married</i>	1	
speaker2:	aah (<i>yeah</i>)	2	(3)
speaker1:	<i>and I was very glad to meet Amrawi at the wedding</i>	3	

speaker1:	<i>They want people to speak in these languages and they record sample text without saying who is speaking</i>	1	
speaker2:	ah ah (<i>yeah yeah</i>)	2	(4)
speaker1:	<i>then they apply computer analysis to the data</i>	3	

speaker1:	<i>The tickets will cost us \$2000-\$2500</i>	1	
speaker2:	yah, ah (<i>oh, yeah</i>)	2	(5)
speaker1:	<i>this is just for the tickets because the baby needs a separate seat</i>	3	

3.4 Borderline Examples

Although the identification of back-channels was unproblematic in the vast majority of cases, there were 17 examples where the classification was difficult. This is unavoidable, as there is no hard-and-fast distinction between back-channels and related phenomena. Although not particularly important for the purpose of identifying prosodic cues to back-channels, here are a few examples of these borderline cases.

speaker1:	<i>I sent you some pictures</i>	1	
speaker2:	<i>and I will send couple of our pictures in each letter</i>	2	
speaker1:	aiwa (<i>exactly</i>)	3	
speaker3:	<i>and don't worry about your stuff, your brother will transfer them to our apartment</i>	4	(6)
speaker2:	tab kowayes awi (<i>ok very good</i>)	5	
speaker3:	<i>ah ah (yeah yeah)</i>	6	

In several cases a back-channel included a nuance of additional meaning. Example 6 illustrates two back-channels which additionally convey a sense of agreement or acceptance of an offer.

speaker1:	<i>we are going as employees not investors</i>	1	(7)
speaker2:	aiwa aiwa (<i>right right</i>)	2	
speaker3:	<i>helw helw (very good), and of course keep your status as it is currently</i>	3	

Similarly in Example 7 the back-channel also conveys a nuance of agreement: speaker 2 is agreeing that speaker 1's plan is wise. Incidentally, speaker 3's initial agreement here is not labeled as a back-channel, since it continues on to be a full turn.

speaker1:	<i>we heard you are coming to Canada so Faten will see her aunt, aren't you?</i>	1	(8)
speaker2:	<i>no no</i>	2	
speaker1:	<i>you are not intending to go?</i>	3	
speaker2:	<i>no no, we were just talking</i>	4	
speaker1:	okay	5	
speaker2:	<i>(laughs) it was not serious</i>	6	

Although back-channels are by definition optional, sometimes it seems that the speaker is expecting (although not requiring) a back-channel, as in Example 8. (Although this is not apparent from the transcription, it is clear from the audio.)

speaker1:	<i>ok, listen Ashraf</i>	1	(9)
speaker2:	ha-	2	
speaker1:	<i>most likely the Canada thing is okay</i>	3	
speaker2:	okay?	4	
speaker1:	<i>and it'll be August next year to do the immigration</i>	5	

Although back-channels by definition do not require acknowledgement by the speaker, sometimes they do incorporate a nuance of asking for more information. In Example 9 the *okay?* is such a case.

speaker1:	<i>I was praying to God that she gets well</i>	1	(10)
speaker2:	<i>exactly, leave her in God's hands and we</i>	2	
speaker1:	ya rab (<i>please God</i>)	3	
speaker2:	<i>we will pray for her too</i>	4	

The response part of a conversational routine (also called an adjacency pair) sometimes falls into the back-channel category. In Example 10 above, the *ya rab* is such a fixed response, but being not required in this context, was considered to be a back-channel. The same is true for the *ha-* back in Example 9.

3.5 Related Phenomena

For the sake of illustrating how back-channels both resemble and differ from other types of utterances, this subsection presents three of the six examples which initially seemed to be back-channels but which were ultimately judged not to be.

speaker1:	<i>please Ashraf do your best to come to Canada</i>	1	(11)
speaker2:	<i>Enaya (you got it)</i>	2	
speaker3:	<i>because we are thinking that if everything is ok in Canada why don't you come and live with us</i>	3	

In Example 11 Speaker 2 is promising that he will do his best to go to Canada; this goes beyond the normal function of a back-channel and contributes new meaning to the dialog.

speaker1:	<i>the expenses for studying a year here would be about \$2000, do you see?</i>	1	(12)
speaker2:	<i>mmm</i>	2	
speaker1:	<i>so I get the, by the way this call is being recorded</i>	3	

In Example 12, speaker 1 waits for an answer, so the *mmm* is in effect a required response; it cannot be considered a back-channel.

speaker1:	<i>I'm thinking of studying there and come back</i>	1	(13)
speaker2:	<i>yarait (I hope so)</i>	2	
speaker1:	<i>for a year or two until we settle</i>	3	

Finally, in Example 13 “I hope so” was said by a mother to her son. Although positionally similar to a back-channel, it conveys the clear meaning that she agrees and wants him to study.

