INTERSPEECH 2009 BRIGHTON

Using Responsive Prosodic Variation to Acknowledge the User's Current State

Nigel G. Ward, Rafael Escalante-Ruiz

Department of Computer Science, University of Texas at El Paso, USA

nigelward@acm.org, rafa.escalante@gmail.com

Abstract

Spoken dialog systems today do not vary the prosody of their utterances, although prosody is known to have many useful expressive functions. In a corpus of memory quizzes, we identify eleven dimensions of prosodic variation, each with its own expressive function. We identified the situations in which each was used, and developed rules for detecting these situations from the dialog context and the prosody of the interlocutor's previous utterance. We implemented the resulting rules and had 21 users interact with two versions of the system. Overall they preferred the version in which the prosodic forms of the acknowledgments were chosen to be suitable for each specific context. This suggests that simple adjustments to system prosody based on local context can have value to users.

Index Terms: Confidence, Emotion, User Modeling, Sensitive, Tutorial, Quiz, Social Interaction, Pace

1. Towards Responsive Dialog Systems

Users of spoken dialog systems often perceive these systems to be robotic and inflexible. We would like dialog systems to instead be sensitive, responsive, and expressive, displaying the social interaction skills that are so important in human-human dialog. Achieving this requires advances in various areas of speech science — including prosody, emotion, user state modeling, and dialog management — and topics in these areas have been addressed by many research projects. However only a few studies have put the pieces together in actual dialog systems. Doing so is important, for at least two reasons. First, the perspective provided by the construction of a complete system can tell us what really matters for dialog, and so provide direction for the component-oriented research. Second, complete systems enable experiments to determine whether these aspects are actually effective and valued by users.

This paper describes the development and evaluation of what may be the first system in which the prosody of the system's responses is chosen dynamically, at run-time, based on the user's state as inferred from the context and the prosody of his or her utterances.

2. Acknowledgments in Tutorial Dialog

We chose to work in the domain of memory quizzes [1, 2], which are dialogs consisting largely of questions or puzzles posed by the tutor, guesses by the student, feedback from the tutor, and the requesting and giving of hints. Figure 1 presents an example. Such dialogs are semantically tractable but still interesting pragmatically, with rapid interplay between the tutor and the student on various "emotional" and control dimensions. They form a convenient domain for a case study, but may also have practical value for helping students memorize or review subjects such as multiplication tables, standard abbreviations, famous people, and dates [3]. The corpus we used [2] has a total of 48 dialogs, between one master tutor and 16 students.

Our interest here is in the acknowledgments produced by the tutor in response to correct answers. In general, back-channels and acknowledgments are known to be important in synchronizing and grounding the contributions of participants. In this corpus the acknowledgments are expressive and rich in variation, although perhaps not as complex as in some other domains [4]. In previous work we focused on the factors governing lexical choice in acknowledgments — for example, when the system should say *very good* or *good job* rather than *uh-huh* or *mm-hmm* — developed rules for which acknowledgment to use when, and experimentally verified their effectiveness [2]. However in debriefings the subjects often commented on the prosody of the acknowledgments, although the prosody was not varied and they had been asked to ignore it anyway. It seemed that they could not help noticing the prosody and having it affect their perceptions. Clearly this was something to look into.

The prosody of back-channels and acknowledgments has lately received much attention [5, 6, 7]. However the specific aspects of acknowledgments prosody needed for dialog systems have not been studied, nor has the actual value of manipulating acknowledgment prosody. These are the topics of this paper.

3. Methods

We started out working bottom-up. As the corpus included much obvious variation in the prosody of the acknowledgments, we set out to determine the reasons for this: what these prosodic variants were expressing. First we listened to the acknowledgments in isolation, trying to label the "emotions" they were expressing. Common labels at this stage included terms such as warm, enthusiastic, helpful, empathetic, impatient, condescending, anxious, and absent-minded. Informal experiments with colleagues showed, however, that such perceptions varied significantly among listeners, so this was something of a false start.

