An Informational Analysis of Echoic Responses in Dialogue

Atsushi Shimojima (ashimoji@mic.atr.co.jp)
Hanae Koiso (koiso@mic.atr.co.jp)[1]
ATR Media Integration & Communications Research Laboratories; Seika Soraku Kyoto, 619-02 Japan

Marc Swerts (swerts@ipo.tue.nl)
IPO, Center for Research on User-System Interaction; P.O. Box 513 5600 MB Eindhoven, The Netherlands

Yasuhiro Katagiri (katagiri@mic.atr.co.jp)
ATR Media Integration & Communications Research Laboratories; Seika Soraku Kyoto, 619-02 Japan

Abstract

Echoic responses abound in dialogues, where a speaker reuses a portion of the text uttered by another in a preceding turn, though semantically they contribute little if any new information. The phenomenon has attracted the attention of researchers from diverse academic fields, ranging from sociolinguistics and developmental psychology, to computational linguistics and human-computer interfaces. This study reports an empirical investigation on echoic responses from an informational perspective. Drawing on statistical analyses of instances extracted from corpora of spoken dialogues in Japanese, we show that echoic responses with different timings, lengths, intonations, pitches, and speeds signal different degrees in which the speakers have integrated the repeated information into their prior knowledge. We further consider dialogue-coordination functions enacted by this informational potential of echoic responses, and identify the function of *display* as distinguished from the functions of acknowledgment and repair-initiation.

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. The general theme of this paper is the functions of echoic responses in dialogues. We can distinguish three different perspectives toward this general theme: *social*, *dialogue-coordinating*, and *informational*.

From a social perspective, we ask how the occurrence of an echoic response in a conversation creates or otherwise changes the social circumstances among participants of the conversation. Tannen (1994) describes the creation of *interpersonal involvement* by the repetition of prior text. Norrick (1994) argues that echoic responses play an important role in interactional achievement of joking.

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 their contributions to the process of *information-sharing*. Clark and Shaeffer (1989) separate out the information-sharing aspect of the coordinating functions of utterances as their *grounding* functions. Traum (1994) 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 (1995) and Walker (1992) 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 still may carry significant information at the *meta-level*, namely, information concerning the conversation process *itself*, as opposed to the topic of the conversation (Gumperz, 1991; Grosz & Hirschberg, 1992; Koiso, Shimojima, & Katagiri, 1996).

Our approach toward echoic responses in this paper is primarily from an informational perspective, and we begin by examining the following hypothesis.

Integration signaling hypothesis: 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.

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. The above hypothesis claims that the degree in which she has succeeded in this, is signaled by the prosodic/temporal characteristics of her utterance (such as its length, timing, speed, pitch, and intonation).

We test our hypothesis in two steps through analyses of Japanese dialogue data. The first analysis focuses on signaling possibilities of prosodic/temporal features of echoic responses, taken individually, for the degrees of the speakers' information integration. The second analysis then focuses on their signaling potentials in more detail. We will use the measures of accuracy and comprehensiveness to determine (1) exactly what ranges of the speakers' integrations are signaled by the prosodic/temporal features, and (2) exactly what prosodic/temporal features, or what *combinations* of these features, signal those integration ranges.

We then discuss the implications of our findings on the informational potentials of echoic responses with respect to their grounding functions. If we can identify the prosodic/temporal features that signal the high and low degrees of integration, then it seems natural to be able to conclude that echoic responses with those respective features are used to perform the grounding acts of acknowledgment and repair-initiation. Contrary to this simple generalization, we argue that we need to posit a new type of grounding function *display* to capture the entire range of dialogue-coordinating functions of echoic responses.

Analysis I

Methods

Data To examine the validity of our hypothesis, we conducted an analysis on actual occurrences of echoic responses

[1] Also with Nara Institute of Science and Technology.

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 16KHz. They were subsequently divided automatically by power measurements into "Utterance Units (UUs)," consecutive stretches of speech bounded by silence. The start time and end time of each utterance unit were also extracted automatically.

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., "mosimosi"(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 response, an information integration rating, which is a measure for the degree of success, indicated by a signal, with 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 of the repeat instances 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 five 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.