4 Analysis Method

Our aim was to find a prosodic feature that commonly appeared just before a back-channel by the other speaker, and then determine that this was in fact functioning as a cue.

To do this we chose to directly examine the speech signal itself, especially the F_0 , rather than using any particular model or theoretical framework to guide us. Although there are number of models and frameworks useful for the study of prosody, these have all been designed primarily for read speech and monolog phenomena. As such, these models are not necessarily adequate for describing the sorts of prosodic features involved in turn-taking. Working directly with the signal is advocated as the “direct method” by Shriberg and Stolcke (2004), and has the advantage that the relationship between prosodic features and the phenomena of interest can be discovered, at least in principle, without requiring any hand-labeling of intermediate features, such as sentence boundaries, pitch-accents, target tones, tunes, or turn yields.

We also chose to analyze the data eclectically. Some practitioners of the direct method avoid any use of linguistic knowledge or hunches by the analyst, preferring to rely only on signal processing, statistical analysis and machine learning methods. However such approaches often result in neural nets or decision trees which succeed in classifying the data at the price of being uninterpretablely complex. This project, however, required both a simple qualitative description *and* a quantitative one. The qualitative description is needed so that the initial tutorial module can explain the desired behavior in a simple way that learners could grasp, and the quantitative description is needed so that drills can incorporate an automatic evaluation of the learner’s performance, and also so that the Trainer’s non-player characters (animated agents) can model authentic Arabic turn-taking behavior in unscripted, real-time interactions with the learner.

Thus we used an integrated method for discovering the prosodic cue involved in back-channeling; this gave both qualitative and quantitative descriptions. This method uses both perceptually-based analysis and quantitative analysis, tightly integrated, for the formulation and testing of hypotheses. These analysis phases included simply listening, visually inspecting graphical representations of the pitch and energy, and writing small programs to detect and evaluate various putative back-channel-inviting features. The process was iterative in that the perceptually based and quantitative phases were alternated. In a sense the purpose of listening was to understand how to improve the quantitative description, and the purpose of improving the quantitative description was to direct attention to informative cases in the corpus. The most informative cases were generally those that a tentative version of the quantitative description did not handle correctly, either by failing to identify a back-channel cueing place where in fact there was a back-channel in the corpus (imperfect coverage), or by

incorrectly identifying a back-channel cuing place where in fact no back-channel occurred (imperfect accuracy).

Further description of the analysis method appears elsewhere (Ward & Al Bayyari 2006).

5 Cues for Back-Channels

Although our focus of attention was on the prosody, we did notice a few lexical phrases which also seem to cue back-channel responses: *wakhed balak*, *shayef ezay*, *bos ya* and *baolak aih*. These explained only 2 or 3% of the back-channel occurrences.

The first, most salient, prosodic cue is a pitch upturn at phrase end. This appears to be a relatively strong cue, functioning perhaps like English uptalk.

A second, rarer prosodic cue, only tentatively identified, is a a low flat pitch associated with a lengthened vowel at a disfluency point.

Finally, a sharp pitch downslope seems to be the most frequent cue. The rest of this section describes it further.

5.1 Pitch Downdash as a Cue for Back-Channels

Back-channels are frequently preceded by a certain feature complex produced by the other speaker. This subsection gives a qualitative description.

The most distinctive feature of this feature complex is a region of pitch which is sharply falling; borrowing Bolinger’s terminology, we call this a “downdash” (1989). This fall is generally steady; almost linear when viewed in log scale.

The downdash is typically set off from preceding and following pitch contours by sharp corners, that is, it does not smoothly transition from or to a different contour. A downdash seems to be generally ineffective as a cue for a back-channel if it is closely followed by a region of nearly level pitch.

The downdash seems most effective as an invitation for a back-channel when it occurs after some substantial amount of talk. It also seems most effective when followed by a pause, sometimes immediately but more typically after another syllable or word.

The downdash often falls on syllables which are lengthened. In some cases the downdash is immediately followed by a region of saliently different speaking style — creaky voice, breathy voice, significantly less energy, or an inbreath — and in some cases this transition falls in the middle of a lengthened vowel.

6 Examples

Example 14 and Figure 1 show two back-channels, both *okay*, one at 302 seconds, overlapping the other’s speech, and one at 304 seconds, occurring during a pause. The first one appears to be responding to the downdash at 301.750 seconds, and the second to the pitch upturn at 303.500 seconds.

speaker 1:	<i>most likely we are going to stay for a month and a week just to see how</i>	1	(14)
	<i>life is there</i>		
speaker 2:	<i>okay</i>	2	
speaker 1:	<i>and probably Uncle and Aunt are coming with us</i>	3	
speaker 2:	<i>okay</i>	4	
speaker 1:	<i>so it'd be a good chance for you to come visit us</i>	5	