Next we listened to acknowledgments in context, again assigning labels, including descriptions (e.g. "mere confirmation, a bit warm" and "empathetic — feels like she is responding even before checking that the answer is correct") and paraphrases of what we felt the prosody was expressing (e.g. "calm down, don't stress, you're still doing okay"). These context-

System: In reverse chronological order, name ten presidents of the United States, starting with Jimmy Carter.
Subject: *Carter, Nixon,*
System: No, that's not it.
Subject: *Kennedy?*
System: No, that's not it. His wife's name is Betty.
Subject: *Ford.*
System: Good job.

Figure 1: Transcript of a subject's interaction with the system

informed perceptions varied less across listeners, and listening in context resolved many small mysteries; for example, it turned out that some of the acknowledgments that had seemed condescending seemed in fact to be expressing empathy in contexts where the student was struggling; some acknowledgments that had seemed energetic were actually signaling completion; and some that had seemed impatient were in fact common when the student was doing well and seemed to express an expectation of continued success and an indication to keep going.

We then set out to identify the specific prosodic features that were expressing these various functions, using qualitative inductive methods, including detailed comparisons of the various tokens and the various contexts they occurred in. This was aided by the existence of approximate minimal pairs, e.g. where the tutor responded *very good* four times in a row with various differences in prosody. We found many differences that were subtle but meaningful, and that most of the prosodic features seemed to be graded rather than categorical, so we decided to model the prosodic variations as directly reflecting the pragmatic functions, rather than serving simply to mark a category (e.g. back-channel vs. acknowledgment vs. stall [6]).

4. Prosodic Features and Functions

This section lists the prosodic features found in acknowledgments and describes their observed contexts of occurrence and inferred functions. The first six appeared to relate primarily to turn-taking, floor control, and pacing.

A *pitch upturn* seemed to serve most commonly to allow and encourage the student to continue. In particular this was used when the student was getting back on track after a period of difficulty, and seemed to invite a speed-up in the pace of the interaction. (More rarely it served as an indication that the tutor was going to continue speaking, as when following the acknowledgment with an echo of the correct answer; this seemed to happen only where the student was clearly not intending to take the floor, so there was no ambiguity.) Pitch upturns come in various forms, ranging from long upslopes to tiny final upticks, that were all perceptually very similar.

In contrast *creaky voice* seemed to convey that the tutor wished to slow down the pace and reassert control, often occurring in response to over-confident or domineering student utterances (marked, for example, with a preceding *oh* or a strong pitch downslope and perhaps thereby indicating the intention to proceed without waiting for acknowledgment). The tutor used creaky voice to various degrees and over various fractions of the acknowledgment, and in general the more confident the student, the longer and stronger the creakiness of the response. She used creaky voice mostly when the student was male.

There were also a few less common prosodic features which seemed to relate to pace and turn-taking. *Loudness* seemed to occur most often as a way to grab the floor to control the pace, for example when the tutor's acknowledgment appears later than usual, and she possibly feels at risk of forfeiting her turn. *Shortness* of duration seemed to indicate that the acknowledgment was intended not to interrupt, but to allow the student to continue on at their own pace. Conversely, greater *length* seemed to indicate a desire to slow the pace of the interaction, sometimes by directly indicating the mental time involved in verifying the correctness of a guess. *Delay* in time from guess to acknowledgment may also have this function. Finally a *low pitch*, especially when present at onset, seemed to indicate dominance or control.

There were four prosodic features which seemed to relate mostly to acknowledgment of the student's state and the tutor's "emotional" response to that state.

Vibrato seemed to be used to provide reassurance when the student was lacking confidence in his guess, typically as indicated by a rising pitch. (An extreme rising pitch is of course in general a hallmark of a yes/no question, but in this corpus there appears to be no categorical difference between the strongly rising-pitch guesses and the mildly uncertain guesses.) Vibrato also seemed to convey warmth and sometimes praise, encouragement, empathy, and pleasure at the student's success, providing a personal rather than a purely businesslike confirmation. This occurred less frequently when the student was male. Vibrato can perhaps be considered a mild case of *pitch range expansion*, which also seemed warm and encouraging.

Variation in syllable length, specifically elongation of one segment of the acknowledgment (e.g. *vvvery good*, *very goooood*, *good jooob*) seemed to convey a similar meaning.

