AN EXPERIMENTAL STUDY ON THE INFORMATIONAL AND GROUNDING FUNCTIONS OF PROSODIC FEATURES OF JAPANESE ECHOIC RESPONSES

Atsushi Shimojima[1] *Yasuhiro Katagiri*[2] *Hanae Koiso*[3] *Marc Swerts*[4]

[1]Japan Advanced Institute of Science and Technology
[2]ATR Media Integration & Communications Research Laboratories
[3]The National Language Research Institute
[4]IPO, Center for Research on User-System Interaction

ABSTRACT

Echoic responses, which reuse portions of the texts uttered in the preceding turns, abound in dialogues, although semantically they contribute little new information. Earlier, we conducted a corpus-based analysis on echoic responses occurring in real-life dialogues, and examined their informational and dialogue-coordinating functions in connection with their temporal/prosodic features. The present study attempts to complement this observational approach with an experimental approach, where particular prosodic/temporal features of echoic responses can be studied in a more controlled and focused manner. In combination, the two lines of analyses provide an evidence that (1) echoic responses with different timings, intonations, pitches, and speeds signal different degrees in which the speakers have integrated the repeated information into their prior knowledge, and (2) the dialogue-coordinating functions of echoic responses vary with the speaker's integration rates signaled by these prosodic/temporal cues.

1. INTRODUCTION

An *echoic response* is an utterance in which a speaker reuses a portion of the text uttered by another in the preceding turn. We invariably do this when we talk, though we know semantically it contributes little new information. This paper studies the functions of echoic responses from the *dialogue-coordinating* perspective and the *informational* perspective.

From a dialogue-coordinating perspective, we ask how echoic responses in a dialogue contribute toward the coordination of the dialogue to a specific goal, particularly to the goal of *information-sharing*. Clark and Shaeffer [3] separate out the information-sharing aspect of the coordinating functions of utterances as their *grounding* functions. Traum [12] lists seven different "grounding acts" including acknowledgment and repair-initiation, that may be performed in an interactive dialogue. Though they consider only acknowledgment for echoic responses, Beun [1] and Walker [14] suggest that both acknowledgment and repair-initiation should be admitted to the variety of grounding functions of echoic responses.

From an informational perspective, we ask what information is carried by the occurrence of an echoic response during a conversation. Even if an echoic response adds little information to the *topic* of a proceeding conversation, it may still carry significant information at the *meta-level*, namely, information concerning the conversation process *itself*, as opposed to the topic of the conversation [5, 4, 8].

Since an echoic response is a reuse of an already uttered text, both the informational contribution and the grounding function of an echoic response must originate from its features other than its text. This characteristics of an echoic response makes it a good candidate for investigating the contributions of prosodic and temporal features of speech to the proceeding of a dialogue. Our main hypotheses to be examined in this paper are the following:

Main Hypotheses

A. The prosodic and temporal features of an echoic response carry information about the degree in which the speaker has integrated the repeated information into her body of knowledge.

B. An echoic response can have more than one grounding functions due to the variability of the speaker's integration rates signaled by these cues.

Suppose a speaker says, "Then go to Keage station," and another speaker responds by saying, "Keage." The first speaker is trying to give a piece of information about where the second speaker should go for the next destination. At the time of producing her echoic response, however, the second speaker may or may not have succeeded in assimilating the part of the information that she repeats, namely, the part represented by "Keage," with the body of her prior knowledge in a consistent manner. Hypothesis A above claims that the degree in which she has succeeded in this, is signaled by the prosodic/temporal characteristics of her utterance. Hypothesis B goes further and claims that an utterance of "Keage" in this timing potentially has different grounding functions, depending on what the above prosodic/temporal cues signal about the speaker's integration rate.

Earlier [11, 9], we conducted an *observational* analysis, designed mainly to test Hypothesis A. We compiled a corpus of spontaneous spoken dialogues, with various tags pertinent to our purpose, extracted instances of echoic responses from this corpus, and investigated their behaviors insofar as they were recorded in the corpus. This approach had the advantage of allowing us to see actual examples of echoic responses in a least biased setting, and thus putting appropriate constraints on our view of their functions. We will present a part of this earlier research in section 2, as far as it is relevant to the purpose of this paper.