Length: Repeat instances were categorized in terms of their lexical make-up. A *long* repeat is a repeat which contains or paraphrases at least all of a repeated part of a UU and possibly contains additional lexical materials. A *short* repeat is a repeat which repeats or paraphrases a strict subpart of a repeated UU.

Boundary Tone: Repeat instances were categorized in terms of their final intonation patterns. A variant of J-ToBI (Venditti, 1997) 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.

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 (Siegel & Castellan, 1988), which takes into account chance level biases, has been widely accepted by many researchers as one of the most useful measures of such reliability (Carletta, 1996; Hirschberg & Nakatani, 1996); 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[2]. 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.

[2] 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 categorical features of echoic responses.

	[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
Long	3	30	11	22	18	15	26	7
Short	10	28	18	20	30	8	32	6

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

Single 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 two categorical features, boundary tone and length, we applied χ^2 tests 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. Similarly, for length, we found significant differences in the [1]-[2345] and [123]-[45] divisions. The results also indicate that a high boundary tone and a short repeat are 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. For all three features, higher values 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 five 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.

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^{**}
Length	$\chi^2(1)$	3.50^{+}	1.44	4.80^{*}	0.35
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^{+}

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

Analysis II

Motivated by this observation, we will now take a closer look at our data, in order to answer the following questions:

- Exactly what prosodic/temporal features have signaling potentials as to the speaker's integration rate.
- Exactly what ranges of integration rates are signaled by those features.

Methods

Before we start investigating these issues, however, we need specify (1) how we measure the signaling potential of a specific feature as to another feature, and (2) what range of prosodic/temporal features we consider as candidates for such signals.

Measures for Signaling Potentials Suppose you are wondering whether a feature α signals another feature β. One natural way to approach this issue is to see how often β occurs when α occurs. This method measures the *accuracy* of α as a signal to β. If β occurs whenever α occurs, then α is a perfectly accurate cue to β.

However, we cannot determine the signaling potential of α with the measure of accuracy alone. Suppose α seldom occurs when β occurs, but when α does occur, β also occurs. In such a case, α is a perfect cue to β in the accuracy measure. We would not, however, call α a good cue to β, simply because α misses most of the occurrences of β. To be counted as a good cue, α must also be a *comprehensive* signal to β.

We compute the accuracy rate (ACC) and the comprehensiveness rate (COM) of α as a signal to β in the following way [3]:

$$\mathrm{ACC}(\alpha/\beta) = \frac{\text{Number of cases where } \alpha \text{ and } \beta \text{ occur}}{\text{Number of cases where } \alpha \text{ occurs}}$$

$$\mathrm{COM}(\alpha/\beta) = \frac{\text{Number of cases where } \alpha \text{ and } \beta \text{ occur}}{\text{Number of cases where } \beta \text{ occurs}}$$

As we can see from the above formulas, the accuracy of α as a signal to β tends to be higher when (1) α occurs less frequently, and when (2) β occurs more frequently. On the contrary, the comprehensiveness of α tends be high in the opposite case. Accordingly, there is a trade-off between the accuracy and the comprehensiveness of α, and they normalize each other's chance level. Accuracy and comprehensiveness thus make a fairly good measure for signaling potentials, when used conjunctively.

We saw before that a perfectly accurate cue may be useless if it is low in comprehensiveness. The converse is also true: a perfectly comprehensive cue may be useless if it is low in accuracy. Therefore, when we compare the cuing potentials of two different features α_1 and α_2, it would be a mistake to simply compare the sum of α_1's accuracy and comprehensiveness with the sum of α_2's accuracy and comprehensiveness. We should rather compare the minimal value of α_1's accuracy and

[3] The measures of accuracy and comprehensiveness correspond to the measures of precision and recall traditionally used in the field of information retrieval to evaluate the efficiency of a search engine in retrieving information satisfying the given query. We will use our own terms of "accuracy" and "comprehension," however, to avoid confusion on the issue of cuing performance with that of search efficiency.

comprehensiveness with the minimal value of α_2's accuracy and comprehensiveness. Ordering the signaling performances of different features in this way would exclude from the top list those useless cues with an extremely high accuracy but with low comprehensiveness, or with extremely high comprehensiveness but a low accuracy. More precisely, we use the following ordering to compare the performances of two features α_1 and α_2 as signals to a feature β:

$$\alpha_1 \sqsupseteq \alpha_2 =_{df} \\ \min(\mathrm{ACC}(\alpha_1/\beta), \mathrm{COM}(\alpha_1/\beta)) \geq \min(\mathrm{ACC}(\alpha_2/\beta), \mathrm{COM}(\alpha_2/\beta))$$

