ON DIFFERENT FUNCTIONS OF REPETITIVE UTTERANCES

Marc Swerts[1], *Hanae Koiso*[2], *Atsushi Shimojima*[2] *and Yasuhiro Katagiri*[2]

[1]IPO, Center for Research on User-System Interaction, Eindhoven, The Netherlands
[2]ATR Media Integration & Communications Research Laboratories, Kyoto, Japan
swerts@ipo.tue.nl, {koiso,ashimoji,katagiri}@mic.atr.co.jp

ABSTRACT

The study reported in this paper focuses on different functions of echoing in Japanese dialogues. Echoing is defined as a speaker's lexical repeat of (parts of) an utterance spoken by a conversation partner in a previous turn. The phenomenon was investigated in three task-oriented, informal dialogues. Repeats in this corpus were labeled in terms of a 5-point scale which expressed the level to which a speaker had integrated the other person's utterance into his/her own body of knowledge. Kappa statistics showed that the labels could reliably be reproduced by three independent subjects. The investigation brought to light that the level of integration is reflected in a number of lexical and prosodic correlates. These features are discussed regarding their information potential, i.e., their accuracy and comprehensiveness as signals.

1. INTRODUCTION

Except for particular settings, the exchange of information through spoken language usually is not an exact data transfer process between a sender and a receiver. A person who talks to another cannot simply take it for granted that all his/her messages are completely understood by the other party. Communication failures may arise because there may be different types of noise on the channel, a speaker may overestimate the other person's knowledge about a given state of affairs, or the listener may simply not have been paying enough attention.

Therefore, dialogue partners constantly negotiate on the information being exchanged in the course of their conversation (Clark and Wilkes-Gibbs, 1986). This is clearly illustrated by the fact that the most prototypical example of language usage, i.e., daily-life conversation, is characterized by different signals that - strictly speaking - do not contribute to the content of the topic at hand, but carry information about the conversational process itself and thus serve to manage the dialogue (Shimojima et al., 1997). Some of these cues may be non-verbal, e.g., like the use of pointing and other hand gestures, nodding, gaze, etc. Conversants also use particular utterances to acknowledge receipt of a message, to repair it or to ask for clarifications.

The current research focuses on the pragmatic use of so-called repetitive or echoing utterances. Dialogues between two or more persons sometimes show instances where a speaker repeats what his or her conversation partner just said in a previous turn, in one way or another. Examples are given below in fragments (1) and (2) of dialogues between speakers A and B:

(1) A and then you transfer to the Keage line …
B Keage line
A which will bring you to Kyoto station

(2) A and that is the Keage line …
B Keage line?
A that's right, Keage line

Such repeats have intuitively clear pragmatic functions in everyday conversations and may even serve multiple goals. In general, they can be explained in terms of dialogue management behaviour. For instance, the purpose for repeating in (1) is probably to acknowledge that information has been received, in that sense being equivalent to the usage of a simple "uhuh" (integration), whereas the repeat in (2) signals a communication error, i.e., B appears to be unsure about the information provided by A and wants to have confirmation that he understood A correctly (non-integration). An important question is what differentiates various usages of repeats in actual conversations, so that conversational partners interpret them correctly. This study looks at prosodic features to see whether these correlate with communicatively different repeats.

2. HYPOTHESIS

From the point of view of information flow, the repeats in the examples above are very distinct. Speaker A's incidental miss of an acknowledgment such as in (1) does not necessarily lead to communication problems afterwards. However, to guarantee successful interaction, it seems more crucial that A really detects the request for repair in (2).

Since the repeats that flag a comprehension problem seem to be more important for communication, one would expect them to have prosodic features that are marked or prominent. This assumption could lead to the prosodic predictions summarized in Table 1, though it does not contain an exhaustive list. With respect to the length category of repeats, one might consider two alternatives. On the one hand, one could assume that repeats signaling non-integration are

Table 1: List of prosodic features and their expected settings for integrating and non-integrating repeats.

Features	Integrating	Non-integrating
pitch range	low	high
intonation	declarative	interrogative
delay	short	long
tempo	fast	slow

embedded in a longer utterance unit. Since it is essential that these utterances are picked up by the information giver, one could argue that they are more likely to occur in a turn in which the comprehension problem is flagged with additional lexical materials. On the other hand, it is also possible that the repeat is limited to the part of the other speaker's turn that was problematic, in other words explicitly focusing on the information for which a repair is requested, which would lead to relatively short repeats. Therefore, to find out which of these opposite expectations is valid, the current study will also deal with the immediate lexical context of repeats.