In section 3, we will start presenting newly conducted research based on the *experimental* method, which works complementarily with the observational analysis. Unlike the observational approach, this approach artificially modifies a particular feature of echoic responses, and investigates the contrast between two groups of echoic responses that differ minimally in that feature. Although this approach sacrifices the naturalness of the samples under investigation, it allows us more focused and controlled investigation into the particular aspects of echoic responses we are interested in.

Our goal is to combine the results of these two lines of analyses and to see to what extent our main hypotheses are to be maintained. We will undertake in sections 4 and 5 separate evaluations of Hypotheses A and B in this light, as well as discussions of the remaining issues.

2. OBSERVATIONAL APPROACH

2.1. Methods

Data We conducted an analysis on actual occurrences of echoic responses extracted from a corpus of dialogue data we earlier collected. Our corpus consists of two-party face-to-face task-oriented dialogues in Japanese in which the participants engage in block construction tasks in a sound-isolated studio, where one participant (*instructor*) verbally gives instructions, referring to a set of pictures for target block configurations, to the other participant (*constructor*), who in turn tries to recreate the configurations out of the set of blocks available to her. Both the target pictures and the blocks were kept invisible from the other party until both sides agreed that they had completed the constructions. Both participants were allowed to make gestures while communicating, but the instructor could not physically touch any of the blocks.

We analyzed three dialogues, each between two participants familiar with one another. The speech materials from both participants were digitally recorded on separate channels, and transferred to a computer at a sampling frequency of 20KHz. They were subsequently divided automatically by power measurements into "Utterance Units (UUs)," consecutive stretches of speech bounded by silence.

Echoic Response Repeats can be classified according to a number of different criteria. They can be classified in terms of who makes the repeats, into self-repetitions, or into other-repetitions. They can be classified in terms of forms of repeats, ranging from an exact repetition to a paraphrase. They can also be classified in terms of the number of intervening turns before them, or into immediate and delayed repetitions.

For the present study, we focused on immediate other-repetitions, e.g., *echoic responses*. Taking the UU as the unit of analysis, "echoing" was operationalized in the following way:

> A sequence of UUs (X) made in a turn and another sequence of UUs (Y) made in the directly following turn are *echoic pairs* if and only if a sequence of morae that occupies a half or more of Y has already appeared in X or is a semantic paraphrase of a part of X.

We imposed two further conditions to guarantee that repeats are genuine instances of echoic responses. First, only repeats coming from the responder were considered, and "initiates" and "repairs," which do not constitute responses to previous utterances, were excluded. Secondly, we omitted repeats in standardized opening/closing sequences, such as those in greetings, e.g., "mosi-mosi"(hello). Given these restrictions, the definition given above resulted in a total of 71 repeat occurrences in our corpus.

Integration Rating We assigned, to each instance of the echoic responses, an information integration rating, which is a measure for the degree in which the responder had integrated the repeated information into her body of knowledge. Integration rating involves a 5-point scale ranging from minimal integration (score 1) to full integration (score 5). Ratings were first made by means of a consensus labeling among three of the authors. Both the speech and transcription of each instance were presented to them, which they examined until a consensus was reached. To test the reliability of the labelings so obtained, they additionally conducted a follow up experiment, in which seven instances of repeats were taken randomly from each of the five integration categories and were subjected to integration ratings by three subjects (two females and one male). Ratings were made several times to guarantee the stability of the rate assignment, and the last ratings were compared with those obtained in a consensus labeling operation.

Prosodic/Temporal Features For prosodic and temporal features of speech, we considered the following four features, which we think are the most significant in their dialogue functions. They cover categorical and continuous features. Categorical features were obtained by manual labeling, and continuous features were obtained through automatic procedures.

