Automatic User-Adaptive Speaking Rate Selection

NIGEL WARD and SATOSHI NAKAGAWA
University of Tokyo[1]

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. It is therefore easy, at least in principle, to make systems which adapt their speed to users' needs.

Keywords: rate, speed, pace, adaptation, number-giving

1 Introduction

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 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 non-native speakers, children, and people in noisy environments, and too slow for others, such as business people in a hurry. There is a time cost either way: 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 waits for a repetition.

[1]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. We thank all who participated in the project, and also the anonymous reviewers of this paper.

2 User Adaptation

This section briefly surveys research in user-adaptive interfaces.

Common practice in interface design is to produce an interface that meets the needs of all members of the target user population. Any such generic interface will, however, be less than optimal for any individual user. One solution to this is to allow the user to personalize or customize the system's behavior to some degree, by explicitly stating his interests and preferences or by explicitly setting system parameters. A second solution is for the system to adapt itself, based on experience with the user.

Within user adaptation research there are two general approaches. The first involves explicit user modeling. This requires determining or planning what content he needs to know, and deciding how to convey it to him. This can be a very knowledge-intensive process, especially if the system aims to adapt after only a brief initial interaction with the user (Langley 1999). The second approach adapts without maintaining an explicit user model by keeping comparable information implicitly in the state of the dialog manager. One advantage of this approach is that the dialog model can be trained by reinforcement learning (Singh *et al.* 2002).

Assuming the system has obtained some idea of what it needs to convey, the next step is conveying it. The field of natural language generation includes a body of work which is concerned with the problem of expressing a message using words and syntactic structures that the user can easily understand (Reiter & Dale 2000; Walker & Rambow 2002). All of this work, however addresses adaptation at a fairly coarse level of granularity, at best that of the word, but more commonly that of the proposition or speech act.

In human-human dialogs, however, there is also adaptation at a much finer level: the participants often adapt their diction, pacing, timing, and tone of voice to meet each other's needs and cognitive abilities.

Deciding how to do such fine-level adjustments does not always require clever inference or exact user-modeling. Rather, in some cases, dialog participants provide clues as to how they would like the interaction to proceed. Since these often take the form of subtle prosodic cues, the details of how this is done are understood in only a few cases. Schmandt's and Iwase's work (Schmandt 1994; Iwase & Ward 1998) 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. Tsukahara (Ward & Tsukahara 2003) 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. For example, if the user is pleased with himself the system can produce a congratulatory acknowledgement, if the user is proceeding swiftly without problems the system can produce a short businesslike acknowledgement, and if the user is unsure the system can produce a firmly reassuring acknowledgement. Tsukahara's demonstration of swift adaptation, operating within a second or less, was the direct inspiration for the current work.

3 Corpus

To determine how a system should adapt the rate of information delivery, we gathered a corpus of human-human dialogs.

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

Figure 1: A Directory Assistance Type Dialog (translated from Japanese)

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. The corpus was designed to let us test some of these hunches. This was done by gathering a corpus with various kinds of variability. As such the corpus is narrowly useful for our purpose, correlation finding, and is not likely to be useful for such other purposes as determining which groups of users would in fact benefit from speaking-rate adaptation.

We chose to gather 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 that they are fairly consistent in structure, which simplifies analysis. Figure 1 is an example from the corpus. As we were gathering the corpus in Japan, we chose to mimic the format of the most popular Japanese directory-assistance service, namely NTT's 104. This follows the same pattern seen in Figure 1 except that the number reading, line 6 of the figure, is done not by the operator but mechanically.

The corpus was gathered for us by Arcadia, a Japanese company specializing in corpus development and other services for the speech systems industry.

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. We recorded users' sex, age decade, language and accent history, occupation, presence of hearing impairments, degree of experience with NTT's 104 service, 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 degree of busy-ness or haste. 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 Sendai 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 (using the rough classification of PHS, portable, normal landline, and public telephone), the location (home, office, outdoors, taxi, train station, etc), and the time. Finally there was a space for the user to mark his impression of the operator's performance, with the suggested responses being "good", "normal", and "bad".

The final corpus includes 508 dialogs, since some users called in less than 9 times. One user called NTT's 104 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, at least in principle, 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 Preliminary Corpus Analysis

Listening to the corpora, it was clear that there was substantial variation in pacing.

There were however few overt communication problems. In particular, there was little or no hyperarticulate speech (Oviatt *et al.* 1998).

Noise is known to cause speakers to talk more slowly (Summers *et al.* 1988). In the corpus there was some noise in some of the dialogs, however not enough to have a noticeable effect on intelligibility. The correlation between the signal-noise ratio for the user's voice and the operator's number-giving duration over a roughly labeled 289 dialog subset was significant but very low, 0.014 (r^2= .0002).