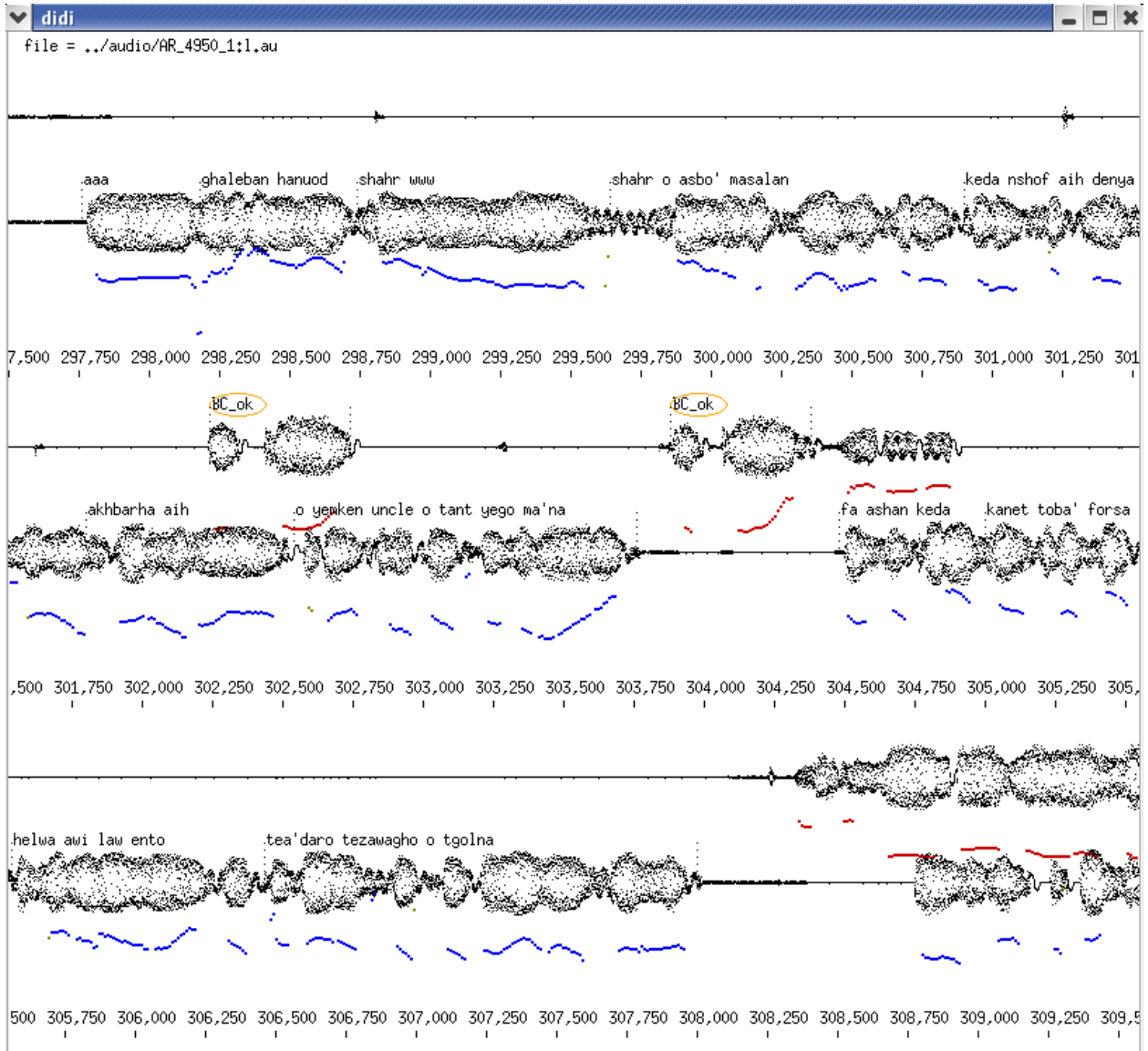


Figure 1: Dialog Fragment Including DOWNDASH, UPTURN, and Two BACK-CHANNELS. Each of the two strips includes two tracks and a timeline. In each strip the top track is one speaker and the bottom track the other. Each track includes: a transcription, the signal, the pitch, and English translation.

Example 15 and Figure 2 show a turn yield, leading the other speaker to take the turn. The phrase *sana yani* at about 54 seconds appears with each syllable at a successively lower pitch level. This is therefore a four-step Kadenz; two and three-step Kadenzes are also seen in the corpus. Although only the first syllable has a clearly flat pitch, perceptually this sounds like a downward staircase of pitches.

speaker 1:	<i>and she was talking, so I told her that you knew for a year</i>	1	
speaker 2:	<i>ok, she's getting married to a Russian guy, but by the time Maha knew, I had not told anybody</i>	2	(15)

Example 16 and Figure 3 show a pitch DOWNDASH that was not followed by a back-channel, although, in the judgment of the second author, it would also have been appropriate had the listener chosen to