Boundary Tone: Repeat instances were categorized in terms of their final intonation patterns. A variant of J-ToBI [13] labels was assigned to repeat instances by an independent researcher who was not aware of the purpose of the current research. We made a simple distinction between high-ending contours, which include a simple rise (H%) and a fall-rise (L%H%), and low-ending contours, which include a simple fall (L%) and a rise-fall (L%HL%).
Pitch Registers: Pitch registers, which refer to the fact that utterances can be made in a low voice or in a high voice, were measured as the F_0 mean per utterance unit.
Tempo: The normalized average mora duration per utterance unit was chosen as a measure of the articulation rate. Using transcriptions of speech data, mora labels were first automatically time-aligned, and average mora durations were calculated and normalized with respect to durational variations among vowels.
Delay: Delay was measured as the duration between the offset of repeated fragment and the onset of a repeating fragment. A large negative number reflects overlap, whereas a large positive number reflects a considerable delay.

2.2. Results

Labeling Reproducibility The reliability of a labeling scheme is a basic, but often hard to confirm, requirement in corpus-based research. The kappa coefficient of agreement [10], which takes into account chance level biases, has been widely accepted by many researchers as one of the most useful measures of such reliability [2, 6]; a value of 0.8 or higher is generally regarded as indicating agreement with a high reliability.

We calculated κ coefficients between integration rate labels obtained in the consensus labeling and those obtained from each of the three independent subjects. Calculations were performed under the "strict match" criterion and the "loose match" criterion. For the former, only strictly equal ratings were considered as indicating agreement, whereas for the latter, up to one point differences were deemed to indicate agreement. We obtained an average pairwise κ score of 0.58 for the strict match, and 0.84 for the loose match[1]. These results showed that even though the inter-labeler reliability for the integration ratings was not high enough for strict five category distinction, we could claim a sufficiently high inter-labeler reliability by slightly weakening the rating agreement criterion.

[1] The loose match condition guarantees a higher observed value than the strict match condition, but it also gives a lower expected value, so we cannot say that the loose match condition necessarily produces higher κ coefficients.

Table 1: Distributions of boundary tones relative to integration ranges.

	[1]	[2345]	[12]	[345]	[123]	[45]	[1234]	[5]
L%	5	42	16	31	28	19	34	13
H%	8	16	13	11	20	4	24	0

Table 2: Distributions of continuous features of echoic responses.

	Temp		Delay		Pitch	
	mean	s.d.	mean	s.d.	mean	s.d.
[1]	4.96	0.66	7.49	0.52	4.90	0.36
[2345]	4.61	0.73	7.00	1.04	4.80	0.25
[12]	4.84	0.52	7.40	0.46	4.89	0.26
[345]	4.55	0.82	6.88	1.18	4.76	0.27
[123]	4.80	0.50	7.31	0.39	4.87	0.28
[45]	4.40	1.01	6.65	1.56	4.70	0.22
[1234]	4.76	0.47	7.17	1.00	4.84	0.27
[5]	4.28	1.34	6.76	0.70	4.69	0.25

Prosodic/Temporal Features and Integration We next looked into the question of whether and to what degree the five prosodic and temporal features, taken individually, of echoic responses reflect the degree of information integration of the responder. To that end, we applied statistical tests to see if we could find statistically significant distributional differences of feature values between integration and disintegration responses.

We first categorized echoic responses into integration and disintegration categories based on the consensus labeling of integration rates. There are four different ways to divide the 5-point scale of integration ratings into binary integration/disintegration categories: [1]-[2345], [12]-[345], [123]-[45], and [1234]-[5]. We examined all of these possibilities.

For the categorical feature, boundary tone, we applied a χ^2 test for distributional differences. Table 1 gives the distribution of features between the integration and disintegration responses. Results of the χ^2 tests are shown in Table 3. The tables show that, for boundary tone, there are significant distributional differences in three out of four possible divisions of integration/disintegration. The results also indicate that a high boundary tone is more probable in disintegration responses.