Listening to the data more closely suggested two hypotheses.

- Slower number-giving is preferable for users who speak slower, and conversely for faster speakers.

  This is an example of convergence or "accommodation" (Giles *et al.* 1987), as has often been observed in human-human dialog.

- 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.

5 Quantifying the Correlations

To investigate these two hypotheses, we wanted to examine only good dialogs, reasoning that we wanted our system to model good operator performance, rather than bad or even just average performance. Overall, 153 dialogs were rated good, 299 normal, and 3 bad, with the rest rated using free text. The significance of a "good" rating is open to question, as about a third of the users rated all of their dialogs the same, and there was clearly no consistency across users. Moreover, some judgements of "good" probably had little relation to dialog pacing, as the free responses included positive coments such as "operator was kind" and "lively" and "had a nice voice". (Incidentally complaints were mostly that the operator was too quiet (14 responses) or too "mechanical" (7 responses).) Despite these limitations of the ratings, we chose to use them and analyze only the dialogs rated "good".

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.

We measured speaking rates in morae per second, where a mora is roughly a syllable. There are various ways to count morae: we simply counted two morae for each double vowel and one mora for each single vowel, syllabic nasal, and geminate consonants. Although the relation between various metrics and perceived articulation rate is complex in general (Koreman 2003), and for Japanese in particular there are more sophisticated ways to relate mora counts to speech rate (Takamaru *et al.* 2000), this served as a convenient approximation. Filled and non-filled pauses, although clearly significant indicators of the speaker's state, were excluded from the computation, as they probably do not affect perceived speaking rate in any simple way. To be specific, pauses longer than 250 ms, and thereby unlikely to be of phonetic origin even for a fairly long geminate consonant closure, were omitted from the denominator. Users' speaking rates ranged from 6 to 10 morae per second.

We defined the "user's initial response time" to be 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). This ranged from 40 to 1600 milliseconds.

Measuring operators' number-giving times was complicated by the fact that there were various patterns. The most common was where the user produced an acknowledgement after each group of digits (75 dialogs). 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. To allow direct comparisons we restricted analysis to dialogs with the most common pattern. Our metric of information-delivery 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.

Insert Figure 2 about here

Figure 2: Correlation between the user's speaking rate (measured from the transcription) and the duration of the operator's number-giving.

Insert Figure 3 about here

Figure 3: Relation between subjective judgment of the user's speaking rate and the duration of the operator's number-giving.

There was a significant negative correlation between the user's speaking rate and operator's number-giving duration, –.25 (r^2= .06), as seen in Figure 2. To see whether it would be worth striving for a more accurate rate estimate, we labeled the users' rates on a scale from 1 to 9, based on the second author's subjective judgment. The correlation this gave was only slightly better, –.28 (r^2=.08, again significant), as seen in Figure 3.

The correlation between the user's initial reaction time and the operator's number-giving duration was positive and somewhat stronger, .32 (r^2 = .10), as seen in Figure 4.

Insert Figure 4 about here

Figure 4: Relation between the user's initial reaction time and the duration of the operator's number-giving.

These two factors can be 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

$b = 9048.25[msec]$.

For example, if the user's speaking rate is 8.25 morae/sec and his initial reaction time is 600 ms, then the predicted operator's number-giving duration is 7.0 seconds.

6 Evaluation

L given by Formula 1 correlates fairly well (.46, $r^2 = .21$, correlation significant at $p < 0.01$) with the actual operators' number-giving durations. Thus information delivery was indeed slower to the extent that the user spoke slowly and to the extent 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 predicted 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 (Ward & Tsukahara 2000), 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 to produce acknowledgements, but the operator had waited,

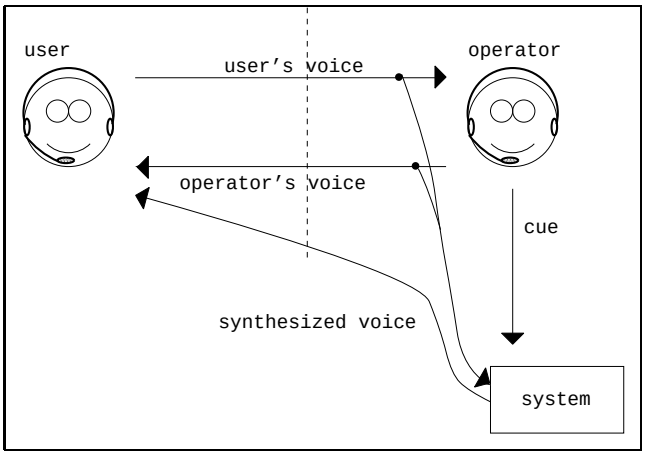

Figure 5: Experiment Set-up.

