AUTOMATIC USER-ADAPTIVE SPEAKING RATE SELECTION FOR INFORMATION DELIVERY

Nigel Ward and Satoshi Nakagawa*

nigelward@acm.org, nakagawa@sanpo.t.u-tokyo.ac.jp
Mechano-Informatics, School of Information Science and Technology
University of Tokyo

ABSTRACT

Today there are many services which provide information over the phone using a prerecorded or synthesized voice. These voices are invariant in speed. Humans giving information over the telephone, however, tend to adapt the speed of their presentation to suit the needs of the listener. This paper presents a preliminary model of this adaptation. In a corpus of simulated directory assistance dialogs the operator's speed in number-giving correlates with the speed of the user's initial response and with the user's speaking rate. Multiple regression gives a formula which predicts appropriate speaking rates, and these predictions correlate (.46) with the speeds observed in good dialogs in the corpus. An experiment with 18 subjects suggests that users prefer a system which adapts its speed to the user in this way.

1. INFORMATION-GIVING BY VOICE

Many commercial telephone dialogs include an information delivery phase, in which the system gives the user information such as a time, a price, a password, directions, a transaction or confirmation number, etc. As far as we know, all IVR and spoken dialog systems today provide information either by playing back a fixed, prerecorded voice, or by using a synthesized voice generated with fixed parameters.

With information delivered at a single speed, invariant across users, it will be too fast for some users, such as nonnative speakers, children, and people in noisy environments, and too slow for others, such as business people in a hurry. In terms of time cost, if the speed is too slow there is a clear loss in user time, system time, and connection time; if the speed is too fast there is again a time loss, as the user has to wait for a repetition.

Whereas the other phases of commercial dialogs (the greeting, call routing, caller identification, content understanding) have been well studied, and are indeed key concerns in the interactive voice response (IVR) business and in spoken dialog research, the information delivery phase has received less attention.

2. USER ADAPTATION

In artificial intelligence, the question of how to adapt the system's output to the individual user is a classic research problem; here the field of user modeling has addressed the question of inferring the user's beliefs, desires, and knowledge in order to provide him with the information that will be most helpful to him. In natural language processing, the field of natural language generation is largely concerned with the problem of expressing a message using words and syntactic structures that the user can understand.

In human-human dialogs, the participants often adapt their production to each other's needs [1]. This fact has, however, not been put to use for information delivery. Schmandt's and Iwase's work [2, 3] shows how simple prosodic properties of the user's utterances can be used to decide when to repeat, wait, or play the next sentence in a sequence of directions. This work has not, however, addressed the question of adapting the speaking rate itself.

Tsukahara [4] showed that it is possible to detect the "ephemeral emotional state" of the user from the timing and prosody of his utterances, and that this information can be used to adapt the system's utterances to make them more pleasing to the user. This demonstration of swift adaptation, operating within a second or less, was the direct inspiration for the current work.

3. CORPUS

We started with some hunches, such as that slower information delivery would be preferred by foreigners, children, people in noisy environments, people who are tired, confused, distracted, or fumbling for a pencil, people living in rural areas, polite people, people in low-pressure occupations, and the elderly.

To test these, we gathered a corpus of directory-assistance-type dialogs, in Japanese. We chose to use directory-assistance-type dialogs primarily since they are short, which allowed us to gather many dialogs from many speakers in many conditions at relatively little cost. The second reason we chose directory-assistance-type dialogs is

*Ward is currently at the University of Texas at El Paso. Nakagawa is currently at IBM Japan. This work was supported in part by the International Communications Foundation, Tokyo, and by the Japanese Ministry of Education's Prosody and Speech Processing Project, headed by Keikichi Hirose.

operator: Directory Assistance, Suzuki speaking. 1
user: oh, hello. I'd like the number for the University of Tokyo, in Bunkyo-ku, Tokyo 2
operator: University of Tokyo, Bunkyo-ku, Tokyo? 3
user: yes. 4
operator: here is the number . . . 5
operator: . . . 03 3812 2111. 6

Fig. 1. A Directory Assistance Type Dialog (translated from Japanese)

that they are fairly consistent in structure, which simplifies analysis. Figure 1 is an example from the corpus.

The corpus was gathered for us by Arcadia.