3. METHOD

3.1. Data

Analyses were based on three elicited Japanese dialogues recorded at ATR-MIC, each time between two male undergraduate students who knew each other, yielding about 45 minutes of speech in total. While they were seated in a sound-isolated studio, one participant was given the task to orally instruct the other on how to build a particular construction, like a 'duck', using differently coloured blocks. The result had to be similar to a construction shown on a picture, which only the instruction-giver could see. Both participants were allowed to gesture during communication, but the instructor could not physically touch any of the blocks. Using a head-set with microphone for both participants, the speech materials were recorded on different channels, so that even when their speech overlapped in time, their voices were separate on tape. The data were first fed into the computer with a 20-kHz sampling frequency and converted into waves+ format. Using the power measurements, the speech materials were automatically divided into 'utterance units' (UUs), defined as consecutive stretches of speech bounded by silence. The beginning and end time of each unit was extracted automatically.

3.2. Labeling

In these speech materials repeats were operationalized in the following way: "Let X be a sequence of UUs made in a single speaking turn, and Y be another sequence of UUs made in the turn following that. Then, X and Y are echoic pairs if and only if a sequence of morae that occupies 50% or more of Y already appears in X or is a semantic paraphrase of X". Next, the dialogue act specification of the different repeats was made by means of a consensus labeling between the three authors affiliated with MIC. To this end, they could listen to the speech and read the transcribed texts of the repeats as often as they liked and take any dialogue context into account, until consensus was reached. The repeats were rated in terms of the degree to which the receiver had integrated the given information into his body of knowledge using a 5-point scale, which ran from 'non-integrated' to 'fully integrated'.

To test the reliability of the labels obtained in this way, 35 utterances of the corpus, i.e., 7 instances randomly selected form the 5 integration categories, were presented to three independent subjects, who were not informed about the purpose of the research. They were instructed to rate the degree of integration on a 5-point scale. Kappa coefficients (k) were calculated between the original consensus labels and those obtained from each of the three independent subjects. Calculations were performed under a "strict match" and a "loose match" condition. For the former, only completely identical ratings were considered to indicate agreement, whereas for the latter, up to one point differences were allowed. In this way, an average pairwise k score of 0.58 was obtained for the strict match, and of 0.84 for the loose match. Given that a value of 0.8 or higher is generally regarded as indicating agreement with a high reliability, one may conclude that the labels can reliably be reproduced if the agreement criterion is slightly weakened.

3.3. Selected features

Both categorical and continuous variables were taken into account. The former were obtained by manual labeling, and comprised specifications of length category and boundary tone:

- *Length category.* Repeats were classified into 'short repeats' and 'long repeats', the former being defined as those that were shorter than the repeated UU, the latter as those that were equal or longer.
- *Boundary tone.* Intonation of the repeats was labeled in terms of their final boundary tone (Venditti, 1995); there appeared to be one set of high-ending contours: the simple rise (H%) and the fall-rise (L%H%), and another set consisting of low boundary tones: the simple fall (L%) and the rise-fall (L%HL%).

The following continuous features were obtained automatically:

- *Pitch register.* Pitch register was measured as the F_0 mean per utterance unit.
- *Tempo.* The normalized average mora duration per utterance unit was chosen as a measure of articulation rate. Using the transcriptions of the speech data, phone labels were automatically time-aligned first. After the phones were further grouped into a smaller set of morae, the normalized mora durations were calculated.
- *Delay.* Delay was measured on the basis of the automatically obtained beginning and end times of the utterance units. In particular, the time distance was calculated between the offset of the repeated fragment and the onset of the repeating fragment. In this way, a negative number reflects overlap, and a positive number a delay.

Since the distribution of the continuous, prosodic variables was sometimes skewed, the original values were converted by logarithmic transformation to satisfy the normality of the distribution.

Table 2: Number of low versus high boundary tones as a function of the different levels of integration.

	Low BT	High BT	% Low BT
1	5	8	38.46
2	11	5	68.75
3	12	7	63.15
4	6	4	60.00
5	13	0	100
Total	47	24	

Table 3: Number of short versus long utterances as a function of the different levels of integration.

	Short	Long	% Short
1	3	10	23.08
2	8	8	50.00
3	7	12	36.84
4	8	2	80.00
5	7	6	53.85
Total	33	38	

4. RESULTS

4.1. Descriptive analysis

Tables 2 and 3 give the data regarding the dependency of integration level on the two categorical variables, boundary tone and length type. Table 2 shows that - overall - utterances provided with a low boundary tone are more frequent than those with a high tone. However, looking at the extreme values of the integration continuum reveals that the low tones are more typical for repeats that signal a high degree of integration, whereas non-integrating repeats are provided more often with a high boundary tone. This finding is in agreement with the predictions. The data for the distribution of the two length categories are less clear, since there is no comparable increase or decrease of the relative frequency of either of the categories as a function of integration level. Therefore, the data do not allow one to choose between the two alternatives discussed in the hypothesis section regarding the lexical context of repeats.