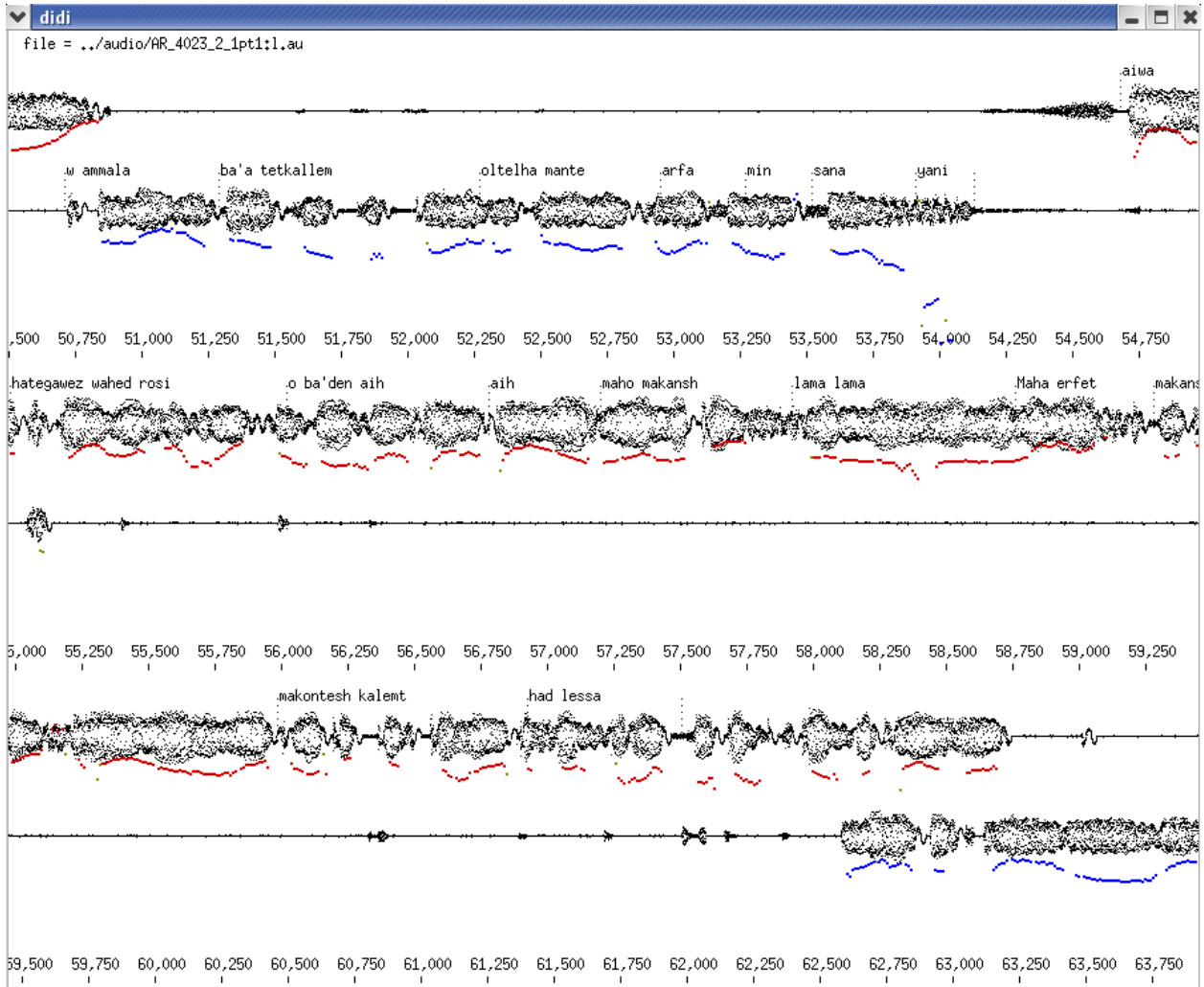


Figure 2: Dialog Fragment Including a Kadenz at a Turn Yield. The pitch just after 54 is so low it is displayed overlapping the time marker. The pitch here at the end is also spotty because the pitch tracker behaves poorly for regions of creaky voice like this one.

respond with a back-channel at this point. In general there is an element of choice, perhaps even of randomness, in back-channel behavior: it seems that a listener typically responds with back-channel feedback at only some fraction of the opportunities given.

speaker 1: <i>Sarah, that little girl, is so cute she's killing me, I told my mom if I were her I would go see her</i>	1 (16)
--	--------

Audio for these examples is available at http://www.cs.utep.edu/isg/Arabic_BC.html.