57 "users" were recruited, chosen to exhibit variety in terms of age, sex, occupation, and native dialect or language. 5 were non-native speakers of Japanese. Most were Arcadia employees or consultants or their family members. From the users we gathered sex, age decade, language and accent history, occupation, presence of hearing impairments, degree of experience with standard directory assistance, namely NTT's 104, and a rating of the acceptability of the automatic number giving phase of this service versus human number-giving.

Users were requested to use the "service" 9 times, 3 times each in the morning, afternoon, and evening. This was intended to give some variety in terms of user's alertness level and feeling of tenseness/rush. However call times were at the user's choice, when he had free time, and thus there were probably no truly rushed calls. We also asked the users to use the "service" from at least two different telephones.

For each user, we prepared a sheet including 9 listings (some imaginary) that they had to get numbers for, such as the Sapporo Central Post Office and the Sendagi City Hall. The city name, and thus the exchange number of the number given, were always the same for each user. This allowed us later to easily sort the dialogs by user. Next to each listing was a blank for the user to write down the number given by the operator. There were also fields for the user to record, for each dialog, the telephone type used (PHS, portable, normal landline, public telephone), the location (home, office, outdoors, taxi, train station, etc), the time, and the user's impression of the operator's performance: good, normal, or bad.

The final corpus includes 508 dialogs, since some users called in less than 9 times. (One user called NTT's directory assistance number by mistake. This was detected after the dialogs were collated, as the user had not realized it at the time; for this user at least our "service" was apparently indistinguishable from "real" directory assistance.)

Each dialog was recorded onto two DAT tapes: one directly from the telephone line and one from a microphone on the operator's side. The operator's microphone also picked up some of the user's voice; thus both channels include both voices. This was convenient to set-up, and it allows, in principle, recovery of the individual voices [5], or at least use of the correlations between the two channels to automatically synchronize them, and use of the volume differences to automatically identify who was speaking when.

8 operators were recruited for us by Denwa-Hoso-Kyoku, where the recording took place. All operators had call-center experience, and all had professional-sounding voices and manners. The operator's task was to behave like a normal directory assistance operator, with the main difference being that the number for the listing requested was found by scanning a short list, rather than searching in a large database. Some operators later reported that, after a few calls, they started to recognize the voices of some of the users; however this did not appear to change their behavior.

Neither the operators nor the users were told the purpose of the experiment.

The dialogs were uploaded from DAT tape, converted to 8 KHz μ-law, and then chopped into .wav files, two per dialog (one for each channel). The .wav files were correlated with the data sheets from the users, and the data on each user and each dialog was entered into Excel. The numbers that users had recorded were checked, and all were correct; thus although the users had no need for the information given, all had taken the task seriously.

4. CORPUS ANALYSIS

Listening to the dialogs showed that the vast majority were fairly straightforward: there were few overt communication problems. In particular, there was little or no hyperarticulate speech [1]. There was nevertheless substantial variation in pacing. The data suggested that:

- Slower number-giving is preferable for users who speak slower, and conversely for faster speakers.
- Slower number-giving is preferable for those who react to the operator's greeting after a delay, and conversely for users who respond more swiftly.

To investigate these hypotheses, we chose to limit attention to the dialogs that users rated "good", reasoning that we wanted our system to model good operator performance, not just average or bad performance. We also chose to work only with dialogs without excessive noise, in order to make analysis easier. This left us with 142 dialogs to analyze. These dialogs, specifically the channels recorded from the telephone line, were labeled by hand.

In the corpus users' speaking rates ranged from 6 to 10 morae per second. For this we calculated the morae rate from the transcribed utterances. Filled and non-filled pauses lasting for more than 250ms were omitted from the denominator.

Users' initial response times, defined as the delay between the end of the operator's greeting (utterance 1 in Figure 1) and the start of the user's first utterance (utterance 2 in Figure 1), ranged from 50 to 1600 milliseconds.

To simplify the measurement of operators' information delivery rates, we further limited attention to the 75 dialogs where the user produced an acknowledgement after each group of digits. (Although this was the most common pattern, there were also 38 dialogs where the user repeated back each group of digits, 20 dialogs where the user listened to the number in silence, and 9 dialogs where the user repeated back some but not all of the digit groups). For these 75 dialogs our metric of information-giving slowness was then simply the overall duration of each number-giving (utterance 6 in Figure 1), including internal pauses and the user's interleaved acknowledgements. These durations ranged from 5 to 11 seconds.