For the continuous features, tempo, delay, and pitch, we applied t-tests for distributional differences. Original feature values were first converted by logarithmic transformation to satisfy the normality of the distribution. Table 2 summarizes the values of the mean and standard deviations of these continuous features. Slow tempo, long delay and high pitch tend to be associated with disintegration responses. Table 3 summarizes the results of the t-tests. The results show that for all three continuous features, there are significant distributional differences in the [123]-[45] division; further differences are also found in [1234]-[5] for tempo, and in [12]-[345] for delay.

These results clearly indicate that the four prosodic and temporal features examined here reflect the degree of information integration, suggesting the possibility that they play important roles in actual dialogues with their signaling potentials.

3. EXPERIMENTAL APPROACH

Our experiment had two parts. As with the observational analysis reported above, experiment 1 aiming at determining (1) whether the prosodic/temporal features of echoic responses have any potential to signal the degrees in which their speakers have integrated the repeated information. Experiment 2 goes further and addresses the issue of the grounding functions of echoic responses. In particular, we examine (2) whether and to what extent the integration rates associated with echoic responses determine their grounding functions, and (3) how the prosodic and temporal features of echoic responses are directly related to their grounding functions.

Table 3: Statistical tests for differences between integration and disintegration.

		[1] [2345]	[12] [345]	[123] [45]	[1234] [5]
B.T.	$\chi^2(1)$	5.47*	2.66	4.09*	8.13**
Tempo	$t(69)$	1.61	1.64	2.20*	2.19*
Delay	$t(69)$	1.62	2.23*	2.75**	1.38
Pitch	$t(69)$	1.28	1.94+	2.55*	1.88+

$^{**}\,p < .01 \quad ^{*}\,p < .05 \quad ^{+}\,p < .1$

3.1. Methods

Experiment 1 In the first experiment, we artificially manipulated prosodic/temporal characteristics of the echoic response portions of the speech, and examined the effects they have on the signaling potentials of the speaker's degree of information integration. Speech analysis, modification and synthesis program STRAIGHT [7] was employed to manipulate speech characteristics of the echoic responses and to obtain high-quality synthesized speech. Among the prosodic/temporal characteristics of speech, we chose to manipulate *tempo*, *delay* and *pitch* because of their ease of controlled manipulation[2] The manipulated speech materials were then presented to a set of subjects to obtain information integration ratings.

Based on the statistical analysis on all 71 instances of echoic responses, the high and the low target values were determined for each of the three features. Total of 27 instances (9 for each feature) of echoic responses were chosen and speech materials were constructed out of them together with their surrounding utterances. Manipulation of speech materials was performed by the following procedure: (1) Analyze each speech material with STRAIGHT speech processing package; (2) Create a high value sample and a low value sample by setting the selected feature parameter for the echoic response portion to the high and to the low target values and by setting the non-selected feature parameters to their average values; (3) Synthesize target speech materials with the modified parameters. Manipulated speech materials were presented in random order to a group of subjects, who were then asked to rate them in terms of the 5-point scale information integration ratings. Both speech and its transcription were presented for ease of comprehension. A total of 16 subjects, 8 males and 8 females, participated in the experiment.

Experiment 2 The second experiment was designed to address the issue of grounding functions of echoic responses. The subjects were asked to assess the grounding functions being performed with the given instances, rather than the integration rates of their speakers.

A total of 19 instances were chosen, out of the 71 echoic responses in our corpora, by taking into account the distributional balance

[2] Even though we fully realize the importance of the boundary tone feature, e.g., it exhibited the strongest correlation with integration ratings in our observational analysis, and it has also been pointed out [15] that intonational contour contributes to determining the functions of redundant utterances, we decided not to manipulate boundary tone feature in this study. This is because there simply are too many variables in constructing intonational contours to reliably explore all the possible intonational patterns and to come up with a pair of minimally contrasting patterns. Identifying the structure of the space of intonational contours is itself an interesting research issue.

Table 4: Distributional differences of mean integration scores for the two conditions on prosodic and temporal features.