Pitch downslope seemed to indicate certainty, to provide confirmation and sometimes reassurance, but without much connotation of warmth. The default, neutral intonation for acknowledgments also generally included slight pitch drop, but the steeper drops sounded more definitive. They occurred commonly in response to the first correct answer of a sequence and in response to the final answer.

There were also a few miscellaneous prosodic features and phenomena. *Syllable-boundary strength*, realized by such features as discontinuous pitch contours and creaky voice, seemed to indicate alertness, attention, control, formality, and distance. *Breathiness* seemed to indicate amusement, and typically occurred in response to breathiness by the student. *Reduction* in pitch range and loudness over repeated acknowledgments was seen when the student was producing correct answers at a steady pace. Finally, the tutor appeared to deploy *variation* to avoid producing more than two identical tokens in a row, perhaps to avoid seeming robotic.

Although we present these features as a list — reflecting our belief that these are, to some extent, independent dimensions of prosody, reflecting independent pragmatic functions — these dimensions are clearly not orthogonal; as evidenced by the fact that certain functions never co-occur in the corpus. Although we list functions relating to pace and control separately from those relating to affective dimensions, no clear distinction can be made; the various types of functions are intertwined.

5. The Experimental System

We wanted to see whether suitably varying the prosody of acknowledgments would be perceived positively by system users. To test this, we needed a specific set of rules for controlling acknowledgment prosody — not necessarily a complete or independently validated or optimized set of rules, but one with at least a few good rules — and these we developed by quantifying some of the above observations.

To do this we switched perspectives, considering what sort of states the user could be in (momentary states, rather than more complex knowledge or learning states [9]), and what the system had to do in each case to appear sensitive and responsive. Thus we grouped the observations of the previous section around the contexts in which they were relevant. We then considered how to compute the relevant aspects of the user's state.

(In doing so we decided to leave out those aspects of prosody which related to tutor-side behaviors, notably those compensating for the occasional delays in responding. This was because our setup for the experiments enabled us to provide ac-

Conditions	Student Feeling	Tutor Feeling	Tutor Prosody
delay $>$ 4 sec & hints $>$ 1	question is hard, possibly wanting praise	warm, praising	elongated
immediate incorrect guesses $>$ 1 & total incorrect guesses $>$ 3	not doing well, possibly discouraged	praising, encouraging	elongated and lively
immediate incorrect guesses = 0 & salient pitch downslope	doing well, possibly feeling dominant	keeping control	creaky
no hints & immediate incorrect guesses = 0 & total incorrect guesses $>$ 3	was not doing well, but now doing better	welcoming a speed-up of pace	upturn
strong pitch upturn & delay $>$ 2 sec	low in confidence on this guess	reasssuring	vibrato
delay $>$ 3 sec	confident but still needing time to recall	expecting good performance to continue	creaky
immediate incorrect guesses = 0 & delay $<$ 2 sec	certain	no time to acknowledge	acknowledgment omitted
delay $<$ 4 sec & immediate incorrect guesses = 0	confident, but still needing time to recall	expecting continuation of good performance, at a slower pace	creaky and elongated
default	neutral	neutral	neutral

Table 1: Rules for Responsive Prosody in Acknowledgments, as Implemented

knowledgments with a fixed and rapid response time. However it could be very useful to control these aspects in systems where variable speech recognition delays do occur.)

Based on previous work, e.g. [2, 8], and intuitions developed by listening to the corpus, we inferred the user's state through indications of three types: 1. Delay, indicating confidence, specifically delay from the closure of the previous round (the time of onset of the tutor's acknowledgment of the previous correct answer) to the onset of the current correct guess. 2. The student's recent level of performance, as indicated by the number of hints needed before he or she got the right answer, by the number of incorrect guesses for the current president, and by the total number of incorrect guesses in the dialog so far. 3. The pitch slope, measured over a linear approximation to the pitch over the last quarter of the guess, quantitized into three categories: strong upturn if rising at a rate of $>$ 50% per 100ms, salient downslope if $<$ –10%, and neutral otherwise.

While it would be elegant to implement a dimensional model directly, in which continuous-valued features of the context and the user's speech determined continuous values for each of the pitch qualities of the response, we instead built a categorical model, mostly for convenience of implementation. Thus the user state and context were classified into one of nine categories, as shown in Columns 1 and 2 of Table 1.

