A Case Study in the Identification of Prosodic Cues to Turn-Taking: Back-Channeling in Arabic

Nigel G. Ward and Yaffa Al Bayyari

Department of Computer Science
University of Texas at El Paso
nigelward@acm.org, yalbayyar@utep.edu

Abstract

Discovering and quantifying the prosodic signals that help manage turn-taking is difficult, in part because of the limitations of commonly used methods. This paper presents an integrated method that uses both perceptually-based analysis and quantitative analysis. The eight activities involved in the method — clarification of aims, problem formulation, corpus preparation, feature discovery, feature combination, hypothesis refinement, tuning, and evaluation — are illustrated using task of finding prosodic cues for back-channel feedback in Arabic.

Index Terms: Discovery, Methods, Dialog, Prosody

1. Introduction

One of the key issues for this special session, The Prosody of Turn-Taking and Dialog Acts, is the discovery problem. Although recent years have seen a number of solid qualitative and quantitative findings showing how prosody can cue turn-taking and indicate dialog acts, today these findings are scattered: for no such phenomenon is there yet a complete understanding of the prosodic features involved and their meanings and pragmatic effects in context. In part this is due to the limitations of the research methods available.

Common methods include traditional descriptive linguistic methods, instrumental methods, conversation analysis, and direct methods [5]. However none of these methods gives the complete picture, and the results of different methods can be forbiddingly difficult to relate to each other.

This paper presents an integrated method for discovering the prosodic cues involved in turn-taking. This method uses both perceptually-based analysis and quantitative analysis, tightly integrated, for the formulation and testing of hypotheses. As such, it exhibits most of the advantages of all previous methods.

Subsequent sections of this paper describe the activities involved in this method: clarification of aims, problem formulation, corpus preparation, feature discovery, feature combination, hypothesis refinement, tuning, and evaluation. Each activity is illustrated with respect to a case study, the problem of identifying the prosodic cues which invoke back-channels (also known as minimal responses or continuers) in Arabic.

2. Project Aims

This study arose from the need to investigate turn-taking behavior in Arabic in order to extend an intelligent tutoring system, the Tactical Language Trainer [2]. 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, thoughtless, discourteous, passive, untrusting, pushy, or worse [8]. 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 project required both a qualitative description *and* a quantitative one: the former so that the initial tutorial module could explain the desired behavior in a way that learners could grasp, and the latter so that drills could quantitatively evaluate the learner's performance, and also so that the Trainer's non-player characters (animated agents) could model authentic Arabic turn-taking behavior in real-time interactions with the learner.

3. Problem Formulation

While there are several perspectives on how to model turn-taking, the initial aim here, as for most engineering and linguistic studies, is simply to identify prosodic cues that, when produced by the speaker, suggest how the listener should respond. Specifically the goal was to identify the prosodic features which indicate to the interlocutor when back-channel feedback is especially welcome, and thereby make it likely that the interlocutor will produce a back-channel in response.

Since the ultimate goal is to enable an automated agent to take one side of a conversation, a good measure of the quality of a model is its ability to predict where back-channels will appear in one track of a dialog, given only the information in the other track so far. This rules out recourse to hand-labeling of events such as tone boundaries, turns, or utterance ends.

4. Corpus Preparation

Corpus-based analysis is necessary for problems like this, 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.

Initially we obtained LDC's CallHome Corpus of Egyptian Arabic Speech, and a native Arabic speaker, the second author, labeled all back-channels in a subset, giving 660 occurrences. Thanks in part to the size of this subcorpus, we were able to find one very nice dialog, where the participants were articulate, at ease

[0] This research was sponsored in part by DARPA and in part by NSF Grant No. 0415150. We thank Thamar Solorio, Olac Fuentes, David Novick, and an anonymous referee.

with each other, and involved in the topic. This dialog was also long enough to contain enough normal conversation, rather than just prosodically special activities such as conversational routines, quotations, negotiations of commitments, and telephone number giving. This dialog was also rich in back-channels, so we made it our favorite and started analysis with it.