7 Quantitative Description of the DOWNDASH Feature Complex

This section gives a quantitative description of this feature complex. As the features noted in the last paragraph of Section 5.1 are more difficult to quantify, and since we are less certain that they are

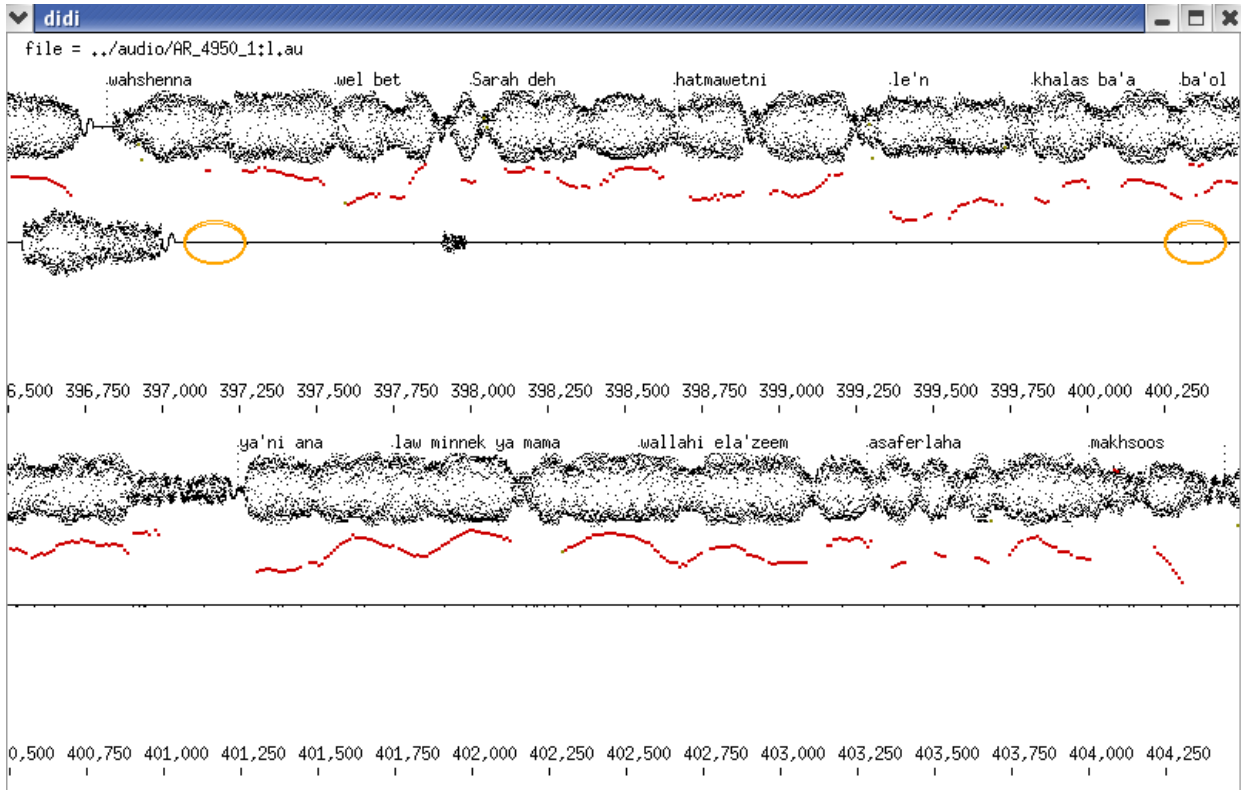


Figure 3: Dialog Fragment Including a DOWNDASH but no Back-Channel. The downdash occurs around second 399.250, and the rule (discussed below) would predict a back-channel 800 milliseconds later.

consistently and distinctively part of the cue, this quantitative description does not incorporate those features.

The feature complex is deemed to be present whenever there is a timepoint which is

- C1** part of an utterance which has lasted at least 1.8 seconds,
- C2** preceded by a downdash lasting at least 30 milliseconds,
- C3** where the pitch in the downdash drops by at least 0.8% every 10 milliseconds,
- C4** followed within no more than 600 milliseconds by a pause (low energy region) which lasts at least 150 milliseconds,
- C5** not followed by a flat pitch region before the pause, where a flat pitch region is one in which the pitch stays within .4% of the average pitch in that region for a period of at least 80 milliseconds, and
- C6** not preceded by another back-channel prediction within 900 milliseconds.

If this feature complex occurs, then a back-channel response is likely to occur some 800 milliseconds later.

This rule is the best of the many variants we have tried, where “best” means performs best according to the criteria described in Section 8. This means that the specific parameters of this rule are tuned to optimize performance, but do not necessarily closely describe the most typical cases of back-channel

cues. In particular, the typical rate of pitch drop is steeper, around 1.2% every 10 milliseconds, which is a semitone every 50 milliseconds. Also, the pitch drop typically lasts much more than the 30 milliseconds required by the rule. Furthermore, the pause usually comes sooner than 600 milliseconds after the pitch drop.

Condition C5 reflects the fact that the downdash functions in contrast to the Kadenz (downward pitch “staircase”) commonly found at turn ends. Although the Kadenz pattern prototypically has several regions of flat pitch, it appears that just a single region of flat pitch is enough to revoke the back-channel invitation associated with a downdash.