Since we aimed to estimate the value of the prosodic manipulations, we held constant the lexical form of the acknowledgment: the system always responded *Good job*. The tutor prosody, column 4 of the table, was realized by various alterations of a neutral token from the corpus [neutral.au]. Elongation was done by adding 4 or 5 additional pitch periods during the vowel of *job*, using Audacity [elongated.au]. The elongated and lively token was another instance of *good job* from the corpus, one that was somewhat breathy and had a large pitch range and high volume [enthusiastic.au]; this was the only one not synthesized from the neutral token. Creakiness was added by superimposing a sawtooth pitch pattern, using Praat [creaky.au]. The pitch upturn was also created using Praat [upturn.au]. Vibrato was added using Sox [vibrato.au]. Finally, the creaky-elongated token was created by adding sawtooth pitch to the elongated token [creaky-elong.au].

Table 1 presents the rules as implemented. Each row shows how the system varied the tutor's response prosody according to the student's recent behavior: computationally the response (4th column) depends on the input (1st column), with the middle two columns serving as explanation. All conditions of a rule had to be true for it to apply. Rules were checked in order.

6. Experiment Set-up

To test our hypothesis, that suitably varying the prosody of acknowledgments does matter to users, we implemented the ruleset in Yesman, an experimental Wizard of Oz testbed in which the experimenter takes the role of the speech recognizer, but everything else is automated [1]. In particular, the system was responsible for producing negative utterances and hints, for computing the prosodic features, and for choosing the form and timing of the acknowledgments

Subjects interacted with two versions of the system, a baseline and one that chose the acknowledgment prosody according to the rules. Knowing that most users prefer varied to unvarying acknowledgments [1], we chose to use a baseline displaying the same prosodic variation, but with the variants chosen randomly with equal probability (except the "omit acknowledgment" option), without regard to the local context. Each subject used the systems to work through two moderately unfamiliar lists of US presidents. The pairing of systems with lists was balanced, as was the order of presentation. We recruited 22 subjects from the Introduction to Computer Science class and compensated with class credit. The procedure was as follows:

1. Subjects gave consent and were told that they would be interacting with two systems. They were then exposed to a previously recorded dialog, so that they understood what sort of dialog to expect. This was necessary because pilot studies showed that, without this, subjects felt awkward when interacting with the system for the first time.

2. Subjects had a minute to study the first list of presidents, which included for each a factoid, which was also one of the hints that the system would provide if they got stuck. Subjects then interacted with the first system. Subjects similarly studied the second list of presidents and then interacted with the second

	Naturalness initially	Friendliness initially	Friendliness after relistening
Random	**5.0** (1.5)	**5.3** (1.3)	**5.7** (1.3)
Rule-based	**5.6** (1.1)	**5.6** (1.5)	**6.0** (1.4)
significance	$p < 0.05$	$p \approx 0.27$	$p \approx 0.24$

Table 2: Ratings (and standard deviations). Significance was computed using one-tailed t-tests.

system.

3. Subjects rated each system in terms of naturalness on a 7-point scale, and indicated which they preferred.

4. Subjects listened to their interactions with each system. This was done because previously we had found that subjects could get so caught up in the game of recalling the items that they were not aware of the system's behavior. Later listening to their own interaction could help them recall how the dialog had been satisfactory or unsatisfactory, moment by moment [1].

5. Subjects again rated each system, this time on friendliness in addition to naturalness. We then probed their perceptions of the systems and the reasons for their choices, and finally debriefed and thanked them.

7. Results

In terms of the ratings, on every metric the rule-based system was preferred (Table 2), but only one difference was statistically significant: subjects' perceptions of naturalness expressed before relistening. In terms of overall preference, most of the subjects preferred the rule-based system (13 of the 21 (one subject being lost due to recorder error), but this was not statistically significant ($p \approx 0.097$ by the sign test)). Thus the hypothesis was supported, although not strongly.

The reasons subjects gave for their preferences were diverse, and revealed little or no conscious awareness, and occasionally mistaken impressions, of the differences between the two systems. We were worried that the baseline system, because its responses were random, might have produced one or two crashingly bad responses that dominated users' perceptions, but in fact no one mentioned specific good or bad responses. Thus it seems that their favorable judgments were based on the cumulative impression of many subtly appropriate acknowledgments.