Insert Figure 6 about here

Figure 6: The Correlation between `mrate` and Transcribed Rate

forcing him to produce them anyway. As these factors are all under operator control, they would not raise problems in an automatic system.

7 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 5) 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.

Since speech recognition was not reliable we used Morgan and Fosler-Lussier's mrate (Morgan & Fosler-Lussier 1998) 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 (Juuten 1995) we found that

Insert Figure 7 about here

Figure 7: Cumulative Duration of Pauses between Digit Groups as a function of Overall Number-Giving Duration. Small boxes represent corpus data; large diamonds represent system behavior, with the synthesizer speed parameter at values from 1 (upper right) to 6 (lower left).

it also correlates well (.68, r^2=.46) for Japanese (Figure 6). 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[$morae/sec$]. For example, if mrate is 5, the infered speaking rate is 8.25 morae/sec.

To generate a number-giving voice of the duration given by the formula, we needed to determine the duration of each digit group and the duration of the pauses. The timing of digit sequences is fairly well understood, (Olaszy & Nemeth 1999), and we entrusted this to to the synthesizer. The duration of the pauses, although known to be important (Ishizaki & Den 2001) seems to be less well understood. It is known that the relative duration of pauses increases as speakers strive for clarity in difficult conditions (Oviatt *et al.* 1998), however in these dialogs we opted for the simple rule of using 40% of the total duration for the pauses between digit groups, slightly higher than the corpus average (Figure 7). We then used the Fujitsu voice synthesizer (Create System Development Company 2001), selecting the rate (using a value from 1 to 6) needed to produce an utterance of roughly the desired duration.

The duration of the digit groups was set by selecting the "speed parameter" of the synthesizer according to 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[]. Speed 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. For example, if the desired number-giving duration was 7 seconds, the total pause length would be 2.8 sec. and the total synthesized voice duration would be 4.2 sec., implying a speed parameter of 6.

Thus, combining equations 1, 2, and 3, 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, r^2=.17, correlation significant at $p < 0.01$) between the predicted values and operators' actual number-giving durations. Overall number-giving durations varied from 4.3 to 8.2 seconds (Figure 7).

8 Evaluation Issues

Ultimately speaking rate adaptation should be tested in the context of use, by real users. We have not yet done this. A pilot study (Nakagawa & Ward 2003) indicates two factors that need to be considered before such an experiment. First, the system needs a sanity check so that it backs-off to a standard speaking rate if the computed parameters are implausible. Second, when the number is given by a synthesized voice, users tend not to repeat back or acknowledge the digit groups; thus the actual use of the system will not exactly match the conditions under which the corpus was gathered. However this may not be a major problem, given that the major determinants of desired speaking rate are probably the times needed to hear and write down the information, which should not depend much on whether the user speaks or is silent. Subjectively, even naive implementation of the equations still gives roughly appropriate speaking rates even in dialogs where the users do not repeat or acknowledge (Ward & Nakagawa 2002).

9 Discussion

This section discusses future prospects and some remaining issues.

Although we have addressed speaking rate adaptation in the context of information delivery, it may also be useful in other contexts, such as prompting and audio browsing. Compared to techniques such as barge-in or explicit control of playback (Resnick & Virzi 1992), rate adaptation allows a factor of 2 speed-up with a simple implementation and without requiring the user to do anything special.

We have only looked at the most obvious factors of the user's speech; many others could be considered. F0 variation has been found to correlate with perceived "busy-ness" (Yamashita & Matsumoto 2002), and various contextual and prosodic features may correlate with perceived "hastiness" (Komatani *et al.* 2003). The duration of filled pauses (Goto *et al.* 1999) or their acoustic content or prosody (Ward 1998; Ward 2004) may indicate the user's degree of understanding or cognitive load. The user's vocabulary, dialect or accent, or inferred age, and also extra-dialog factors, such as time of day and originating exchange, may also be informative.

Our system uses the information in the user's speech, but for a pure interactive voice response (IVR) system it may be possible to do similar adaptation by considering the timing and rate of the user's keypad input. Users familiar with the system, for example, often press keys immediately after, or even during, the system prompt; such users would probably also welcome a faster speaking rate from the system.

Speaking rate adaptation is probably not universally a good thing, especially if taken to extremes (Suzuki 2001). For example, overadaptation in the direction of slow output may be perceived as patronizing (Giles *et al.* 1987), and variability may clash with the image or personality the system is designed to project.

It would be interesting to explore automatic speaking-rate adaptation for other languages.

10 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.

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

References

Create System Development Company (2001). Linux Library for Japanese Voice Synthesis. http://www.createsystem.co.jp/linux.html.

Giles, Howard, Anthony Mulac, James J. Bradac, & Patricia Johnson (1987). Speech Accommodation Theory: The First Decade and Beyond. In M. L. McLaughlin, editor, *Communication Yearbook 10*, pp. 13–48. Sage.

