THE EFFECT OF REPETITION ON WORDS IN RECORDED DICTATIONS

E. G. Bard, A. J. Lowe, G. T. M. Altmann

Centre for Speech Technology Research, University of Edinburgh

ABSTRACT

Two experiments on words isolated from recorded dictations show that the effect of repetition on intelligibility depends on the discourse roles of the tokens being compared. When the later token adds no new information to the discourse, as in self-corrections and other coreferential repetitions, intelligibility falls with repetition. When the later token introduces a new discourse entity, intelligibility may rise with repetition. The results are discussed with reference to the effects of the speaker's experience in machine dictation and to the consequences for automatic speech recognition.

INTRODUCTION

Fowler and Housum [1] have recently shown that in radio monologues and television interviews second tokens of repeated words are shorter and less intelligible than the corresponding first tokens. If these findings apply to the speech produced for automatic speech recognizers, two undesirable consequences may follow. First, whenever the input is long enough to allow words to be repeated, the task of recognition will become more difficult. Second, whenever, even in short input sequences, a user repeats some part of the input to correct an error, the system will be presented with a more, not a less, difficult task.

The likelihood of these consequences depends on how generalizable the repetition effects are. First, speech produced for the sorts of tasks which speech recognizers may perform and speech produced for media broadcasts may not be equally susceptible to this effect. Second, repetitions in self-corrections may be more intelligible than what they correct, while other repetitions are less intelligible. Third, the Fowler and Housum result may apply to any second mention, or only to those which are coreferential with their corresponding first mention. In other words, the repetition effect may be merely a matter of the repeated use of a word or it may depend on the role each token plays in the developing discourse. Finally, it may be the case that loss of intelligibility with repetition is a counterproductive tendency which speakers learn to correct as they gain experience interacting with a speech system. The means which must be taken to protect an automatic speech recognizer against the effect will depend on which of these is the case.

This paper examines the effects of repetition on the intelligibility of words in speech dictated to a recording device for later transcription as a letter, report, or memo. Speakers were either experienced users of dictating equipment or novices. Unlike the speakers in Fowler and Housum's experiment who were either delivering monologues or responding to questions, these speakers were composing documents as part of their normal work. For both of the experiments reported here, panels of listeners attempted to identify word tokens which had been instrumentally isolated from their original dictated speech contexts and subsequently overlaid with noise. The results of two experiments indicate that repetition does affect word intelligibility in dictation and that the nature of the effect depends on the speaker's experience at the task and the word token's role in discourse.

EXPERIMENT 1

This experiment was designed to determine whether repetition leads to loss of intelligibility in dictation and, if so, whether words repeated in the course of a self-correction are less susceptible to such loss than words repeated in the course of fluent dictation.

METHOD

Materials were taken from a digitally recorded corpus of dictations. All recordings were made using the same recording equipment in office conditions. Sixty-one speakers are represented in the corpus, each contributing three or four dictations, for which the hard copy document produced in the speaker's office was also available. From each recorded dictation a verbatim orthographic transcription was prepared.

Thirty-two quadruples of word tokens were selected, sampling unequally from the speech of 14 speakers, with each quadruple composed of tokens of a single word form.
Eighteen of the quadruples came from experienced dictation-givers, 14 from inexperienced. Within any one quadruple, two test words were earlier and later Disfluent tokens; the second was produced in a sentence or phrase which immediately followed and revised the sentence or phrase in which the first had been spoken. In no case did the final document include both members of such pairs. The two other tokens in each quadruple were earlier and later Fluent tokens of the same word. Both members of the Fluent pair appeared in the final document. Wherever possible the earlier Fluent token was the first occurrence of the word in the dictation and the later was the second.

All 128 word tokens were isolated from their dictation contexts, and their onsets and offsets determined with reference to the recordings and corresponding waveforms. The isolated tokens were redigitized at a signal to noise ratio of +2dB, which had yielded an average recognition rate of about 50% in pre-tests.

Four audio tapes were produced, each containing one token from each of the 32 quadruples, distributed by Latin Square to represent each combination of Fluency (fluent, disfluent) and Token (earlier, later) equally. Test words were blocked by speaker and preceded by an "orienting" sentence from the same speaker, also overlaid with noise at +2dB. Where a speaker had contributed only one quadruple, another word token from the same dictation was supplied so that no speaker block contained less than two stimuli. With these dummy items, each tape presented 40 words. Two tones preceded the orienting sentences and one tone preceded each test word. The ISI was 4 seconds.

Subjects and procedure. Subjects were 32 members of the University of Edinburgh undergraduate and postgraduate community all of whom would have encountered the accents represented in the experiment. Eight Subjects heard each tape.