Range of Candidate Signals According to analysis I, a higher pitch, faster tempo, and longer delay of an echoic response reflect a lower degree in which the speaker has integrated the repeated information, while a lower pitch, slower tempo, and shorter delay reflect a higher degree of the speaker's integration rate. Therefore, we take the prosodic/temporal features in the left column of Table 4 as candidate signals to integration and those in the right column as candidate signals to disintegration. The variable X appearing in Table 4 indicates the threshold value of the relevant continuous feature, ranging over -0.3, -0.2, - 0.1, 0, 0.1, 0.2, and 0.3. These values are normalized in units of standard deviations from the mean.

Table 4: Candidate signals.

Integration	Disintegration
Length = Long, Short	Length = Long, Short
BT = High, Low	BT = High, Low
Tempo $<$ X	Tempo $>$ X
Delay $<$ X	Delay $>$ X
Pitch $<$ X	Pitch $>$ X

It is certainly possible that the prosodic/temporal features in Table 4 individually have certain signaling potentials to the speaker's integration rate. However, it is also possible, or even natural, that two or more of these features *work together* to make accurate and comprehensive cues. Let's think of the case features α and α' are fairly accurate cues to β, but neither covers the cases of β comprehensively enough, namely, neither occurs frequently enough when β occurs. Even under such a circumstance, however, it might well be that α and α' complement each other to cover the cases of β fairly comprehensively, in such a way that when β occurs and α does not occur, α' occurs instead; and when β occurs and α' does not, α occurs instead. This is the case where the *disjunction* of α and α' makes a comprehensive cue to β, even though α and α' are not comprehensive individually. As the dual to this disjunctive complementation, we can also think of the possibility of *conjunctive* complementation, where α and α' work together to enhance the *accuracy* of the signaling.

In order to take these possibilities of disjunctive and conjunctive complementations into account, we decided to include the conjunctive and disjunctive combinations of the features in Table 4 in our list of candidate signals to the speaker's integration rate. To prevent the explosion of the search space, however, we had to limit ourselves to the combinations of up to three different features in Table 4. As a result, the combinations that we considered were of one of the following forms: $\alpha \vee \gamma$, $\alpha \wedge \gamma$, $\alpha \vee (\gamma \wedge \delta)$, $\alpha \wedge (\gamma \vee \delta)$, $\alpha \vee \gamma \vee \delta$, and $\alpha \wedge \gamma \wedge \delta$.

Procedures As we indicated at the outset, we are interested in not only what prosodic/temporal features have signaling potentials as to the speaker's integration rate, but also what ranges of integration rates are cued by those features. There might be a prosodic/temporal feature that signals *strong* information that the speaker's integration rate is exactly 5, but of course, there may be only a cue with *weak* information indicating that the speaker's integration rate is in the range of [2345]. We wish to address this issue of signaled ranges of integration rates explicitly.

Assuming that the signaled ranges are closed in the direction of integration or disintegration, there are eight possible ranges that may be signaled: [1], [12], [123], [1234], [5], [45], [345], and [2345].[4] For each of these ranges, we order our candidate prosodic/temporal features according to their potentials as a cue to the range in question. The candidate features and the method of ordering are as described above. Thus we can compare the overall cuing performances of prosodic/temporal features with respect to different ranges of integration rates, and hence can determine what ranges, if any, are the targets of prosodic/temporal cuing. By studying the features in the top places of the relevant orderings, we can also determine what particular prosodic/temporal features do the cuings.

Results

Signaled Information Table 5 shows results of ordering candidate signals for the following integration ranges: [1], [12], [123], [1234], [5], [45], [345], and [2345]. (Only the top five candidates are shown for each integration range.)