Later we collected 112 minutes of face-to-face dialogs in Iraqi Arabic [9]. This corpus included 689 back-channels. The last 15% of each dialog was reserved for testing.

5. Feature Discovery

5.1. Where to Look

Any plausible prosodic cue for back-channels must occur at least 200 milliseconds before back-channel onset, based on the minimum human reaction time, and probably not more than a second earlier, so this is where we looked.

It would be convenient if the exact places to look for cues could be known in advance. However this is not possible because the delay between the time the listener hears the cue and the time he or she responds is quite variable, a problem which, incidentally, makes it difficult to apply machine learning methods to this problem.

There is, however, an assumption that is sometimes made in order to allow machine learning techniques to be applied: that the cue must occur at the boundary (utterance end) that most closely precedes the back-channel. However this assumption is problematic because in some languages back-channels and even turn starts often overlap the interlocutor's ongoing speech, because often the listener action starts less than 200ms after a boundary, and because preceding utterance ends may not actually co-occur with the cue, for example in cases of "post-completion", as in *"At the mall it was crazy. Just crazy."* where the effective turn end may be after the first word *crazy*, with the subsequent comment not intended to hold the floor. Experimental manipulations also show that the prosodic cues to turn-taking are not always located immediately pre-boundary [4]. (There is one way in which boundary-based analysis can work: if the boundaries are labeled by someone who has listened to the whole dialog, and thus knows the locations of the upcoming back-channels. Approaches which use this as a starting point are exploiting future information and hand-labeled information, and so are solving only half of the problem.) As the boundary-based assumption is not tenable, our search for cues ranged widely over the regions preceding back-channels.

5.2. What to Look At, What to Look For

Pitch is of course the primary prosodic feature, but a tricky one to work with, thanks to problems such as doubling, halving, dropouts, unvoiced consonants, and creak. Rather than using techniques to automatically convert raw F_0 data to a better approximation of the actual perceived pitch, we took the approach of trusting to the size of the corpus to have enough cases where the raw F_0 provides useful information.

One challenge in feature hunting is the abundance of possible features that one might consider. Even simple ones, such as average pitch, pitch slope, maximum pitch, and pitch range, can be computed over various intervals, creating a multitude of possible features. Beyond that, arbitrarily complex pitch features could be involved, such as number of pitch peaks over the past 500ms, height of highest pitch peak in the last 400ms relative to the baseline computed over the past 2000ms, first coefficient of a second-order approximation to the pitch curve over the last three syllables before a pause of at least 200ms, and so on, where all the feature-defining parameters can range over many values. The abundance becomes breathtaking when one includes features computed from other dimensions of prosody — energy, voicing type, duration, rate and timing — let alone composite features combining multiple dimensions.

With so many possible features, it seems unlikely that systematic, exhaustive examination can be productive. We set out to discover cues by instead seeking some commonality in the prosody of all (or most) of the regions which precede back-channels.

5.3. How to Look

To find such commonalities, it is advantageous to use both auditory and visual methods.

In analyzing dialog, looking seems to be more effective than listening for things like navigating to points of interest, scanning large amounts of data looking for recurring prosodic patterns, and detecting commonalities in pitch between two utterances. However visual-only analysis can lead one astray. For one thing, features that are visually salient, such the degree of smoothness vs. jaggedness of the pitch contours, are not necessarily salient as auditory features. Visual analysis is of course also limited to features that are easy to compute and display; these typically do not include features such as speaking rate, creaky voice, breathy voice, and spectral tilt.

On the other hand, listening is effective for perceiving more information, but it has the disadvantage of being harder to focus on specific places in the signal and on specific dimensions of the prosody. In addition, unfocused listening can easily be distracted by more salient phenomena, for example the prosody of the syllable immediately preceding a pause.