Condition C6 means that a downdash occurring shortly after another downdash is treated as a reinforcement of the invitation to back-channel, rather than as an invitation for another back-channel.

8 Evaluation

As mentioned in Section 4, our analysis method relied on the use of the quantitative description to identify informative examples. For this each putative rule, such as the one given above, was treated as a predictive rule. The task accomplished by such a rule is this: given some initial portion of one track of a two-person dialog, predict whether or not a back-channel is about to appear in the other track.

This formulation makes it easy to identify false predictions, that is, places where the cue occurred but a back-channel did not follow, and also missing predictions, where a back-channel occurred but no cue preceded it. The specific criterion for deciding whether a prediction was a success or failure was whether the predicted back-channel point occurred within half a second of the actual onset of a back-channel in the corpus; this window of opportunity was allowed since a back-channel can be produced slightly earlier or slightly and have the same pragmatic effect (Ward & Tsukahara 2000).

8.1 Rule Performance

According to the F Measure, a way to combine accuracy and coverage into one unified performance measure, the rule given above is better than all the other rules we have considered so far. It accounts for 44% of the back-channels in our subcorpus (44% coverage). One reason why this is less than 100% is that in Egyptian Arabic there also seem to be other prosodic features which cue to back-channels, as noted above. Another is that some fraction of the back-channels seem to be produced more in response to the semantic content of the speaker’s utterances, with prosody less important.

The accuracy of the rule is 15%: it makes many predictions at places where a back-channel was not actually present. Since the accuracy to be expected by random predictions is 3.3%, this rule clearly has some predictive power. The reasons why the accuracy is less than 100% probably include inter-speaker differences (no single rule can be expected to model the behavior of all speakers in the corpus), the random element in back-channeling as noted above, and the limitations of the corpus. A more interesting reason is the inability of our current speech processing environment to handle pitch in regions of creaky voice, which are common in this corpus.

8.2 Utility

The rule clearly has some validity, but it’s actual utility is not known: we do not yet know what level of performance would count as performance (Tsukahara & Ward 1997). Quantitatively, we do not know what level of coverage and accuracy would be adequate to enable a computer system to be judged “as good as” (that is, an acceptable model of) a native speaker of Arabic in back-channeling ability, nor to permit a rule to be judged as “good enough” to use for pedagogical purposes.

9 Open Questions

The first priority for future work is to refine this rule. It is doubtless not the best possible, and needs to be improved by further tuning the parameters or adding new parameters. This should improve the accuracy and coverage of the quantitative rule, and also improve the quantitative description. Another obvious priority is an examination of how the occurrence of this prosodic cue and back-channels relates to various semantic, pragmatic, and interpersonal dimensions of dialog. The rest of this section mentions some other topics that should be addressed.

9.1 Experimental Confirmation

Although this paper has shown a correlation between the occurrence of this feature complex in one track and the subsequent appearance of a back-channel in the other track, the existence of a causal relationship remains unproven. Demonstrating this may require controlled perceptual experiments with synthesized speech.

9.2 Iraqi Arabic

Given that many of the prosodic properties of Arabic seem to hold across dialects, it is natural to wonder whether this is also true of pitch downdash as a cue for back-channels. We have begun to examine this for Iraqi Arabic, using our own corpus of dialogs (Ward *et al.* 2006). Unlike the Callhome corpus, this corpus was recorded with the two speakers in the same room. This has two advantages for our purposes: First, since the recordings are not as band-limited as telephone conversations, it is possible to get more accurate and complete pitch estimates. Second, there is no significant line delay (unlike that probably present in some of the cell phone dialogs in Callhome).

The best downdash-based predictive rule found so far achieves a coverage of 51% and an accuracy of 16% for this corpus. There are a number of small differences between the parameters for this rule and the rule for Egyptian presented above. The details appear in (Ward & Al Bayyari 2006). There is also one large difference: for the Iraqi corpus there is typically a delay of 300 milliseconds between the occurrence of the downdash and the appearance of the back-channel, but in the Egyptian corpus the typical delay is 800 milliseconds. This difference may be due to the absence of line delay or to the effects of simultaneous gestural cues to back-channels in the Iraqi data collection.

9.3 Other Factors

In the course of analysis we considered many features as possible back-channel cueing factors. Three deserve further mention.

We have also sought a correlation between vowel lengthening and a subsequent back-channel response, without much success. This may mean that lengthening is not an independent cue for back-channels. Indeed, it may be the lengthening observed with pitch downdashes is there merely to provide enough phonetic content to realize this pitch pattern.

Our colleague Thamar Solorio has examined the relationship between back-channels and energy patterns. In the Iraqi data she has found that a loud syllable or two just before a pause is a fairly reliable indicator of a position where a back-channel would be welcome.

