Factors Affecting Speaking-Rate Adaptation in Task-Oriented Dialogs

Nigel G. Ward and S. Kumar Mamidipally

Department of Computer Science
University of Texas at El Paso
nigelward@acm.org, neosrk@gmail.com

Abstract

Dialog systems should be able to automatically determine an appropriate speaking rate for each utterance. In a small corpus of billing support dialogs we identified factors accounting for 18% of the variation in agent speaking rate. Simple adaptation to the user accounts for little, rather it is the dialog state and dialog acts in the local context which seem to matter more. [1]

Index Terms: speech rate selection, tempo, utterance-level, user modeling, accommodation

1. Motivation

Today most spoken dialog systems do not adapt the rate of their speech output. This is a problem in that different users may prefer to hear speech at different rates, and in that they may prefer different rates at different places in the dialog. Non-adaptive speaking rate is largely unavoidable for recorded prompts, but it also seems that synthesized voices today are typically generated at a rate fixed for all utterances and for all users. To acheive adaptive speaking rate we need a model of what rate is appropriate in what circumstances.

This paper builds on an earlier study of speaking rate adaptation [1]. That work found two factors predictive of the agent's speaking rate: the speed of the user's initial response and the user's speaking rate. Specifically, in a corpus of simulated directory assistance dialogs in Japanese, the agent's speaking rate during number-giving was faster to the extent that the user had responded more swiftly and to the extent that the user had spoken more quickly. Multiple regression on these factors gave a formula which predicted the appropriate speaking rate, and the predictions correlated fairly well (.46) with the rates observed in good dialogs in the corpus. This work suggested that automatic speaking rate adjustment is feasible.

The current paper describes an exploration of whether speaking rate adaptation is also feasible for longer, more complex dialogs.

2. Method and Corpus

We set out to develop a predictive model of speaking rate, one using information about the course of the dialog so far to determine an appropriate rate for the next system utterance. Although factors affecting speaking rate in monologs and unstructured conversations have previously been identified [2, 3], our interest is in task-oriented dialogs, in which people communicate to accomplish some business.

Our approach was corpus based, using a billing-support corpus collected for another purpose [4], in which subjects were instructed to obtain balance information, to review recent transactions, and to make a payment. There were 20 students in the user role and one in the agent role. There were 733 utterances in total. All subjects spoke in American English, although some with heavy Spanish accents. The person chosen for the agent role was a student who had a pleasant manner, had customer service experience, and seemed generally socially adept.

An initial impressionistic review of this corpus revealed that the agent's speaking rate varied greatly, both across dialogs and within dialogs. This seemed to generally be deliberate, not a mere artifact of the words she was saying, and not a mere artifact of performance problems (although there were cases where she slowed as she fumbled to look up information for the customer).

3. Measuring Speaking Rate

The proper estimation of speaking rate is a matter of some complexity [5, 6, 7, 8], but many of the complicating factors probably are less significant for rate at the utterance level, our focus here. Preferring to have a convenient measure over a perfect one, we used the estimates of syllables per second given by the "mrate" program [9]. However mrate is unsuitable for very short utterances, so we excluded utterances lasting less than .5 seconds from the analysis, except where noted. We were also concerned that the values given by mrate may be misleading unless corrected for filled pauses, so we did a quick validation, labeling all utterances in the corpus using a 4 step perceptual judgment of rate. We found that even raw mrate had a .84 correlation with these judgments, which we decided was adequate for our purposes.

On this measure the agent's speaking rate averaged 4.31 and had a standard deviation of 0.67. The per-dialog averages ranged from 4.10 to 4.86.

4. User Speaking Rate as a Predictive Factor

Based on the Japanese study we expected the user's speaking rate to be a strong predictive factor. Specifically we expected the agent to talk faster in response to users who were talking faster. At the level of dialogs there was indeed a strong correlation, .60.

However, looking at the rates of individual utterances, the picture is more complex and the correlations far weaker. The correlation between the agent's speaking rate on an utterance and the user's speaking rate on the immediately previous utterance was a mere .025. The correlation with the cumulative average of the user's rate across all previous utterances was some-

[1] This research was sponsored in part by NSF Grant No. 0415150. We thank David Novick and Anais Rivera for discussion and Guarav Garg for data analysis.