To get the advantages of both looking and listening we used a tool that makes it easy to both view and listen to arbitrary parts of the signal, Didi [7]. This makes it easy to discover something of interest by listening and then take a closer look at how its pitch and energy contours appear visually, and conversely. This tool also provides the ability to navigate quickly to the context of a back-channel and to move back and forth in the local vicinity, among other features.

Examining the Arabic dialogs in this way, it soon became clear that there were several different cue forms. The most salient was a pitch upturn at turn end. Also salient was a low flat pitch associated with a lengthened vowel at a disfluency point. These accounted for only some of the back-channels. Directing attention to the remaining cases, we noticed that many were preceded by a steep pitch downslope, a "downdash".

5.4. Quantification

The next step, quantifying a feature, involves several activities. First is producing a numeric description of the prosodic feature. Starting from the initial auditory and visual impressions perceptions. We did this by alternately considering the sound, the visual display, and the underlying numeric data.

The next activity is operationalizing the description. We programmed each feature detector in C. (It would clearly be better to do this in some higher-level "feature-definition language", perhaps based on Matlab, but this does not seems to be supported by any existing dialog analysis tool.) Doing this was not straightforward, because there can be more than one way to formally characterize the same feature. For example, we later needed to model a staircase-like pitch pattern, one with successive periods of nearly flat pitch interleaved with down jumps. A feature detector for this pattern could work by picking up the flat pitch regions, or the downward jumps, or both. It is impossible to know which to use without programming both options and seeing which does a better job. Even if logically equivalent, two alternative descriptions may not behave the same with real signals, for example when their input includes frames with missing pitch. Ultimately choices of how to quantify features should probably be based on psychoacoustics; instead we used a rough preference for simplicity — for example, preferring conjuncts that refer to change or its absence, in F_0, delta-F_0, energy, or other simple features — and for low computational cost.

The third activity is verifying that the feature detectors as programmed are behaving as expected. We did this again by both looking and listening. A useful technique was to superimpose a symbol on the display at each place when the code detected the presence of a feature; this made it easy to do a quick check of a dozen or so places to verify that the symbol appears when it is supposed to, and only when it is supposed to. It generally takes an iteration or two to make a feature detector work as intended.

6. Feature Selection and Combination

Having identified a likely prosodic feature, the next step is to test whether the presence of this prosodic feature in one track does in fact predict the subsequent occurrence of a back-channel in the other track: that is, to evaluate the hypothesized relation.

A preliminary evaluation can be done by computing the accuracy, that is, the percent of predictions which are valid. If this is no higher than the baseline, the feature is probably not relevant. This happens surprisingly often: it is easy to be seduced by prosodic features that occur frequently in the regions of interest, but that turn out to also occur frequently in other places. This danger may be reduced by first devoting some time to simply listening to the favorite dialog repeatedly, for the sake of familiarization with the common prosodic patterns.

For Arabic we discovered in succession that none of our initial features — lengthened vowels, pitch downdashes, and jagged pitch contours — actually performed well as predictors. After each failure we began another iteration of the feature discovery process.

Rejecting a hypothesis may not mean that a feature should be abandoned entirely, however. Individual prosodic features may be relevant only in certain contexts or in conjunction with other prosodic features. This was the case for the pitch downdash: when we gave it a second look, based on its prevalence before back-channels, we found that the correlation with the presence of a subsequent back-channels was broken mostly in cases where it occurred near the start of an utterance. Thus the same prosodic feature appears to have different functions depending on the context.

This led to the addition of another “prosodic” feature, time-into-utterance, and the creation of a two-clause predictive rule, predicting a back-channel if the time-into-utterance was greater than some threshold and the steep pitch downslope was present. Performance of this rule, although still poor, was better than random.

7. Hypothesis Refinement

Having an initial predictive rule, the next step is to improve it. This involves the close inspection [3] of missed predictions and incorrect predictions (false alarms) to diagnose the cause of each problem. This was supported by the Didi features for quick navigation to the contexts of missing predictions and false alarms; these made it easy to jump quickly to familiar cases in the favorite dialog, and also to find informative places in the less-studied data. The diagnoses were of various kinds.