		25%	50%(Median)	75%
tempo	slow	2.75	3.10	3.40
	fast	3.12	3.45	3.67
delay	long	2.90	3.50	3.75
	short	3.60	4.00	4.45
pitch	high	2.75	3.10	3.25
	low	3.00	3.25	3.77

of both integration rates and temporal/prosodic features. Out of each instance, the "negative" sample was created by changing the text of the UU directly preceding the echoic response. This modification was realized by completely replacing utterances of one participant with utterances produced by another speaker who mimic the original speaker's speech in all respect except for the textual content of the target UU. The negative version of a sample containing the following exchange would be created by substituting A's utterance with "ano hiroi houo maemukini site (Well...the larger side should face front)." B's response, "uemukini site (Face up)," would then cease to be an accurate echoic response to A's preceding instruction, while all of the temporal and prosodic features remain untouched.

(1) A: ano hiroi houo uemukini site
(Well...the larger side should face up.)
B: uemukini site
(Face up.)

The reason why we took pains to create these negative versions is to compare the subjects' judgments of grounding function in the positive cases and the negative cases. We are concerned with the possibility that the subjects are inclined to assess an echoic response to perform an acknowledgment whenever it correctly duplicates the information given before, and to perform a request-repair whenever it duplicates the information incorrectly. We want to see whether and to what extent the subjects' judgments of the grounding function of an echoic response is affected by this factor of "informational accuracy," as opposed to the integration rate and the prosodic/temporal features associated with the instance. All positive and negative samples were combined and made into two complementary sets of samples. They were presented in random order to a group of 16 subjects, who were then asked to assess whether the speaker is trying to do an acknowledgment or a request-repair.

3.2. Results

Prosodic/Temporal Features and Integration Table 4 shows the distributional differences of mean integration rate scores obtained from our subjects for the two target conditions on each of the three features: tempo, delay and pitch. The figures in the table are the integration rate scores at 25%, 50% and 75% points in the respective distributions. Wilcoxon signed-ranks tests were applied to the data, and significant differences were found in all three features (tempo: $z = 2.35, p < .01$, one-tailed test; delay: $z = 1.71, p < .05$, one-tailed test; pitch: $z = 2.73, p < .01$, one-tailed test). Fast tempo, short delay and low pitch were rated significantly higher in terms of information integration signaling than slow tempo, long delay and high pitch, respectively. These results clearly confirmed our findings obtained in our corpus-based observational analysis.

Table 5: Distributions of the judgments of repair-request (Req-rep) and acknowledgment (Ack) in relation to the speakers' integration ranges; the results are shown separately for the positive samples, the negative samples, and the entire samples.

	Positive		Negative		Entire	
	Req-rep	Ack	Req-rep	Ack	Req-rep	Ack
[12]	52	12	54	10	106	22
[3]	19	13	22	10	41	23
[45]	5	51	13	43	18	94

Table 6: Distributions of the judgments of repair-request (Req-rep) and acknowledgment (Ack) in relation to the boundary tones (B.T.), tempos, delays, and pitches of echoic responses.

		Positive		Negative		Entire	
		Req-rep	Ack	Req-rep	Ack	Req-rep	Ack
B.T.	H%	54	18	54	18	108	36
	L%	21	59	35	45	56	104
tempo	slow	58	38	66	30	124	68
	fast	18	38	23	33	41	71
delay	long	31	33	38	22	69	59
	short	45	43	51	37	96	80
pitch	high	50	30	58	22	108	52
	low	26	46	31	41	57	87

Integration and Grounding Acts Tables 5 and 6 show the subjects' judgments of grounding acts in experiment 2, where the judgments on the positive versions and the negative versions of the samples are displayed separately, along with the merged results.

In Table 5, the judgments are sorted horizontally by integration ranges associated with the assessed instances. As the merged result shows, there are clear distributional differences of grounding-act judgments over different ranges of integration rates: echoic responses with higher integration rates tend to be judged to perform acknowledgment while those with lower integration rates tend to be judged to perform request-repair ($\chi^2(2) = 110.34, p < .01$). Furthermore, this distributional difference is preserved both in the positive samples ($\chi^2(2) = 63.91, p < .01$) and the negative samples ($\chi^2(2) = 47.77, p < .01$).