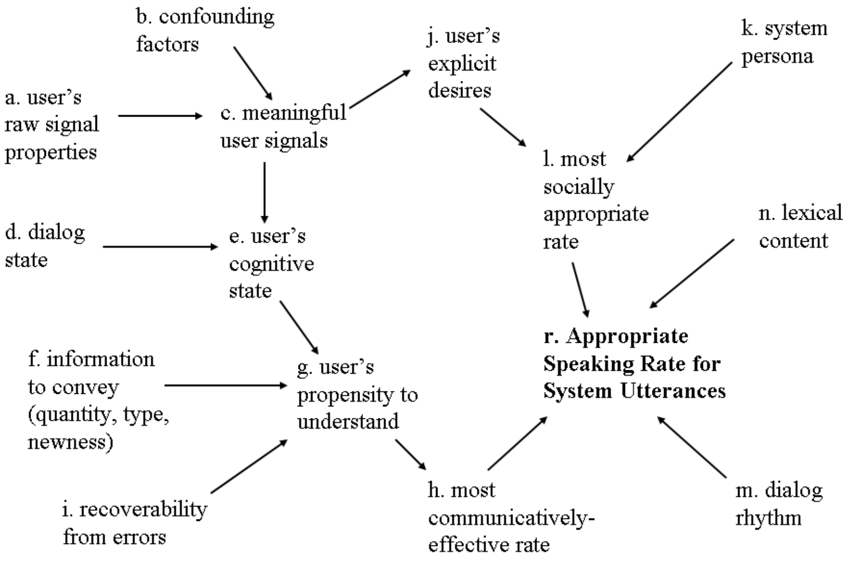
b. confounding factors
a. user's raw signal properties
c. meaningful user signals
j. user's explicit desires
k. system persona
l. most socially appropriate rate
d. dialog state
e. user's cognitive state
n. lexical content
f. information to convey (quantity, type, newness)
g. user's propensity to understand
r. Appropriate Speaking Rate for System Utterances
i. recoverability from errors
h. most communicatively-effective rate
m. dialog rhythm

Figure 1: Relations among Factors involved in Inference of the Appropriate Speaking Rate for the System. Observable and known factors are at the far left and far right.

what better: .13. Thinking that the pattern of adaptation might be obscured by the diversity of dialog acts in the corpus, we examined also one specific kind of utterance, that in which the agent reads out recent transactions (e.g. *you have a charge of $120 to Wal-Mart* . . .). As these were similar across all the dialogs, were among the longest agent utterances, and occurred near the end in almost all dialogs, we expected to see a clear correlation between cumulative average user rate and the operator rate on these utterances, however the correlation was negative: –.20.

At this point we realized that simple adaptation was inadequate to explain what was going on, and took a step back to consider what other factors might be involved.

5. Relations among Factors Involved in Speaking Rate Adaptation

Based on a literature review and our own observations, Figure 1 shows how agent speaking rate could be affected by user speaking rate and other factors, displaying some of the likely causal and inferential relations. In this section we discuss two reasons to expect agent speaking rate (r) to correlate with user speaking rate (a), and also discuss factors that may complicate this relation.

The most direct causal path, a-c-j-l-r, represents pure social accommodation [10, 11]. While it is conceivable that speakers have a reflex response to match their dialog partner's speaking rate, here we assume that such adaptation is generally mediated by inference. As seen in the figure, the user's choice of speaking rate (c) may reflect his (current) information-uptake capabilities or display his chosen dialog personality, such as whether he wants to be in control of the dialog. Such factors are represented as "user's explicit desires" (j) in the figure, and good agents can probably form a user model including these from the user's speech.

However such factors are not directly inferable from the raw speaking rate, which can be affected by extraneous factors (b), such as the time the user takes to formulate the utterance before he begins talking (the reaction time); momentary user confusion, for example when trying to find an account number to read aloud; and syntactic, lexical, and phonetic features such as the number of stressed syllables in the utterance. Normalization for typical user behavior is doubtless also required, as user speaking styles differ. These differences can be complex, for example it seems that some users handle formulation difficulties by reducing their speaking rate while others tend to start quickly but then pause and repair; here the information content and the impression given might be similar, but the raw speaking rate could be very different.