8. Discussion

This work has shown how it is possible to dynamically adjust of prosody and shown that doing so can make the responses of a dialog system more appropriate and satisfying to users.

It is often said that *what you say* often matters less than *how you say it*, and in particular it seems likely that appropriate prosody can be more effective and less obtrusive than words for dialog functions such as conveying affect, enabling smooth interpersonal relations, and managing turn-taking. Although we did not test the advantage of prosody explicitly, support comes from comparing the results here with those of varying acknowledgments' lexical forms [2]: The ruleset developed for prosodic variation was easier to develop and simpler than that for lexical variation, and led to a stronger overall preference. The improvement in perceived naturalness was 0.3 for both systems (measured after relistening), but changes in the procedure (no item-by-item pausing during relistening, instructing the subjects to judge based on the quality of the acknowledgments, which was intended to guide them away from basing their judgments on their own memory performance, but which may have been misinterpreted as instructions to consider only naturalness of the acknowledgments as acoustic objects, without reference to the flow of the dialog) are likely to have led to understatement of the effect of the prosodic variations. Thus there is some support for the idea that prosodic variation is more effective than lexical variation.

This work indicates that current models of user states, of affect, and of the expressive uses of prosody do not tell the whole story. While the specific mappings and response rules identified here relate to known functions, including turn-taking and pacing control and the classic three dimensions of social interaction in dialog — dominance, interest, and valence — the specifics could not have been predicted from such models. Our findings are likely to have practical applications, for example, in commercial dialogs with a login phase in which the user is prompted for his or her name, account number, secret code, desired transaction type, etc. Inferring the user's knowledge, confidence, and willingness to control the pace of the interaction, and acknowledging that with suitable prosodic variations, as done here, might make such dialogs more efficient and easier for the users.

Although the specific rules developed here are probably of limited generality, we think that aspects of our method are likely to be of general value, especially the focus on bottom-up analysis, the focus on those aspects of prosody that respond to the user's state, and the focus on the dynamics of affective interaction on short time-scales, as they play out second-by-second and utterance by utterance. Future research using similar methods may help lead to spoken dialog systems that are more sensitive, more responsive, and more usable.

9. Acknowledgements

This research was sponsored in part by NSF Grant IIS-0415150 and by RDECOM via USC ICT. We thank David Novick, Jaime C. Acosta, Christian Servin, and Agustin Gravano.

10. References

[1] Ward, Nigel and Wataru Tsukahara. 2003. A Study in Responsiveness in Spoken Dialog. *International Journal of Human-Computer Studies*, 59 (6): 959-981.

[2] Hollingsed, Tasha K. and Nigel Ward. 2007. A Combined Method for Discovering Short-Term Affect-Based Response Rules for Spoken Tutorial Dialog. SLaTE. ISCA.

[3] Higashinaka, Ryuichiro, Kohji Dohsaka *et al.*. 2007. Effects of Quiz-style Information Presentation on User Understanding. Interspeech pp 2725-2728.

[4] Porayska-Pomsta, Kaska and Helen Pain. 2004. Providing Cognitive and Affective Scaffolding Through Teaching Strategies. in James C. Lester, Rosa Maria Vicari, Fabio Paraguacu (Eds.): *Intelligent Tutoring Systems*, Springer pp 77-86.

[5] Ward, Nigel. 2004. Pragmatic Functions of Prosodic Features in Non-Lexical Utterances. Speech Prosody '04, pp 325-328.

[6] Gravano, Agustin, 2009. Turn-Taking and Affirmative Cue Words in Task-Oriented Dialogue. Columbia Univ. Ph.D. Dissertation.

[7] Stocksmeier T., S. Kopp and D. Gibbon. 2007. Synthesis of prosodic attitudinal variants in German backchannel 'ja'. Interspeech.

[8] D'Mello, Sidney K., Scotty D. Craig *et al.* 2008. Automatic detection of learner's affect from conversational cues. *User Modelling and User-Adaptive Interaction*, 18(1-2), pp 45-80.

[9] Forbes-Riley, Kate and Diane Litman. 2007. Investigating Human Tutor Responses to Student Uncertainty for Adaptive System Development. ACII.