Prosodic/Temporal Features and Grounding Acts Table 6 shows the subjects' judgments on the grounding functions in relation to the prosodic/temporal features of the instances being assessed. The χ^2 test shows significant distributional differences for boundary tones ($\chi^2(1) = 47.21, p < .01$), tempos ($\chi^2(1) = 21.20, p < .01$), and pitches ($\chi^2(1) = 22.69, p < .01$) of the entire samples, indicating that echoic responses with rising boundary tones or slow tempos or high pitches tend to be judged to perform repair-request, while those with falling tones or fast tempos or low pitches tend to be judged to perform acknowledgment. The same level of distributional differences are also found in the positive cases and the negative cases separately, indicating that the observed correlation overrides the difference in contextual accuracy of the instances. On the other hand, no significant distributional difference are found for delay, either in positive, negative, or entire samples.

4. HYPOTHESES EVALUATION

Let us review the observational and the experimental results presented so far, and examine to what extent our main hypotheses are supported by them in combination.

4.1. Hypothesis A

Hypothesis A claims that the prosodic and temporal features of an echoic response carry information about the degree in which the speaker has integrated the repeated information into her body of knowledge. Do our results support this claim?

Delay Our observational analysis shows, for the actual 71 instances of echoic responses in our corpus, that there is a strong tendency ($p < .01$) that the speaker's integration is in the range [123] when an echoic response start with a longer delay, while the integration is in the range [45] when the response starts with shorter delay. An analogous tendency was found for the division [12]-[345], although with a moderate strength ($p < .05$). Thus, a longer delay can be used as a fairly reliable signal to a lower degree of integration, and a shorter delay as a fairly reliable signal to a higher degree of integration.

In fact, when we created minimal pairs of echoic responses that differ only in length of delay, we found a strong tendency ($p < .01$) that our subjects assess the speakers' integration highly for the samples with shorter delays and lowly for those with longer delays. Since our subjects had no other cues than length of delay that differentiate the minimal pairs, we must conclude that short delays functioned as cues to high integrations while long delays functioned as cues to low integrations.

Thus, in the case of delay, both observational and experimental analysis strongly support Hypothesis A.

Pitch and Tempo Our observational analysis shows that there is a moderate tendency ($p < .05$) that the speaker's integration rate is in the range [123] when an echoic response has a higher pitch or a slower tempo, while the integration is in the range [45] when a response has a lower pitch or a faster tempo.

Our experimental analysis provides a stronger evidence. Our subjects evaluated the speaker's integration differently in the cases where two instances differ only in their pitches or tempos, while there is a strong uniformity in their differentiation ($p < .01$): evaluations are uniformly low for those with higher pitches or slower tempos and high for those with lower pitches or faster tempos.

Thus, in the cases of pitch and tempo, our experimental analysis strongly supports Hypothesis A, while the support provided by our observational analysis is more moderate.

Boundary Tone In our observational analysis, we found a strong tendency ($p < .01$) that the speaker's integration rate is in the range [1234] when an an echoic response has a rising boundary tone, while the integration is in the range [5] when an echoic response has a falling boundary tone. Similar tendencies, though more moderate ($p < .05$), were also observed for integration range pairs [1][2345] and [123][45]. Although we could not obtain an experimental confirmation of this tendency, due to technical difficulties involved in creating good minimal pairs of echoic responses that differ only in their boundary tones, our observational analysis strongly support Hypothesis A in the case of boundary tone.

4.2. Hypothesis B

Hypothesis B claims that an echoic response can have more than one grounding functions due to the variability of the speaker's integration rates signaled by its temporal and prosodic features.

The results of experiment 2 show a clear distributional difference ($p < .01$) of judgments on acknowledgments and repair-requests over the three integration ranges [12], [3], and [45], where instances falling in the lower integration range [12] tend to be judged to perform repair-request, while those falling in the higher integration range [45] tend to be judged to perform acknowledgment. This strong tendency is preserved in the group consisting only of positive samples, and in the group consisting only of negative samples. This indicates that echoic responses are judged to perform acknowledgment even when they fail to be accurate duplications of the information given in the preceding turn, and to perform repair-request even when they duplicate the given information. Thus, the speakers' integration rates are a dominant factor that divide the subjects' classifications of acknowledgment and repair-request, irrespective of the contextual accuracy of the echoic responses being assessed.