We have also examined the possible role of speaker gestures as an additional way to indicate when back-channels are welcome, using our Iraqi data. Contrary to expectation, no correlations were found (Al Bayyari & Ward 2007).

9.4 Relation to Other Prosodic Phenomena

Although the role of pitch-downdash as a back-channel cue could not have been inferred from previous descriptions of Arabic prosody, it does not contradict them, since grammatical and semantic finality or completion are not closely correlated with interactional completion or invitations. It would be very interesting to attempt a unified description of the prosody of dialog acts, utterance types, finality, and turn-taking.

In terms of theoretical significance, the existence of a significant difference in interactional import between pitch downdash and the Kadenz, although both are ways of reaching a low pitch, suggests that target-based models of prosody are inadequate for describing these phenomena.

In future we would like to understand how the prosodic feature complex identified here relates to other prosodic phenomena, such as declination and lexical stress.

9.5 Pedagogy

Insofar as producing back-channels at appropriate times is part of being a good listener, this skill should be taught to learners of Arabic. In particular, they should be taught to recognize the downdash feature complex, and to understand how it contrasts with the Kadenz pattern. They also should be trained to respond quickly when they hear the downdash in conversation. As both of these abilities are quite different in nature from most language skills, the development of novel teaching methods may be required.

We have developed a prototype 30 minute training sequence. This includes exposure to dialogs from the corpora, rather than professionally recorded conversations, since people acting out dialogs typically follow different rules. It also includes practice in both the speaker role and the listener role, using software that provides feedback on learners' attempts to produce the cue themselves, and feedback on learners' performance as they play the role of an attentive listener in response to one side of a pre-recorded dialog. Preliminary results are positive (Escalante *et al.* 2007).

There is also the question of when back-channeling skills should be taught. If taught early, there is the advantage that learners needing to interact with Arabic speakers can show polite attention through active listening, thereby increasing the chance that the speaker will produce more comprehensible utterances (Kraut *et al.* 1982), and the chance that the dialog will continue long enough for the listener to understand.

On the other hand, teaching back-channeling skills early has risks, given the current state of knowledge. We do not know whether or how Arabic speakers talking to non-natives change their use of back-channel cue; perhaps it is abandoned in favor of a more crude (pause-based) form of turn-taking, or perhaps replaced it with gestural cues, or perhaps the cues are exaggerated. We also do not know whether an early-stage learner whose skill at attentive listening exceeds his other language skills will be perceived by Arabic speakers as polite or phony.

9.6 Cross-Cultural Impressions

Finally, we are interested in the ways in which the prosodic contours we have identified are interpreted by non-native speakers.

To English speakers, some languages of the world can seem beautiful to listen to, but Arabic is probably not one of them. This is doubtless because certain features of Arabic have different roles, including some with negative connotations, in English. Obviously examples include pharyngeals, which may sound harsh to speakers of English, and the absence of de-accenting for given information (Hellmuth 2005), contributing to frequent pitch variation or wider pitch range, which may connote anger to speakers of English (Murray & Arnott 1993). Less obviously, some pitch contours used in questions have similar semantic functions but starkly different "attitudinal meanings" in English and Arabic (El-Hassan 1988).

The prosodic features identified in this paper also have negative connotations in their roles in English dialog. The pitch downdash resembles the utterance-final pitch pattern used in American English to make an accusation or authoritative imperative (Bolinger 1986:208), and the downward staircase or Kadenz (or “terraced monotone”) resembles a final pitch pattern used in American English to express discouragement or resignation (Bolinger 1986:231, Bolinger 1989:324). Since naive English speakers listening to Arabic speech are likely to interpret these features as reflections of the speaker’s personality or attitude, the potential for inter-cultural misunderstandings is clear.

It would be interesting to examine in detail these divergences in interpretation and their significance. We hope that a clear understanding of such differences could help prevent some misunderstandings in interactions between early-stage Arabic learners and Arabic speakers; or even counteract some false impressions of Arabic speakers that casual listeners may pick up from Arabic soundbites on radio or television.

10 Summary

This paper has reported an identification of a prosodic feature, a pitch downdash, which can cue back-channel feedback from the listener in Egyptian Arabic dialogs.