Sometimes the problem was due to an aspect of one of the feature detectors. For example, one speaker in the corpus had a tendency to produce noisy in-breaths. Although this behavior may have had communicative significance, for our rule the primary implication was that it caused problems for our simple speech/nonspeech detector, and therefore made both the time-into-utterance feature and the pause detector more noisy.

Sometimes the problem was traced to something beyond the scope of this study, such as semantic aspects, unusual contexts, apparently idiosyncratic speaker behavior, and so on. These cases had to be identified and marked as uninformative, that is, being without implications for the task of refining the formulation of the prosodic cue.

The most interesting problems were those which could be traced to a missing prosodic factor, something that needed to be added as a new conjunct to the predictive rule. As our rule was fairly easy to understand and simulate by hand (unlike, say, large decision trees), it was possible to understand why it succeeded or failed in a particular case, and then determine what needed to be changed or added.

For example, although a sharp downdash was a cue in itself, a downdash with a more moderate slope was a reliable cue only when followed by a pause. Interestingly the two events were not always synchronized: sometimes the pause came immediately after the pitch downslope but more typically it came after another syllable or word.

Another factor was discovered less directly. Many of the false alarms were due to cases where the downslope came at a turn-end, that is, cases where it was followed by a full utterance from the other person, rather than a back-channel. Attempting then to characterize the turn-end signal using the methods described above, we found many cases of a staircase pattern of flat pitches interleaved with down jumps (similar to pattern earlier identified as a mark of finality [1]). Re-examining the false alarms, we found that some of them occurred where the steep downslope was immediately followed by a flat pitch; we interpreted these as a turn-end signal cancelling a previous turn-hold signal. Interestingly, this turned out to provide an easy way to understand our earlier observation that jagged pitch contours tended to precede back-channels: as jagged pitches are intrinsically not flat, the correct (more natural, easier to perceive, easier to compute) feature seems to be the absence of flat pitch, rather than the presence of jagged pitch.

After each refinement to the hypothesis, typically involving the addition of a conjunct to the rule, the performance was evaluated by examining its behavior on the cases that motivated the refinement, and its effect on the overall accuracy or coverage. This step typically also took several iterations. In a sense the purpose of listening is to understand how to improve the quantitative description, and the purpose of refining the quantitative description is to direct attention to informative cases in the corpus.

8. Tuning

In the refinement phase the focus is on identifying ways to improve either the accuracy or coverage. In the tuning phase, the aim is to improve both together: to maximize the F-metric or some other combination of accuracy and coverage. We did this by systematically varying, one by one, the key parameters of each conjunct in a rule. While not a generally a reliable way to optimize a function, this does not appear problematic for identifying prosodic cues.

Tuning must be done last, after the features are reliably identified and the basic form of the prediction rule is set. If not, tuning can lead to implausible parameter values. For example, when we did a premature first pass at tuning the parameters defining the downslope, before adding the conjunct requiring it not to be turn-initial, the best performance was obtained at a threshold for slope that was so loose as to allow gentle pitch rises: a clearly implausible outcome.

Our qualitative and quantitative descriptions of the rule for Egyptian Arabic appear in [8]. A slightly different version was found to better model of the Iraqi data; specifically, the downdash feature complex is deemed to be present whenever there is a time-point which is:

C1 part of an utterance which has lasted at least 1.2 seconds,

C2 preceded by a downdash lasting at least 40 milliseconds,

C3 where the pitch in the downdash drops by at least 0.7% every 10 milliseconds,

C4 followed within no more than 500 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 at least 80 ms., and

C6 not preceded by another back-channel prediction within 1.3 seconds.

A back-channel is predicted to occur in response to this feature complex, 300 milliseconds later.