Subjects listened to the tapes over headphones and were instructed to write down test words. One orienting sentence and two dummy test words were given as practice items.

RESULTS

Responses which corresponded to target words and their homophones, with or without any sub-syllabic inflections in the original, were scored as correct and were used in the analysis presented below.

ANOVAs were performed in which Subjects or quadruples of words were cases. Subjects were nested in groups and crossed with Fluency and Token. Word quadruples were divided into 4 groups on the basis of the tape containing their first Fluent token and each quadruple was crossed with Fluency and Token.

**TABLE 1: Percent Correct Recognitions for Fluent and Disfluent Pairs, Experiment 1**

<table>
<thead>
<tr>
<th>Category</th>
<th>Token</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Earlier</td>
</tr>
<tr>
<td>Fluent</td>
<td>50</td>
</tr>
<tr>
<td>Disfluent</td>
<td>57</td>
</tr>
</tbody>
</table>

Table 1 shows that repetition does reduce intelligibility in dictated speech but not for all classes of repetition. While there was no overall effect of repetition on intelligibility ($F_1 < 1.00; F_2 < 1$) there was a significant by-Subjects interaction between Fluency and Token ($F_1 = 13.26, df = 1, 28; p = .001; F_2 = 2.69, df = 1, 28, p = .11$). Words lost intelligibility with repetition only if the repetition was Disfluent (Newman-Keuls test at $p = .05$). They actually gained intelligibility if the repetition was Fluent (Newman-Keuls test at $p < .05$).

DISCUSSION

Dictated speech is susceptible to loss of intelligibility with repetition, but the effect is not a unitary phenomenon. Unfortunately, it seems to be restricted to the situation where it ought to matter most in ASR: speakers produce less intelligible word tokens when they repeat themselves in the course of making a correction. Moreover, as this apparently counterproductive behaviour came from a sample where the speech of experienced dictation-givers predominated, it is unlikely to be the result of the speakers' initial disorientation. In contrast, fluent self-repetition appeared to increase intelligibility.

To understand why these two kinds of repetition should have such different results, consider the roles which each kind of later token might play. When a word is repeated in the course of a self-correction, the second token adds no information to the message. It almost always denotes the same object, event or relationship just established by the earlier token. Since such tokens usually appear within a repetition of several earlier words, they tend to be produced in local contexts very similar to those in which they have just been said. In the sense in which Lieberman [2] used the term, Disfluent repetitions are
redundant or predictable from their linguistic contexts. Lieberman and others [3, 4] found that the predictability of a word from its linguistic context correlated negatively with the intelligibility of the word token as measured in isolation. The result for Disfluent pairs appears to be a replication of Lieberman's effect.

It would follow that the later members of the Fluent pairs, which show higher intelligibility, should be less predictable from context. This would certainly be the case if the later tokens added new information to the discourse which was not predictable from what had been said earlier. In fact, some 60% of the later Fluent tokens introduced new information: they were not coreferential to their earlier mate. The preponderance of such items may have produced the unexpected reversal of the repetition effect. Experiment 2 pursues this possibility.

EXPERIMENT 2

This experiment examined the effects on intelligibility of a word's discourse role. Some second tokens of words were coreferential with first tokens, and so denoted only Given information. Others second tokens introduced an object, event, or relationship which was New to the dictation. If the increase in intelligibility for repeated Fluent words in Experiment 1 was due to the preponderance of New second tokens, then only the New tokens should show an increase intelligibility with repetition here. Pairs with Given second tokens should, like the Disfluent pairs in Experiment 1, show a decrease.

The speaker's level of experience with the task is an explicit factor here in order to determine whether speakers discard potentially unhelpful speech habits as they accumulate experience.

METHOD

Materials. Materials were taken from the corpus described earlier and consisted of pairs of tokens of a word. The first token of each pair introduced a new discourse entity according to the definition set out above. Each first token was the first use of its word form with any meaning in its dictation. Second tokens were of two sorts. New second tokens introduced some new entity, while Given second tokens were coreferential with their corresponding first tokens. Pairs were thus New-Given or New-New.

Twenty-eight New-Given pairs were selected from experienced speakers' dictations. A further set of 28 New-New pairs, loosely matched to the first set for frequency and for length in syllables of the citation form of the word, were then selected from the same materials. Fifty-six word pairs were selected from the dictations of the inexperienced speakers in the same way. The 112 pairs of test words were isolated from their dictation contexts and redigitized at the (pre-tested) signal to noise ratio of +4dB.

Two audiotapes were produced, each including either the first or the second token of each pair, and each balanced for Speaker Experience, Token, and Given/New status and roughly balanced for word frequency and length. The remaining details of materials and procedure were as described for Experiment 1.

Subjects. Twenty-eight members of the University of Edinburgh undergraduate and post-graduate community participated.