Now, our observational analysis and experiment 1 have already confirmed that certain temporal and prosodic features of echoic responses have the potential of signaling the speaker's integration rates. Thus, the above finding completes the two-step connection asserted by Hypothesis B, namely, the connection from temporal/prosodic features to integration rates and to grounding acts. Moreover, experiment 2 also shows that there is a direct correlation from boundary tones, pitches, and tempos to our subjects' judgments on acknowledgment and repair-request.

5. DISCUSSIONS

Thus, both hypotheses A and B are supported by our observational and experimental results, indicating that with their informational potentials to the speakers' integration rates, the temporal and prosodic features of echoic responses contribute to the performance of repair-request and acknowledgment. This, however, does not mean that for every echoic response, single temporal/prosodic features (such as long delay, high pitch, or slow tempo) uniquely classify its grounding function into acknowledgment or repair-request. In particular, there remain two important possibilities, namely, the possibilities of combinatorial signals to integration rates and of neutral echoic response.

5.1. Combinatorial Signals

Although long/short delays, high/low pitches, rising/falling tones, and fast/tempos were all shown to have some informational potentials to the relevant speaker's integration rate, it is possible that none of these temporal/prosodic features signal a sufficiently narrow range of integration required for the performance of acknowledgment or repair request, and that the required range is signaled only when such a temporal/prosodic feature is combined with some other features.

To understand how this is a possibility, assume a single prosodic/temporal feature, say long delay, indeed signals the speaker's integration to be in the range [345]. Then certainly, long delay has an informational potential to the speaker's integration rate, as our observational and experimental analyses indicate. But suppose that for an echoic response to act as acknowledgment, the speaker must be signaled to be in a narrower range of integration, say [5]. That is, the speaker's integration must be signaled to be "very high," rather than just "not low." In this case, the signal provided by long delay is not strong enough to let the echoic response perform acknowledgment. Thus, it may well be the case that long delay contributes to the performance of acknowledgment only when it co-occurs with some other feature that complements it to signal the sufficiently narrow integration range [5].

All these assumptions are compatible with our findings in this paper, since the statistical analyses we applied to the relationship between prosodic/temporal features and integration rates are designed to check the existence of global distributional differences, and are just not fine-grained enough to determine the exact ranges of integration rates signaled by individual prosodic/temporal features. Likewise our statistical analyses on the relationship between integration rates and grounding functions fail to specify what range of integration rates need be signaled for an echoic response to perform acknowledgment or repair-request. It seems important, therefore, to fill this gap in future research and to seriously explore the informational potentials of multiple prosodic/temporal features to signal narrower integration ranges.

5.2. Neutral Cases

Given the above considerations, it can be easily seen to be possible that for some echoic response, no features or combinations of features signal sufficiently narrow ranges of integration for the performance of acknowledgment and repair-request. Obviously, the grounding act performed by such an utterance should not require the same strong signals as required for these standard acts. In fact, an observational analysis reported in [9] positively suggested the existence of a large number of such instances, and we characterized the grounding function performed by those instances as *display*. There, the act of display is a non-committal and conditional act, which demands, figuratively, "take this as an acknowledgment if the repetition is correct, otherwise take this as a request-repair." Due to this non-committal nature, the act of display can be performed without any strong signal to the speaker's integration or disintegration, and thus it appears to be a natural grounding function to be attributed to those "neutral" echoic responses.

In fact, experiment 2 described above was originally conceived to establish the validity of this notion of display. The expectation was that if our samples contain a significant number of instances acting as display, then their conditional nature should increase the number of acknowledgment judgments in the positive condition and increase the number of request-repair judgments in the negative condition. But this prediction was not born out by the experiment. The judgments of acknowledgment and request-repair were almost uniform across the positive and the negative conditions.