Another complication is that the socially appropriate rate (l) may be constrained by the persona of the dialog system (k). This may be determined statically by the corporate image that needs to be portrayed or by the need to model for the user how to talk slowly and clearly, so as increase the chances of successful speech recognition. It may also also be determined dynamically, for example an agent may deliberately talk more slowly to encourage the user to feel comfortable or calm down.

There is a second causal path by which the user's speaking rate (a) could affect the agent's speaking rate (r): by indirectly giving clues to the user's cognitive state (e). That is, people who are speaking faster can be assumed to be more alert and have a lighter cognitive load. A less loaded user would of course be generally more able to quickly understand (g), and as a result the most effective rate, that is the highest rate which would probably have little risk of causing the user to misunderstand (h), would be higher. There is a trade-off here — an agent or system should talk fast to save time, but not so fast as to cause time-wasting user misunderstandings — and we could in principle compute the rate which maximizes expected efficiency.

With reference to the figure, it is also possible to see why this might not turn out as we expected: other factors are likely to disrupt the correlation. For example the amount of new information to be conveyed in the agent's utterance (f) could affect how likely the the user would be to readily understand it (g).

It is also likely that aspects of the dialog state (d) — such as time into dialog (as a measure of the user's familiarity with the agent's voice), dialog act type, criticality (i), or the presence of recent misrecognitions [12] — could affect the user's cognitive load or his propensity to understand.

Apart from such user-related considerations, the agent's speaking rate could also be affected by the lexical content of the utterance (n) or considerations of dialog rhythm (m) such as desired utterance duration. In particular, an agent wishing to reduce the information density of an utterance may do so without changing speaking rate, by instead pausing more before talking, or by interpolating low content semi-fixed phrases, for example by prefacing some information with *okay, if you're ready I'll go ahead and read them off to you* rather than just *here they are*.

6. Other Predictive Factors

Based on this understanding of the likely relationships among the factors, ultimately we would like model the relations between the intervening variables and the observables. Doing this would, however, require numerous difficult subjective judgments, so we chose to do something simpler, to use Figure 1 merely to suggest what factors to examine, retaining the basic strategy of looking for correlations between observable dialog properties and the agent speaking rate. This section summarizes the tendencies seen; the details are in [13].

We thought that swifter user reaction times would correlate with faster agent speaking rates, as seen in the Japanese study [1]. As with user speaking rate, these could plausibly affect agent speaking rate both via a direct adaptation path and via an inferred representation of user cognitive state, assuming that users who respond faster are more alert. User reaction times were measured as the time from the end of the agent's previous utterance to the start of the user's utterance, with a lower bound of zero if this was negative, that is if there was an overlap, as happened 27% of the time. There was a positive correlation, of .09, meaning that if the user took longer to respond then the operator would speak more quickly on the next turn, contrary to expectation. Correlation with cumulative average reaction time so far was also positive, although weaker, at .014.

We also examined other properties of the user's utterances. We thought that disfluent user utterances would cue the agent to slow down, and there was a weak tendency for this. We thought that dispreferred user utterances, for example something other than an answer in response to a direct question from the agent, would cue the agent to slow down, but in fact there was no such effect. We thought that longer user utterances would enable the agent to speak faster on her next utterance, reasoning that this would give her more time to formulate her next utterance, and in fact there was a slight correlation, .06, between the rate of the subsequent agent utterance and the length of the previous user utterance.

We also looked for effects of the agent's cognitive state and cognitive load. There was a tendency for longer utterances to have a slower speaking rate (– .22 correlation), contrary to what is seen in unstructured conversations [3]. There was also a tendency for the agent to speak faster if she had more time to prepare, that is, a correlation (of .04) between her speaking rate and her reaction time (again, with overlapped utterances counted as having zero reaction time).