Turning to the discussion of the continuous variables, the results for pitch, tempo and delay are visualized in Figure 1. In general, one can observe that level of integration is reflected in each of these three features: the respective average values decrease as a function of integration level. There are of course some slight differences between the features in that the decrease for delay is not strictly monotonic, whereas it is for the other two. Also, the decrease for pitch is less extreme than for delay and tempo. Summarizing, one may conclude that high pitch, long delay and slow tempo are more likely to reflect non-integrating repeats, and vice versa. They are all in the expected direction, more prominent/marked features being more typical for the repeats that flag a (potential) communication problem.

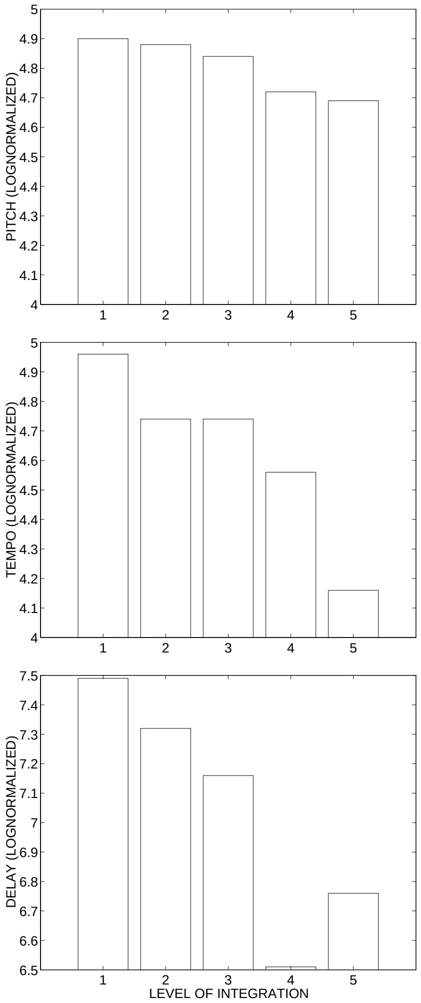

Figure 1: Log-normalized values of pitch (top), tempo (middle) and delay (bottom) for different levels of integration

4.2. Information potential

To analyse the signaling value of the different features in relation to integration level, the repeats were further explored in terms of two concepts borrowed from information-retrieval theory, i.e., comprehensiveness and accuracy. The former is a measure of the coverage of the information signaled, the latter refers to the correctness of the signaling (see also Koiso et al., 1997). More precisely, given a supposed cuing from a feature α to another feature β, then its accuracy is computed as the frequency of $\alpha \wedge \beta$, divided by the frequency of α, and its comprehensiveness as the frequency of $\alpha \wedge \beta$, divided by the frequency of β. In the current analysis, β reflects the level of integration, whereas α refers to one individual prosodic variable or to a particular combination of features (see below).

The following procedure was applied. First, due to the sparsity of the data, the original 5-point scale was reduced to a two-fold distinction between integrating and non-integrating repeats. Though

Table 4: Accuracy and comprehensiveness for signals of integrating and non-integrating repeats (pt=pitch; dl=delay; tp=tempo)

Integrating?	Optimal cue combination	Acc.	Comp.
No	$[tp>.2] \vee [pt>.2] \vee [dl>.2]$	87.8%	89.6%
Yes	$[tp<.2] \wedge [pt<.2] \wedge [dl<.2]$	66.7%	76.9%

the original scale could be split up in different ways, the current paper will focus on a division that takes the original categories 1-3 to represent one class (non-integration) and 4-5 another (integration). This binary categorization appeared to be a very "natural" one in that it coincided with significant differences for each of the categorical and continuous features, whereas the other possible divisions (e.g. 1 versus 2-5, or 1-2 versus 3-5) were reflected in only one or two significant differences. See Shimojima et al (1998) for an analysis which also explores other divisions.