As it turns out, the integration range [1234] is the target of the most accurate and comprehensive signaling from prosodic/temporal features (91.67% and 94.83%). The integration range [2345] receives the second best signaling (91.38% accuracy and 91.38% comprehensiveness). The integration range [123] also seems to be the target of a fairly good signaling (87.76% accuracy and 89.58% comprehensiveness). On the other hand, the accuracy and comprehensiveness of the top candidate signals for the integration ranges of [1], [12], [345], and [45] are significantly lower; they are, respectively, (61.54%/61.54%), (68.97%/68.97%), (66.67%/76.92%), (77.27%/73.91%), and (78.57%/78.57%).

These discrepancies in cuing performances are visualized in Figures 1 (a) and (b), where the best twenty features for each range are plotted according to their accuracy and comprehensiveness. Figure 1 (a) shows a clear discrepancy in cuing performances between the group of features targeted at [2345] and the group targeted at [5], [45], or [345]. Figure 1 (b) shows a discrepancy between the group targeted at [123] or [1234] and the group targeted at [1] or [12].

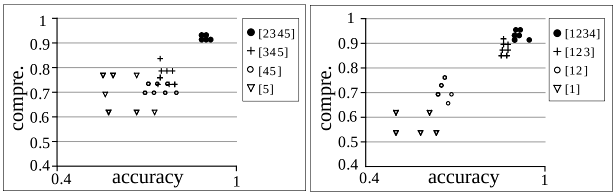
Consequently, we must conclude that the integration ranges signaled by the prosodic/temporal features of echoic responses are [1234], [123], and [2345], but *not* [1], [12], [5], and [45]. Intuitively, the information that one receives is either that the speaker has not completely integrated the information she repeats (the case of [1234] signaled), or that she has not completely failed to integrate it (the case of [2345] signaled), or that she has not integrated it well (the case of [123] signaled). In either case, the information made available through this signaling is rather weak.

Signaling Features What exactly, then, are the temporal/prosodic features that signal these pieces of information?

[4] Of course, we might well lift this assumption, and consider "intermediate ranges" such as [234], [34], and [3]. This amounts to sorting our integration scale into three or more categories, rather than focusing on the binary classification of integration and disintegration. In this paper, however, we adopt that simplifying assumption, because it allows a more explicit test on the ability of the prosodic features of echoic responses for the binary classification in question. In fact, the test reveals an important fact about the coordinating functions of echoic responses, as we will see shortly.

Table 5: Results of ordering candidate signals (LN: length, DL: delay, TP: tempo).