Finally we examined the effects of utterance type (dialog act), and here many correlations were found. There was a strong tendency for closings (e.g. *is there anything else I can help you with?*, and *thank you for calling*) to be much slower than the rest of the dialogs; the average rate was 3.6 versus 4.3 overall. There was a tendency for scripted prompts (e.g. *how may I help you?*, and *can I have your account number please?*) to be slightly faster than those generated on the fly. There was a tendency for answers to user questions to be slower. There was a tendency for the agent to speak slower when reading out the list of transactions. There was a tendency for the agent to speak faster after a very short user utterance, which in these dialogs was usually an acknowledgment of an item in a list, letting the agent know that she could continue on swiftly. There was a tendency for the agent to speak faster when producing utterances that started by acknowledging data provided by the user. There was a strong tendency for questions to be produced faster. This was probably due to the fact that, in this domain, agent questions generally bring little new information (e.g., *okay, now how much would you like to pay?*), although it may also be due in part to the fact that our agent, being bilingual in Spanish and English, may have been using the final rate speed-up sometimes used to mark questions in Spanish.

7. The Predictive Equation

Running multiple regression over a total of 17 factors, including all of those mentioned in Sections 4 and 6, we obtained a linear equation predicting agent speaking rate as a function of 14 terms [13]. This model explained 18% of the variance and had a significance F of $< .0001$, meaning that this performance was almost certainly not due to chance. However, given the small size of our corpus this probably represented overfitting, so we also created a simpler model, using only the five factors which individually had significant correlations ($p < .05$) with agent speaking rate:

Predicted Agent Speaking Rate

$$= 3.79 + .74T_2 + 0.68T_3 + 0.63T_4 + 0.82S + -0.04D$$

In this equation T_2, T_3, and T_4 are binary features indicating which subtask the user is currently engaged in, with the coefficients reflecting the relative speeds, T_2, responding to the balance inquiry being faster than T_3, eliciting payment instructions, and T_4, giving recent transactions. The lack of an explicit factor for T_5, the closing subtask, encodes the fact that it was by far the slowest. S is a binary feature that is true if the utterance was a scripted one: the positive weight here could reflect the ease of producing these, the prevalence of scripted prompts in simple question-answer exchanges, or the prevalence of such utterances early in the dialog. D is the duration of the agent utterance in seconds. This simpler equation had high significance (F $< .000001$) but accounted for only 15% of the variance, which is probably not enough to be directly useful.

8. Directions for Future Work

Our equations are unsophisticated linear combinations of factors which are obviously not independent, failing to reflect the likely interactions (Figure 1). Creating a quantitative model that includes the intervening variables is one promising direction for future work.

It might also be useful to consider the reciprocal process: not only must the agent adapt to the user, the agent may also need to consider how the user is simultaneously adapting to the agent.

Incorporating better characterizations of the dialog states and dialog acts would probably also be valuable. For example, the agent appeared to speed up to indicate completion of a sub-task, and she appeared to slow down when giving dispreferred responses, for example *uh, I don't have any information for that date*. With a finer characterization of states and acts and a larger corpus we could examine such possible relations. We also should of course re-examine the tendencies noted above using a larger corpus.

It may also be valuable to attempt to tie the speaking rate adaptation problem to the user modeling problem. Various sources of information useful for modeling the user's knowledge, needs, and desired interaction style (e, g, j) are likely to be useful here also. For example, his familiarity with the domain of discourse may be inferable from his vocabulary choice; his degree of understanding and comfort in the dialog may be inferable from his prosody; his desired pace of interaction may be inferable from his turn-taking style, for example the propensity to overlap and the propensity to back-channel; and his language comprehension ability may be inferable from his accent, although in this corpus a foreign accent in speech does not seem to reliably predict comprehension difficulties.

The method of analysis assumed that all utterances are equally informative, but this is probably not the case. On the one hand, properties of some utterances may reflect merely transient states. For example, after the user slows down at a disfluency point (due to formulation problems, or looking up a number), he often saliently speeds up, as if to say "that was just a momentary problem; now I'm fast and alert again"; his momentary slowing probably should not affect the agent's rate. Similarly a crisp *excuse me?* can cue a repetition, which should of course come slower, but this should not necessarily affect the way the agent behaves in subsequent utterances. On the other hand there may be "benchmark" utterances, perhaps greetings for example, in which the user's speaking rate or other speech properties reliably indicate the user's personality, state, and conversational style, and thereby "set the tone" for the whole conversation.