This could be either (1) because the prosodic and temporal characteristics of speech determine the kinds of grounding acts (including display) much more strongly than we expected, and the subjects' judgments themselves are not affected by the textual contexts surrounding the echoic responses in question, or (2) because changing the text conditions produces other types of grounding functions, such as repair, in some of the negative condition instances, and might have obscured the subjects' judgments.

In any case, no matter whether these assessments based on our previous work may turn out to be true, it is important to realize that interesting issues still remain concerning the possibility of neutral instances and their possible grounding functions. Furthermore, although we have been focusing our attention on backward-looking grounding acts such as acknowledgment or repair-request, the possibilities (1) and (2) above suggest that echoic responses could perform *forward-looking* grounding acts, such as initiate and repair. The effect of prosodic and temporal features of echoic responses upon these new candidates of grounding functions presents us another interesting set of issues.

6. CONCLUSIONS

On the basis of our earlier observational analysis and our new experimental analysis, we evaluated two hypotheses, claiming (a) the potentials of the prosodic and temporal features of echoic response to signal the degrees in which the speaker has integrated the repeated information into her body of knowledge, and (b) the co-variability of the grounding functions of echoic responses with the speakers' integration rates indicated by prosodic and temporal cues. We could confirm (a) for delay, pitch, tempo, and boundary tone, and the claim (b) was also strongly supported. Our results, however, leave rooms for the possibility of echoic responses for which no strong information is available about the speakers' integration rates, and for the possibility of more than two features of echoic responses complementing each other to signal stronger information.

7. REFERENCES

1. R. J. Beun. "The function of repetitions in information dialogues." Technical Report 20, IPO annual progress report, 1995.
2. J. Carletta. "Assessing agreement on classification tasks: the kappa statistic." *Computational Linguistics*, 22:249–254, 1996.
3. H. H. Clark and E. F. Schaefer. "Contributing to discourse." *Cognitive science*, 13:259–294, 1989.
4. B. Grosz and J. Hirschberg. "Some intonational characteristics of discourse structure." In *Proceedings of ICSLP*, pages 429–432, 1992.
5. J. Gumperz. "Contextualization and understanding." In A. Daranti and C. Goodwin, editors, *Rethinking context*. Cambridge University Press, Cambridge, 1991.
6. J. Hirschberg and C. H. Nakatani. "A prosodic analysis of discourse segments in direction-giving monologues." In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pages 286–293, 1996.
7. H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne. "Restructuring speech representations using STRAIGHT-TEMPO: Possible role of repetitive structure in sounds." IJCAI-97 Workshop on Computational Auditory Scene Analysis, 1997.
8. H. Koiso, A. Shimojima, and Y. Katagiri. "Informatinal potentials of dynamic speech rate in dialogue." In *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society*, pages 394–399. Lawrence Erlbaum Associates, 1996.
9. A. Shimojima, H. Koiso, M. Swerts, and Y. Katagiri. "An informational analysis of echoic responses in dialogue." In *Proceedings of the 20th Annual Conference of the Cognitive Science Society*, pages 951–956, 1998.
10. S. Siegel and N. J. Jr. Castellan. *Nonparametric statistics for the behavioral sciences*. McGraw Hill Text, New York, 2nd edition edition, 1988.
11. M. Swerts, H. Koiso, A. Shimojima, and Y. Katagiri. "On different functions of repetetive utterances." In *Proceedings of the 5th International Conference on Spoken Language Processing*, pages 483–486, 1998.
12. D. R. Traum. "A computational theory of grounding in natural language conversation." Technical Report 545, University of Rochester, 1994.
13. J. J. Venditti. "Japanese tobi labelling guides." Technical report, Ohio State University, 1995.
14. M. A. Walker. "Redundancy in collaborative dialogue." In *Fourteenth International Conference on Computational Linguistics*, pages 345–351, 1992.
15. M. A. Walker. "Inferring acceptance and rejection in dialogue by default rules of inference." *Language and Speech*, 39-2:265–304, 1996.