The correlation between the user's speaking rate and operator's number-giving duration was –.25; the correlation between the user's initial reaction time and the operator's number-giving duration was .32. These two factors are combined in the following formula

$$L = m_1 R + m_2 D + b. \tag{1}$$

where
R is the user's speaking rate in $[morae/sec]$,
D is his initial reaction time in $[msec]$, and
L is the operator's number-giving duration in $[msec]$,
and the parameters, obtained by multiple regression, are
$m_1 = -355.95[msec \cdot sec/morae]$,
$m_2 = 1.50[]$, and
$b = 9048.25[msec]$.

5. EVALUATION

L given by formula 1 correlates fairly well (.46, correlation significant at $p < 0.01$) with the actual operators' number-giving durations. Thus information delivery was slower to the extent that the user spoke slowly and to the extend that he was slow to respond to the initial greeting.

To find out what other factors are involved, we listened to all cases where the number-giving duration given by the formula differed by more than 2 seconds from the actual duration in the corpus. We noticed three phenomena. First, in some dialogs the operator seemed to actively solicit acknowledgements, prosodically [6], which seemed to drive the dialog faster. Second, in some dialogs the operator paused after every digit. Third, in some dialogs the user's acknowledgements came slowly; sometimes it seemed that he had not intended intended to produce acknowledgements, but the operator had waited, forcing him to produce them anyway. As these factors are all under operator control, they would not raise problems in an automatic system.

6. SYSTEM

To see whether users would actually prefer speaking rate adaptation, we built a semi-automated directory assistance system. In this system, as in most directory assistance systems today, a human operator handles the call up to the point of the final information delivery. The novel aspect of our set-up is that the system listens in on the user-operator interaction (Figure 2) to compute the user's initial response time and his speaking rate, and then uses this to give the user the number at an appropriate rate.

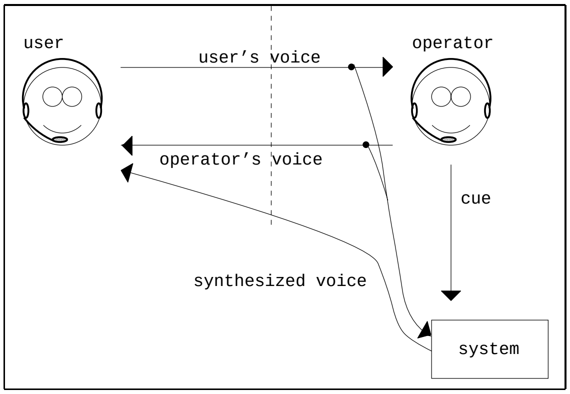

Fig. 2. Experiment Set-up.

Since speech recognition was not reliable we used Morgan and Fosler-Lussier's mrate [7] to estimate the user's speaking rate. mrate is known to correlate well (.67) with the transcribed speaking rate in English; using three speakers from a labeled corpus we found that it also correlates well (.68) for Japanese. Thus R in Equation 1 was computed using

$$R = m_r M + b_r. \tag{2}$$

where M is the value given by mrate, coefficient m_r is $2.76[morae/sec]$, and intercept b_r is $-5.55[morae/sec]$.

We are unaware of any research directly addressing the question of how to set the timing of information that the user is to write down. However it is clear that the pauses in the operator's number-giving are important [8], and that the relative duration of pauses increases as speakers strive for clarity [1].

To generate a number-giving voice of the duration given by the formula, we chose a simple rule of thumb, to use 40% of the total duration for the pauses between digit groups, becase this was approximately the corpus average. We then used the Fujitsu voice synthesizer [9], selecting the rate (using a value from 1 to 6) needed to produce an utterance of roughly the desired duration. Overall number-giving durations varied from 4.3 to 8.2 seconds.

Specifically, the "speed parameter" for the synthesizer was given by the formula below, obtained by regression:

$$S = round(m_L L + b_L). \tag{3}$$

where the round function is used to convert to an integer. Coefficient m_L is $-0.001275[1/msec]$ $(=-1.275[1/sec])$, and intercept b_L is $12.432[]$. Parameters which were too fast or too slow were scaled down to the nearest ordinary speed for this synthesizer, namely 1, 2, 3, 4, 5 or 6.