	Integration	acc.	comp.		Disntegration	acc.	comp.
[5]	[BT = L%] ∧ [DL < 0.0]	77.3%	73.9%	[1]	[LN = S] ∧ [TP > 0.2] ∧ [PT > -0.3]	61.5%	61.5%
	[BT = L%] ∧ [PT < 0.1] ∧ [DL < 0.0]	73.9%	73.9%		[LN = S] ∧ [TP > 0.2] ∧ [PT > -0.2]	61.5%	61.5%
	[BT = L%] ∧ [PT < 0.2] ∧ [DL < 0.0]	77.3%	73.9%		[LN = S] ∧ [TP > 0.2] ∧ [PT > -0.1]	61.5%	61.5%
	[BT = L%] ∧ [PT < 0.3] ∧ [DL < 0.0]	73.9%	73.9%		[LN = S] ∧ [TP > 0.2] ∧ [PT > 0.0]	63.6%	53.9%
	[DL < 0.0] ∧ ([BT = L%] ∨ [PT < -0.3])	70.8%	73.9%		[LN = S] ∧ [TP > 0.2] ∧ [PT > 0.1]	63.6%	53.9%
[45]	[TP < 0.2] ∧ [PT < 0.2] ∧ [DL < 0.2]	66.7%	76.9%	[12]	[TP > -0.3] ∧ ([BT = H%] ∨ [DL > 0.2])	69.0%	69.0%
	[TP < 0.2] ∧ [PT < 0.2] ∧ [DL < 0.3]	72.7%	61.5%		[BT = H%] ∨ ([TP > -0.3] ∧ [DL > 0.2])	66.7%	75.9%
	[TP < 0.2] ∧ [PT < 0.3] ∧ [DL < 0.2]	66.7%	61.5%		[BT = H%] ∨ ([TP > -0.2] ∧ [DL > 0.2])	66.7%	75.9%
	[TP < 0.2] ∧ [PT < 0.3] ∧ [DL < 0.3]	66.7%	61.5%		[BT = H%] ∨ ([TP > -0.1] ∧ [DL > 0.2])	66.7%	75.9%
	[TP < 0.3] ∧ [PT < 0.2] ∧ [DL < 0.2]	58.8%	76.9%		[TP > -0.3] ∧ ([PT > 0.2] ∨ [DL > 0.2])	65.6%	72.4%
[345]	[TP < -0.3] ∨ ([BT = L%] ∧ [DL < 0.2])	78.6%	78.6%	[123]	[TP > 0.2] ∨ [PT > 0.2] ∨ [DL > 0.2]	87.8%	89.6%
	[TP < -0.2] ∨ ([BT = L%] ∧ [DL < 0.2])	76.7%	78.6%		[TP > 0.2] ∨ [PT > 0.3] ∨ [DL > 0.2]	87.8%	89.6%
	[TP < -0.3] ∨ ([BT = L%] ∧ [DL < 0.3])	75.0%	78.6%		[TP > 0.2] ∨ [PT > 0.2] ∨ [DL > 0.3]	87.5%	87.5%
	[TP < -0.1] ∨ ([BT = L%] ∧ [DL < 0.2])	74.5%	83.3%		[TP > 0.2] ∨ [PT > 0.3] ∨ [DL > 0.3]	87.5%	87.5%
	[TP < -0.1] ∨ ([PT < 0.2] ∧ [DL < 0.2])	74.4%	76.2%		[TP > 0.1] ∨ [PT > 0.2] ∨ [DL > 0.2]	86.3%	91.7%
[2345]	[LN = L] ∨ [TP < 0.2] ∨ [PT < -0.3]	91.4%	91.4%	[1234]	[BT = H%] ∨ [PT > 0.1] ∨ [DL > 0.0]	91.7%	94.8%
	[LN = L] ∨ [TP < 0.2] ∨ [PT < -0.2]	91.4%	91.4%		[BT = H%] ∨ [PT > 0.2] ∨ [DL > 0.0]	91.5%	93.1%
	[LN = L] ∨ [TP < 0.2] ∨ [PT < -0.1]	91.4%	91.4%		[BT = H%] ∨ [PT > 0.3] ∨ [DL > 0.0]	91.5%	93.1%
	[LN = L] ∨ [TP < 0.2] ∨ [PT < 0.0]	90.0%	93.1%		[BT = H%] ∨ [DL > 0.0]	94.6%	91.4%
	[LN = L] ∨ [TP < 0.2] ∨ [PT < 0.1]	90.0%	93.1%		[BT = H%] ∨ [PT > 0.0] ∨ [DL > 0.0]	90.2%	94.8%



(a) (b)

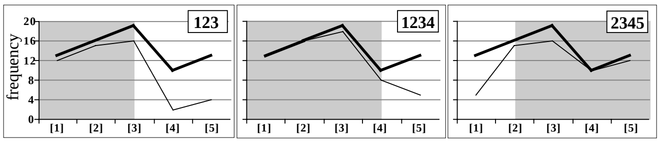
Figure 1: Distributions of signaling performances.

For the integration range [123], one of the best signals is clearly the disjunction of a higher tempo, a higher pitch, and a longer delay, because the top eight features targeted at this range are all of this form. Our data, however, does not clearly differentiate the cuing performances of these features, and so we cannot determine the exact threshold values for tempo, pitch, and delay that define the best signal of this form.

Our data suggest that the integration range [1234] is signaled by the disjunction of a high boundary tone, a longer delay, and a higher pitch. However, the top six features targeted at this range also contain a simpler disjunction consisting of a high boundary tone and a longer delay. Since the accuracy and comprehensiveness of the feature are not discernible from those of the other five, a higher pitch might not really play a role in signaling [1234]. In either case, our data do not allow us to determine the exact threshold for pitch that defines the best signal of this form.

As for the integration range [2345], the disjunction of longness of a repeat, a slower tempo, and a lower pitch have a strong signaling potentials (the top fifteen candidates are all of this form), although, again, our data do not allow us to determine the exact thresholds for the tempo and delay for the best signal of this form.