9. Summary and Conclusions

We set out to develop a model of agent speaking rate adaptation, suitable for determining how fast the utterances of a spoken dialog system should be. We identified factors which account for some of the variation in speaking rate.

However the model developed is only weakly predictive. The main reasons for this result, in contrast to the success in the earlier study, probably relate to the nature of the dialogs. The earlier study used directory assistance dialogs, which are almost the simplest possible task-oriented dialogs. Such dialogs have little variation on most of the factors seen in Figure 1. In a sense, the agents in those dialogs had so little information to go on that they could do no more than simple adaptations, and these were easy to model. Another likely partial explanation relates to the subject populations and recording environments: in the earlier study the subjects varied in occupation and age, called from a variety of environments and used a variety of telephony equipment, thus there were more large differences in user dialog style and in agent behavior, and these were easier to model.

Among the various factors affecting speaking rate, we expected the major determinants to be those related to adapting to the user. This was based on our belief that speaking rate is largely independent of the semantic and pragmatic course of the dialog, and is instead part of a separate dimension of social dynamics. Had this been true it would have been good news for spoken dialog systems development: it would have meant that we could build an autonomous plug-in enabling the dialog manager to chose an appropriate rate for each utterance, which would support the easy retrofitting of existing spoken dialog systems to do adaptive speaking rate selection.

But in fact the most predictive factors were those relating to the dialog state and the quantity, type, and newness of the information to convey. This suggests that proper control of speaking rate for task-oriented systems will probably depend on carefully modeling and representing the interactions between dialog acts and dialog state. The bad news is that this is can be domain-specific; the good news is that this can be done with existing technology.

10. References

[1] Ward, Nigel and Satoshi Nakagawa, 2004. Automatic User-Adaptive Speaking Rate Selection. *International Journal of Speech Technology*, 7, pp 235–238.

[2] Quené, Hugo. 2005. Modeling of between-speaker and within-speaker variation in spontaneous speech tempo. Interspeech 2005.

[3] Yuan, Jiahong, Mark Liberman, and Christopher Cieri. Towards an Integrated Understanding of Speaking Rate in Conversation. Interspeech 2006, pp 541-544.

[4] Ward, Nigel, Anais G. Rivera, Karen Ward and David G. Novick, 2005. Root Causes of Lost Time and User Stress in a Simple Dialog System. Interspeech 2005.

[5] Pfitzinger, Hartmut R., 1998. Local Speech Rate as a Combination of Syllable and Phone Rate. ICSLP 1998, vol. 3, pp. 1087-1090.

[6] Pellegrino, Francois, J. Farinas and J.-L. Rouas, 2004. Automatic Estimation of Speaking Rate in Multilingual Spontaneous Speech. Speech Prosody 2004, pp 517–520.

[7] Koreman, Jacques. 2006. Perceived speech rate: the effects of articulation rate and speaking style in spontaneous speech. *Journal of the Acoustical Society of America*, 119(1), pp 582-596.

[8] Dekens, Tomas, Mike Demol, Werner Verhelst, and Piet Verhoeve. A Comparative Study of Speech Rate Estimateion Techniques. Interspeech 2007.

[9] Morgan, Nelson and Eric Fosler-Lussier, 1998. Combining Multiple Estimators of Speaking Rate. ICASSP '98, pp 721-724. IEEE.

[10] Giles, Howard, Anthony Mulac, James J. Bradac and Patricia Johnson, 1987. Speech Accommodation Theory: The First Decade and Beyond. in *Communication Yearbook 10*, M. L. McLaughlin, ed., Sage Publications, pp 13–48.

[11] Buller, David B. and R. Kelly Aune. The Effects of Speech Rate Similarity on Compliance: Application of Communication Accommodation Theory. *Western Journal of Communication*, 56, pp 37–53, 1992.

[12] Dioubina, Olga I, 2004. Prosody of Dialogues: Influence of Recognition Failure on Local Speech Rate. Speech Prosody 2004, pp 275–278.

[13] Mamidipally, Soujanya Kumar, 2006. Speaking Rate Adaptation for Task-Based Spoken Dialogue Systems. University of Texas at El Paso, Computer Science Department Masters Thesis.