Second, an algorithm automatically computed the accuracy and the comprehensiveness of the signaling relation between integrating/non-integrating repeats and particular (combinations of) prosodic features. To prevent the explosion of the search space, only combinations of maximally three features were considered. The combinations could be of a *conjunctive* or *disjunctive* nature, giving possibilities such as $\alpha^1 \wedge \alpha^2 \wedge \alpha^3$, $\alpha^1 \vee \alpha^2 \vee \alpha^3$, $\alpha^1 \wedge (\alpha^2 \vee \alpha^3)$, etc. Also, all the continuous features were transformed to standard scores, i.e., as deviations from the mean divided by the standard deviation, and the values -0.3, -0.2, -0.1, 0, 0.1, 0.2 and 0.3 were taken as different thresholds.

Given these conditions, table 4 gives the best results in terms of accuracy and comprehensiveness for the two types of repeats. First, the table reveals that the cues for non-integrating repeats are sufficiently accurate and comprehensive, whereas the signaling of integrating cues is rather weak. Second, as expected, only the non-integrating repeats are signaled by means of prosodic features that are "marked", i.e., features that are prominent because they have higher than average values. Finally, non-integrating repeats are best signaled by a *disjunctive* combination of marked settings for the continuous features pitch, tempo and delay, meaning that the presence of only one such feature is sufficient to signal a non-integrating repeat. Integrating repeats are (moderately) cued by a *conjunction* of "low" values for the same features that all need to be present at the same time. A more detailed informational analysis of echoic responses, including a discussion of their relation to grounding acts, can be found in Shimojima et al. (1998).

5. DISCUSSION AND CONCLUSION

Summarizing the results of this investigation, it appears that echoing utterances in Japanese, informal dialogues may serve at least two distinct communicative goals: (1) to signal that information has been integrated successfully by the receiver, or (2) to express the fact that he/she has some difficulty integrating it into his or her body of knowledge. Phonetic measurements reveal that these repeat categories are reflected in different prosodic and lexical features: the non-integrating cases are more likely to have one or more marked prosodic variables. Explorations of their information potential brought to light that these features have significant signal capacity in terms of accuracy and comprehensiveness, especially as cues to non-integrating repeats. The results thus show that repeats are potentially useful in spoken communication, because they represent different dialogue management acts: they function as different types of evidence from the receiver about the information being presented by the communication partner.

This leads one to reflect on the differences between human-human and human-machine interaction. In the former, all the conversants are very much aware of each other's limited resources, which is demonstrated by the fact that they constantly seek and provide evidence about mutual mental beliefs (Brennan, 1990). Repeating is a clear example of such communicative behaviour. In this respect, it seems unrealistic to expect spoken dialogue systems to act as 'perfect' communication partners that can achieve errorless understanding, since there will always be types of noise that are too severe to be solved by machines. Alternatively, to make the interaction with spoken-dialogue systems more efficient, it might be worthwhile to model particular strategies that are typical of human-human interaction, such as repeating, which has proven useful in handling the intrinsic uncertainty of spoken communication.

6. ACKNOWLEDGMENTS

Marc Swerts is also affiliated with the Flemish Fund for Scientic Research (FWO) and with Antwerp University (UIA). The research reported upon in this paper was carried out while he was a visiting researcher at Department 2 of ATR-ITL in summer 1997. Marc Swerts would like to thank Nick Campbell and Norio Higuchi for giving him the opportunity to visit their lab, and Nick Campbell for his constant help. Further thanks to Harald Singer, Kyohko Shimoda, Satoshi Kitagawa, Miwako Kurihara and Yoshinori Michijiri for technical assistance and labeling efforts.

7. REFERENCES

1. Brennan, S. Seeking and providing evidence for mutual understanding. Ph.D. thesis (unpublished), Stanford University, Stanford, CA, USA. 1990
2. Clark, H.H. and D. Wilkes-Gibbs. "Referring as a collaborative process". *Cognition*, 22, 1-39. 1986.
3. Koiso, H., A. Shimojima, and Y. Katagiri. "Informational potentials of dynamic speech rate in dialogue." *Proc. 19th Annual Conference of the Cognitive Science Society*, Stanford, CA, USA, August 7-10, 1997, 394-399.
4. Shimojima, A., Y. Katagiri and H. Koiso. "Scorekeeping for conversation-construction." *Proceedings of the Munich Workshop on Formal Semantics and Pragmatics of Dialogue*, Mundial-97, Munich, Germany, March 10-12, 1997, 172-194.
5. Shimojima, A., H. Koiso, M. Swerts and Y. Katagiri. "An informational analysis of echoic responses in dialogue." *Proc. 20th Annual Conference of the Cognitive Science Society*, Madison,WI, USA, August 1-4, 1998, 951-956.
6. Venditti, J.J. "Japanese ToBI labeling guides." Technical Report. Columbus, OH: Ohio State University. 1995.