9. Evaluation

An initial evaluation of the rule can be done by discussing it with native speakers. Although prosodic cues for turn-taking are not generally consciously known, they are salient enough so that after they are pointed out, they are readily apparent to the unaided ear, meaning that a native speaker can judge whether a prosodic feature does bear the hypothesized turn-taking significance, perhaps by imagining how an utterance would sound with versus without the feature. In this respect discovery was hard but verification was easy.

The rule can also be evaluated by measuring its ability to predict, from one side of a dialog, the places where back-channels could appear in the other track. Scoring by the ability to predict the actual back-channels in held-out test data from the Iraqi corpus, using the criteria described in [10], the coverage of our rule was 51% and the accuracy was 16% on the held-out test data. Although perhaps not above that obtainable with other methods [6], this performance was comfortably above the accuracy of random predictions, 5%. The primary causes for missing predictions were back-channels cued by the pitch rise pattern (about 14% of the total), largely semantically- governed back-channels, and pitch detector failures in regions of creaky pitch. The primary cause for incorrect predictions was the fact that any actual listener typically responds with back-channel feedback at only some fraction of the opportunities given, although the rule identifies all opportunities; other causes included situations where both people were talking at once, places where back-channels were semantically inappropriate, and clear individual differences in back-channel behavior.

10. Conclusions

This paper has presented an integrated method for the discovery of prosodic cues to turn-taking behavior. The method combines perceptual and quantitative methods, with analysis proceeding though iterative cycles combining both. For Arabic, as earlier for Japanese, this method has led to discovery of relevant prosodic features that may not have been discovered using any other method.

This method is currently rather labor-intensive, but it should be possible to reduce the effort required by improving the tool suite to better support the analyst's workflow and by proving integrated access to machine methods for feature discovery, refinement, and combination.

11. References

[1] Bergsträsser, G., 1968. *Zum arabischen Dialekt von Damaskus* (On Damascene Arabic), Hildesheim: Georg Olms Verlagbuchhandlung. Originally published in 1924 by Orient-Buchhandlung Heinz Lafaire, Hannover, in the series Beitraäge zur semitischen Philologie und Linguistik.

[2] Johnson, W. Lewis, Carole Beal, *et al.*, 2005. Tactical Language Training System: An Interim Report. USC ISI, adapted from a paper presented at the Intelligent Tutoring Systems Conference, September 2004.

[3] Local, John and Gareth Walker, 2005. Methodological Imperatives for Investigating the Phonetic Organization and Phonological Structures of Spontaneous Speech. *Phonetica*, 62, pp 120-130.

[4] Ohsuga, Tomoko, Masafumi Hishida, Yasuo Horiuchi, Akira Ichikawa, 2005. Investigation of the Relationship between Turn-taking and Prosodic Features in Spontaneous Dialog. *Proc. Interspeech 2005*, pp 33-36.

[5] Shriberg, Elizabeth E. and Andreas Stolcke, 2004. Direct Modeling of Prosody: An Overview of Applications in Automatic Speech Processing. *Proc. International Conference on Speech Prosody*, Nara, Japan, pp 575-582.

[6] Solorio, Thamar, Olac Fuentes *et al.*, Prosodic Feature Generation for Back-channel Prediction. in these Proceedings.

[7] Ward, Nigel, 2005. Didi, a Dialog Display and Labeling Tool, http://www.cs.utep.edu/nigel/didi/

[8] Ward, Nigel and Yaffa Al Bayyari, 2006. A Prosodic Feature that Invites Back-Channels in Egyptian Arabic. presented at the 20th Arabic Linguistics Symposium, Kalamazoo, Michigan.

[9] Ward, Nigel, David G. Novick and Salamah I. Salamah, 2006. The UTEP Corpus of Iraqi Arabic. Univ. of Texas at El Paso, Dept. of Computer Science, Tech. Report UTEP-CS-06-22.

[10] Ward, Nigel and Wataru Tsukahara, 2000. Prosodic Features which Cue Back-Channel Feedback in English and Japanese *Journal of Pragmatics*, 23, pp 1177–1207.