Goto, Masataka, Katunobu Itou, & Satoru Hayamizu (1999). A Real-time Filled Pause Detection System for Spontaneous Speech Recognition. In *Eurospeech '99*, pp. 227–230.

Ishizaki, Masato & Yasuharu Den (2001). *Danwa to Taiwa (Conversation and Dialog)*. University of Tokyo Press.

Iwase, Tatsuya & Nigel Ward (1998). Pacing Spoken Directions to Suit the Listener. In *International Conference on Spoken Language Processing*, pp. 1203–1206.

Juuten (1995). Monbusho (Japanese Ministry of Education) Juuten (Intensive) Research Project on Speech, Language and Concepts, Dialog Corpus volume 4. CD-ROM.

Komatani, Kazunori, Shinichi Ueno, Tatsuya Kawahara, & Hiroshi G. Okun (2003). User Modeling in Spoken Dialogue Systems for Flexible Guidance Generation. In *Eurospeech*, pp. 745–748.

Koreman, Jacques (2003). The Perception of Articulation Rate. In *International Congress of the Phonetic Sciences*, pp. 1711–1714.

Langley, Pat (1999). User Modeling in Adaptive Interfaces. In *Proceedings of the Seventh International Conferenc on User Modeling*, pp. 198–205. Springer.

Morgan, Nelson & Eric Fosler-Lussier (1998). Combining Multiple Estimators of Speaking Rate. In *ICASSP*, pp. 721–724. IEEE.

Nakagawa, Satoshi & Nigel Ward (2003). Adaptive Number-giving for Directory Assistance (in Japanese). *Human Interface*, 5:391–396.

Olaszy, Gabor & Geza Nemeth (1999). IVR for Banking and Residential Telepone Subscribers using Stored Messages Combined with a New Number-to-Speech Synthesis Method. In Daryle Gardner-Bonneau, editor, *Human Factors and Voice Interactive Systems*, pp. 237–256. Kluwer.

Oviatt, Sharon, Margaret MacEachern, & Gina-Anne Levow (1998). Predicting Hyperarticulate Speech during Human-Computer Error Resolution. *Speech Communication*, 24:87–110.

Reiter, Ehud & Robert Dale (2000). *Building Natural Language Generation System*. Cambridge University Press.

Resnick, Paul & Robert A. Virzi (1992). Skip and Scan: Cleaning Up Telephone Interfaces. In *CHI '92*, pp. 419–426. ACM.

Schmandt, Chris (1994). *Computers and Communication*. Van Nostrand Reinhold.

Singh, Santinder, Diane Litman, Micheal Kearns, & Marilyn Walker (2002). Optimizing Dialog Management with Reinforcement Learning: Experiments with the NJFun System. *Journal of Artificial Intelligence Research*, 16:105–133.

Summers, W. Van, David B. Pisoni, Robert H. Bernacki, Robert I. Pedlow, & Micheal A. Stokes (1988). Effects of Noise on Speech Production: Acoustic and perceptual analyses. *Journal of the Acoustical Society of America*, 84:917–928.

Suzuki, Noriko (2001). Social Effects on Vocal Rate with Echoic Mimicry using Prosdy-only Voice. In *Eurospeech*, pp. 2431–2435.

Takamaru, Keiichi, Makoto Hiroshige, Kenji Araki, & Koji Tochinai (2000). A Proposal of a Model to Extract Japanese Voluntary Speech Rate Control. In *International Conference on Spoken Language Processing*, pp. 255–258.

Walker, Marilyn A. & Owen C. Rambow (2002). Spoken Language Generation. *Computer Speech and Language*, 16:273–281.

Ward, Nigel (1998). The Relationship between Sound and Meaning in Japanese Back-channel Grunts. In *Proceedings of the 4th Annual Meeting of the (Japanese) Association for Natural Language Processing*, pp. 464–467.

Ward, Nigel (2004). Pragmatic Functions of Prosodic Features in Non-Lexical Utterances. In *Speech Prosody 04*.

Ward, Nigel & Satoshi Nakagawa (2002). Automatic User-Adaptive Speaking Rate Selection for Information Delivery. In *International Conference on Spoken Language Processing*.

Ward, Nigel & Wataru Tsukahara (2000). Prosodic Features which Cue Back-Channel Feedback in English and Japanese. *Journal of Pragmatics*, 32:1177–1207.

Ward, Nigel & Wataru Tsukahara (2003). A Study in Responsiveness in Spoken Dialog. *International Journal of Human-Computer Studies*, 59:603–630.

Yamashita, Yasuki & Hiroshi Matsumoto (2002). Acoustical Correlates to SD Ratings of Speaker Characteristics in Two Speaking Styles. In *International Conference on Spoken Language Processing*, pp. 2577–2580.