Figures 2 (a)–(c) show, for the integration ranges [123], [1234], and [2345], the distributions of the best signaling features over the integration scale 1–5. The bold curve indicates the distribution of all echoic responses over the scale, while the non-bold curve indicates the distribution of the signal.



(a) (b) (c)

Figure 2: Distributions of the best signals over the integration scale.

Accordingly, a prosodic/temporal feature is a good cue for the relevant integration range if its curve follows the bold curve in the shaded area as closely as possible and goes down as low as possible in the non-shaded area. Note that the non-bold curve in each graph does a fairly good job in this respect, endorsing our observations.

To summarize the observations made in this section:

- The integration range [123] is disjunctively signaled by a faster tempo, a higher pitch, and a longer delay.
- The range [1234] is disjunctively signaled by a high boundary tone, a longer delay, and possibly a higher pitch.
- The range [2345] is disjunctively signaled by a long repeat, a slower tempo, and a lower pitch.
- The ranges [1], [12], [5], [45], and [345] receive no significant prosodic/temporal signals.

Discussions

So far in this paper we have concentrated on the *informational functions* of echoic responses, with respect to the degree in which the responder has integrated the repeated information into her body of knowledge. In this section, we discuss what our findings on the informational potentials of echoic responses imply about their dialogue-coordinating functions, or more specifically, about the varieties of grounding acts that might be performed with them.

For brevity, let us denote the prosodic/temporal features that signal the integration ranges [123], [1234], and [2345] by S_{123}, S_{1234}, and S_{2345}, respectively. Now, if we assume that echoic responses may be used to perform the acts of *acknowledgment* and *repair-initiation*, it seems natural to suppose that, generally, an echoic response cuing high integration for the responder performs the act of acknowledgment and that cuing low integration performs the act of repair-initiation. Therefore, on the basis of the observations 2 (a)–2(c) above, one might be tempted to infer that echoic responses with feature S_{123} or with feature S_{1234} are used to perform repair-initiation, while those with feature S_{2345} are used to perform acknowledgment. The story is not that simple, however.

A serious problem with this story arises from the fact that according to our observations, the alleged integration cue *overlaps* with each of the alleged disintegration cues *in what they signal*. That is, the cue S_{123} and the cue S_{2345} both contain the integration rates 2 and 3 in their targets, and the cue S_{1234} and the cue S_{2345} both contain the integration rates 2, 3, and 4 in their targets. Therefore, no matter how we may interpret the integration rates 1–5 and divide them into those representing integration and those representing disintegration, there will always be a large number of echoic responses for which these cues will fail to classify as either integration cases or disintegration cases.[5]

Now if we are right about this observation and the information about the integration/disintegration status is absent in many instances of echoic responses, then what sort of grounding act, if any, is performed with these instances? Note that in these instances, information concerning whether the responders have integrated or failed to integrate the repeated information is *absent*. Therefore, acknowledgment and repair-initiation cannot be the acts performed by them, assuming that the performances of the acts require the availability of positive information about the responders' integration and disintegration. Saying that echoic responses do not perform any grounding acts is hardly a satisfying answer, since clearly, the presence of the echoic responses make *some* difference in the process of putting information into the common ground.

To capture this sort of grounding act, we propose the notion of *display*. When performing the act of display, we reproduce the information that we believe to have been received in the directly preceding turn by the other party. The point of the reproduction is to expose the other party to the information again and, more importantly, let her make corrections *if* the reproduced information is not what she had intended, and then to let her continue on to the next turn *if* the reproduced information is correct.

The act of display is fundamentally different from the acts of acknowledgment and repair-initiation in that while the latter positively *guides* the other party to perform a certain act (repair or the initiation of the next turn), the former lets the partner *choose* an appropriate action depending on whether the reproduced information is correct. Therefore, for the purpose of displaying, the responder does not have to positively indicate whether she has integrated the reproduced information, while for the purpose of acknowledgment or repair initiation, the speaker must indicate it to guide the partner into a specific act. In fact, an echoic response would not constitute the act of display (and become acknowledgment or repair-initiation) if either integration or disintegration on the responder's part were indicated strongly. Consequently, the act of display is not only possible in the absence of such information, it is ideally done in its absence. This explains the existence of a large number of echoic responses that are signaled neither as integration cases nor disintegration cases.