RESULTS

Response criteria were those used in Experiment 1. Results from one Subject, whose success rate was more than two s.d below the mean for all Subjects, were discarded. One word pair was discovered to have been misclassified. The corrected data were for 27 Subjects, 55 New-Given pairs, and 57 New-New pairs.

<table>
<thead>
<tr>
<th>Category</th>
<th>Speaker</th>
<th>Token</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>New-Given</td>
<td>Experienced</td>
<td>76</td>
<td>71</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Novice</td>
<td>69</td>
<td>61</td>
<td></td>
</tr>
<tr>
<td></td>
<td>All</td>
<td>72</td>
<td>66</td>
<td></td>
</tr>
<tr>
<td>New-New</td>
<td>Experienced</td>
<td>66</td>
<td>72</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Novice</td>
<td>56</td>
<td>53</td>
<td></td>
</tr>
<tr>
<td></td>
<td>All</td>
<td>61</td>
<td>63</td>
<td></td>
</tr>
</tbody>
</table>

Using the materials produced by all speakers, separate ANOVAs were run using Subjects and word-pairs as cases. Subjects were nested in groups and crossed with Speaker Experience (novice, experienced), Discourse Status (New-Given, New-New) and Token (first, second). Materials were crossed with Token and nested in groups determined by the tape containing their first token, in Discourse Status, and in Speaker Experience. Two more pairs of ANOVAs considered words of experienced and novice speakers separately. No effects were significant by
 materials, though robust effects were found by Subjects.

Two questions were addressed in this experiment. First, does the effect of repetition in fluent speech depend on the discourse role of the word tokens involved? Second, does experience with the dictation task tend to protect a speaker against any of the loss of intelligibility associated with repetition?

In the means for all speakers considered together, Table 2 suggests the answer to our first question: New and Given second tokens do behave differently. In New-Given pairs, the second token was, as predicted, less intelligible than the first (Discourse Status x Token: \(F_1 = 16.05, df = 1, 25, p = .0005;\) Newman-Keuls, \(p < .05\)). Contrary to prediction, however, there was no effect of repetition on New-New pairs.

This overall null result on New-New pairs, however, hides qualitatively different results which answer our second question. In the means subdivided for Speaker Experience, Table 2 shows that New-New pairs from novice dictation-givers lose intelligibility with repetition, as do their New-Given pairs (Token: \(F_1 = 8.19, df = 1, 25, p = .0084;\) Discourse Status x Token: n.s.). The New-New pairs from more experienced speakers, on the other hand, gain intelligibility (Discourse Status x Token: \(F_1 = 12.05, df = 1, 25, p = .0019;\) Newman-Keuls at \(p < .05\)) while the fall in intelligibility for New-Given pairs does not reach significance. Thus, experience seems to make the effort of repetition depend on the discourse role of the repeated word.

**DISCUSSION**

Repetition does affect the intelligibility of words in dictation, but the effect is not uniform. Only for speakers new to the dictating task did repetition always seem to involve loss of intelligibility. For more experienced speakers, the intelligibility of a word token depended on what information it brought to the discourse. Tokens referring to an already-established entity lost intelligibility. Disfluent repetitions appear to belong to this category, since they make immediate reference to whatever their first tokens represented. Tokens referring to some new item, on the other hand, gained intelligibility. The preponderance of such New-New pairs among the Fluent repetitions in Experiment 1 and the preponderance of experienced dictation-givers among the speakers will help to explain why Fluent repetitions increased in intelligibility.

This pattern should come as no surprise. Phenomena subsumed under the terms phonological focus and sentence accent [5] mark important new items in speech and the acoustic realizations of these processes are likely to make word tokens more intelligible than they might otherwise be. If novice dictation-givers are not following this normal pattern, it may be because they do not yet understand the communicative demands of the task or appreciate the possibility that what they say may go unrecognized.

For ASR systems it will not be straightforward to deal with the effects of repetition on intelligibility. The user's self-corrections may be harder to recognize than the token which they are intended to replace. Experience with speaking to a machine seems to create a more, rather than a less complex pattern of behaviour. Words which are difficult to identify cannot merely be referred preferentially to a list of words so far successfully identified. Whether the clearer token has been said or is yet to be said depends on the word's relationship to the rest of the discourse and that is unlikely to be determined without good analysis at higher levels of interpretation.

**REFERENCES**


**ACKNOWLEDGEMENT**

This work was supported by the Information Engineering Directorate/Science and Engineering Research Council as part of the IED/SERC Large Scale Integrated Speech Technology Demonstrator Project (SERC grants D/29604, D/29611, D/29628, F/10309, F/10316) in collaboration with Marconi Speech and Information Systems and Loughborough University of Technology.