References

- Al Bayyari, Yaffa & Nigel Ward (2007). The Role of Gesture in Inviting Back-Channels in Arabic. In *presented at the 10th Meeting of the International Pragmatics Association*.
- Bergsträsser, G. (1968). *Zum arabischen Dialekt von Damaskus (On Damascene Arabic)*. Georg Olms Verlagbuchhandlung, originally published in 1924 by Orient-Buchhandlung Heinz Lafaire, Hannover, in the series *Beiträge zur semitischen Philologie und Linguistik*.
- Bolinger, Dwight (1986). *Intonation and Its Parts*. Stanford University Press.
- Bolinger, Dwight (1989). *Intonation and Its Uses*. Stanford University Press.
- Canavan, Alexandra, George Zipperlen, & David Graff (1997). *CALLHOME Egyptian Arabic Speech*. Linguistic Data Consortium. LDC Catalog No. LDC97S45, ISBN: 1-58563-114-0.
- Corvetto, di Ines Loi (1982). L’intonazione nell’arabo siriano (The Intonation of Syrian Arabic). *Lingua E Stile*, 17:371–393.
- de Jong, Kenneth & Bushra Adnan Zawaydeh (1999). Stress, duration, and intonation in Arabic word-level prosody. *Journal of Phonetics*, 27:3–22.
- El-Hassan, Shahir (1988). The intonation of questions in English and Arabic. In *Papers and Studies in Contrastive Linguistics, Volume Twenty-Two*, pp. 97–108.
- Eldin, S. Nasser & A. Rajouani (1999). Analysis and Synthesis of Interrogative Intonation in Arabic. In *International Congress of the Phonetic Sciences*, pp. 1509–1512.
- Escalante, Rafael, Nigel Ward, Yaffa Al Bayyari, & Tamar Solorio (2007). Learning to Show You’re Listening: A Back-Channel Trainer for Arabic. presented at the Calico Symposium.
- Fujie, Shinya, Kenta Fukushima, & Tetsunori Kobayashi (2005). Back-channel feedback generation using linguistic and nonlinguistic information and its application to spoken dialogue system. In *Proc. 9th European Conf. on Speech Communication and Technology, Interspeech2005*, pp. 889–892.
- Hafez, Ola Mohamed (1991). Turn-taking in Egyptian Arabic: Spontaneous speech vs drama dialogue. *Journal of Pragmatics*, 15:59–81.
- Hellmuth, Sam (2005). No de-accenting in (or of) phrases: Evidence from Arabic for cross-linguistic and cross-dialectal prosodic variation. In Sonia Frota, Marina Vigarío, & Maria Joao Freitas, editors, *Prosodies: With special reference to Iberian languages*, pp. 99–121. Mouton de Gruyter.

- Johnson, W. Lewis, Carole Beal, Anna Fowles-Winler, Ursula Lauper, Stacy Marsella, Shrikanth Narayanan, Dimitra Papachristou, Andre Valente, & Hannes Vilhjalmsson (2005). Tactical Language Training System: An Interim Report. USC ISI, adapted from a conference paper presented at the Intelligent Tutoring Systems Conference, September 2004.
- Kraut, Robert K., Steven H. Lewis, & Lawrence W. Swezey (1982). Listener Responsiveness and the Coordination of Conversation. *Journal of Personality and Social Psychology*, 43:718–731.
- Kulk, Friso, Cecilia Odé, & Manfred Woidich (2005). The intonation of colloquial Damascene Arabic: a pilot study. In *IFA Proceedings 25; The Proceedings of the Institute of Phonetic Sciences, Amsterdam*, pp. 15–20.
- Maynard, Senko K. (1989). *Japanese Conversation*. Ablex.
- Murray, Iain R. & John L. Arnott (1993). Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *Journal of the Acoustic Society of America*, 93:1097–1108.
- Rifaat, Khaled (2005). The Structure of Arabic Intonation: A preliminary investigation. In Mohammad T. Alhawary & Elabbas Benmamoun, editors, *Perspectives on Arabic Linguistics XVII-XVIII*, pp. 49–67. John Benjamins.
- Shriberg, Elizabeth E. & Andreas Stolcke (2004). Direct Modeling of Prosody: An Overview of Applications in Automatic Speech Processing. In *Proceedings of the International Conference on Speech Prosody*, pp. 575–582.
- Tsukahara, Wataru & Nigel Ward (1997). Rikai o Kaisanai Kaiwa Gensho to shite no Aizuchi (Back-channel Feedback as a Reflex Phenomenon). *Gengo*, 26:90–97.
- Ward, Nigel (2004). Pragmatic Functions of Prosodic Features in Non-Lexical Utterances. In *Speech Prosody 04*, pp. 325–328.
- Ward, Nigel (2006). Non-Lexical Conversational Sounds in American English. *Pragmatics and Cognition*, 14:113–184.
- Ward, Nigel & Yaffa Al Bayyari (2006). A Case Study in the Identification of Prosodic Cues to Turn-Taking: Back-Channeling in Arabic. In *Interspeech 2006 Proceedings*.
- Ward, Nigel & Wataru Tsukahara (2000). Prosodic Features which Cue Back-Channel Feedback in English and Japanese. *Journal of Pragmatics*, 32:1177–1207.
- Ward, Nigel G., David G. Novick, & Salamah I. Salamah (2006). The UTEP Corpus of Iraqi Arabic. Technical Report UTEP-CS-06-02, University of Texas at El Paso, Department of Computer Science.
- Yngve, Victor (1970). On Getting a Word in Edgewise. In *Papers from the Sixth Regional Meeting of the Chicago Linguistic Society*, pp. 567–577.