What then is the overall picture of the grounding acts that are performed with echoic responses? Precisely speaking, the picture varies depending on how we decide to interpret the scale 1–5. If we were to consider [123] as the disintegration range and [45] as the integration range, then we would assign the act of repair-initiation to those echoic responses with the prosodic/temporal feature S_{123} while assigning the act of display to others; similarly, if we were to adopt the division [1234]-[5], we would assign the act of repair-initiation to those echoic responses with the feature S_{1234} while assigning the act of display to others. On the other hand, the division [1]-[2345] would let us assign the act of acknowledgment to those echoic responses with the feature S_{2345} while assigning the act of display to others. Finally, if we were to adopt the division [12]-[345], we would have to assign the act of display to all echoic responses.

We will not commit ourselves to either of these specific cases in this particular paper. Although the last one may appear to be too extreme a position to take, our discussions so far have given nothing that precludes the possibility. However, our empirical investigations into the informational potential of echoic responses does suggest, as a logical consequence, that the notion of display must play a significant, even central, role in explaining the varieties of grounding acts that may be performed with echoic responses.

Conclusion

On the basis of statistical analyses on the signaling accuracy and comprehensiveness, we examined the informational potential of echoic responses found in corpora of the Japanese spoken language. We found that variations in the speed, pitch, boundary tone, length, and timing of an echoic response have a definite potential in signaling the responder's informational state, concerning how successfully she has integrated the repeated information into her existing body of knowledge.

On the other hand, we did not find any pair of temporal/prosodic features that divide the scale of the speaker's integration rates at a unique point, and had to abandon the simple picture that partitions the instances of echoic responses into those performing of acknowledgment and those performing repair-initiation. This finding (or non-finding) required us to postulate a grounding act that a speaker may perform without signaling positive information about her integration rate, and made us argue that the notion of display defines such a grounding act and explains the phenomenon in a natural way.

References

Beun, R. J. (1995). The function of repetitions in information dialogues (Tech. Rep. 20). IPO annual progress report.

Carletta, J. (1996). Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, *22*, 249–254.

Clark, H. H., & Schaefer, E. F. (1989). Contributing to discourse. *Cognitive Science*, *13*, 259–294.

Grosz, B., & Hirschberg, J. (1992). Some intonational characteristics of discourse structure. *Proceedings of ICSLP* (pp. 429–432).

Gumperz, J. (1991). Contextualization and understanding. In A. Daranti & C. Goodwin (Eds.), *Rethinking context*. Cambridge: Cambridge University Press.

Hirschberg, J., & Nakatani, C. H. (1996). A prosodic analysis of discourse segments in direction-giving monologues. *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics* (pp. 286–293).

Koiso, H., Shimojima, A., & Katagiri, Y. (1996). Informational potentials of Dynamic Speech Rate in Dialogue. *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society* (pp. 394–399).

Norrick, N. R. (1994). Repetition as a conversational joking strategy. In B. Johnstone (Ed.), *Repetition in discourse interdisciplinaryperspectives* Vol. 2. Ablex Publishing Corporation.

Siegel, S., & Castellan, N. J. J. (1988). *Nonparametric statistics for the behavioral sciences* (2nd edition). New York: McGraw Hill Text.

Tannen, D. (1994). *Talking voices: repetition, dialogue, and imagery in conversational discourse*. Cambridge: Cambridge University Press.

[5] Interpret, for example, [123] as representing disintegration and [45] as representing integration. Then, although those echoic responses with the feature S_{123} are classified as disintegration cases, the other instances are classified neither as integration nor disintegration. For even when some of them have the feature S_{2345}, the feature simply classifies them as being in the range of [2345], which, on the current interpretation, contains both the rates 2 and 3 representing disintegration and the rates 4 and 5 representing integration. The cue S_{1234} does not work either, because its target contains [4] as well as [123]. The other interpretations of the scale 1–5 will encounter analogous problems.

Traum, D. R. (1994). A computational theory of grounding in natural language conversation (Tech. Rep. 545). University of Rochester.

Venditti, J. J. (1997). Japanese ToBI labeling guidelines (Working papers in linguistics 50). Ohio State University.

Walker, M. A. (1992). Redundancy in collaborative dialogue. *Fourteenth International Conference on Computational Linguistics*, pp. 345–351.