Thus, the predictive formula implemented in the system was:

$$S = round(m_L(m_1(m_r M + b_r) + m_2 D + b) + b_L). \quad (4)$$

Running this system on the corpus, we found a fair correlation (.41, correlation significant at $p < 0.01$) between the predicted values and operators' actual number-giving durations.

7. EXPERIMENT

Ultimately speaking rate adaptation should be tested in the field, with a large, varied population, in a real task. So far we have just run 18 students in the laboratory.

Subjects were given a list of 10 businesses and institutions, and asked to obtain and write down telephone numbers for any 9 from the operator. Presumably since the number was given by a synthesized voice, subjects tended not to repeat back or acknowledge the digit groups; thus the situation here did not exactly match the conditions under which the corpus was gathered, however this probably had little effect on subject perceptions regarding speaking rate.

Subjects used three systems: the one described above, one with an unchanging speaking rate, and one with backwards adaptation, which gave the information faster in cases where our formula predicted that the user would prefer it slower, and conversely. Systems were presented in random order. Subjects used each system 3 times; once in a normal way, once pretending to be in a rush, and once pretending to be elderly. After each dialog, subjects rated the suitability of the output speed on a scale from –4 to 4. Finally, after using each system for the 3 times, subjects rated the overall quality of each system. Further details appear elsewhere [10].

15 of 18 users ranked our system more highly than the invariant-rate system, and the difference was significant ($p < 0.01$ using the Wilcoxon matched-pairs signed-ranks test). Our system was similarly preferred to the system with backward adaption.

We attempted to identify the factors which account for some users preferring the non-adaptive system. In at least one case this was because the adaptation did not operate correctly. Our algorithm clearly needs some refinement and/or some sanity checks so that it backs-off to non-adaptive information-giving if the computed parameters are implausible. Also, in one case we think that the user was not truly evaluating the system from the perspective of an elderly person, which is understandable, as several subjects pointed out the difficulty of trying to evaluate a system from that perspective.

8. SUMMARY AND CONCLUSION

We have shown that simple, easily computable features of the user's voice can be used to adapt the system's speaking rate to be more appropriate.

Combination with other features, such as the time of day, environmental and channel noise levels, originating exchange, user's inferred dialect or accent, user's inferred age, and so on, would probably allow even better adaptation.

We expect that speaking rate adaptation will find utility in many automated and semi-automated IVR and spoken dialog systems.

9. REFERENCES

[1] Sharon Oviatt, Margaret MacEachern, and Gina-Anne Levow: Predicting Hyperarticulate Speech during Human-Computer Error Resolution. Speech Communication, 24, pp.87–110, 1998.

[2] Christopher Schmandt: Voice Communication with Computers, VNR Computer Library, pp.199-204, 1994.

[3] Tatsuya Iwase and Nigel Ward: Pacing Spoken Directions to Suit the Listener, International Conference on Spoken Language Processing (ICSLP-98), pp.1203-1206, 1998.

[4] Wataru Tsukahara and Nigel Ward: Responding to Subtle, Fleeting Changes in the User's Internal State, CHI 2001: Conference on Human Factors in Computer Systems, pp.77-84, 2001.

[5] Yosuke Matsusaka: Convclean. http://www.tk.elec.waseda.ac.jp/convclean/

[6] Nigel Ward and Wataru Tsukahara. Prosodic Features which Cue Back-Channel Feedback in English and Japanese. Journal of Pragmatics, 32, pp. 1177–1207, 2000.

[7] Nelson Morgan and Eric Fosler-Lussier: Combining Multiple Estimators of Speaking Rate, ICASSP-98, Seattle, pp.721-724, 1998.

[8] Masato Ishizaki and Yasuharu Den. Danwa to Taiwa (Conversation and Dialog). University of Tokyo Press, 2001.

[9] Linux Library for Japanese Voice Synthesis, http://www.createsystem.co.jp/linux.html.

[10] Satoshi Nakagawa and Nigel Ward. Adaptive Number-giving for Directory Assistance (in Japanese). in 40th Spoken Language Information Processing Workshop Notes (SIG-SLP-10), pages 25–30. Information Processing Society of Japan, 2002.