REPETITION AND ITS PHONETIC REALIZATIONS: INVESTIGATING A SWEDISH DATABASE OF SPONTANEOUS COMPUTER-DIRECTED SPEECH

Linda Bell and Joakim Gustafson
Centre for Speech Technology (CTT)
Department of Speech, Music and Hearing, KTH, Stockholm

ABSTRACT

This paper is an investigation of repetitive utterances in a Swedish database of spontaneous computer-directed speech. A spoken dialogue system was installed in a public location in downtown Stockholm and spontaneous human-computer interactions with adults and children were recorded [1]. Several acoustic and prosodic features such as duration, shifting of focus and hyperarticulation were examined to see whether repetitions could be distinguished from what the users first said to the system. The present study indicates that adults and children use partly different strategies as they attempt to resolve errors by means of repetition. As repetition occurs, duration is increased and words are often hyperarticulated or contrastively focused. These results could have implications for the development of future spoken dialogue systems with robust error handling.

1. INTRODUCTION

Repetition in spoken language has recently been discussed from a number of different points of view. Why do people repeat themselves and how does repetition affect their manner of speaking? Aitchison [2] has suggested that repetition is a central phenomenon in the study of language: "In one sense, the whole of linguistics can be regarded as the study of repetition, in that language depends on repeated patterns" (p.16). In a recent paper, Swerts et al [3] discuss the numerous possible functions of repetition in the context of human-human dialogues. The study reported in this paper, however, deals exclusively with human-computer interaction and the realization of repetition in that context. It is often the case that repetition to a spoken dialogue system occurs when the users fail to make themselves understood. Repetition, then, is one of the strategies available to speakers who wish to resolve errors in human-computer interaction. We will here assume that the main function of repetitions in the database is to resolve such errors.

Studies by Oviatt et al [4] and Levow [5] have shown that speech during error resolution tends to be clearer, contain fewer disfluencies and that the total utterance duration is increased significantly. In a study of multimodal human-computer interaction, Oviatt and VanGent [6] argue that users distinguish repetition from the original input by means of linguistic contrasts and switching modalities. Modality switching was not an option in the current study since the users of our spoken dialogue system had no other means of communicating except by using their voice. Any contrast between a repetition and what the user first said to the system, the original input, would have to be indicated by the manner of speaking only. This paper compares repetitions and original utterances by measuring several acoustic and prosodic parameters.

2. METHOD

2.1. Material

A spoken dialogue system with an animated agent was set up in a public location and recordings of spontaneous human-computer interactions took place for a period of six months [1]. The material analyzed in this paper is extracted from a database consisting of 4647 spontaneous utterances spoken by 1380 users. These utterances were all recorded during the first three months of this period. The utterances were transcribed orthographically and some basic speaker characteristics were manually labeled. This made it possible to distinguish adults and children among the users of the system. In the database, repetitions of all kinds make up approximately 10% of all utterances. In order to make the acoustic and phonetic comparisons more accurate, the present study is based on sequences of original input and repetitions that were lexically identical. This applied to half of all the repetitive utterances in the database. Thus, 452 utterances (200 originals and 252 repetitions) were manually extracted from the database. 339 of the utterances are spoken by adults, while the remaining 113 utterances are spoken by children. Results from studies of the remaining utterances will be presented in a forthcoming paper [7]. Although single pairs of original input/repetition are by far the most common pattern in the study, Figure 1 shows that a single utterance was repeated up to five times in a row.

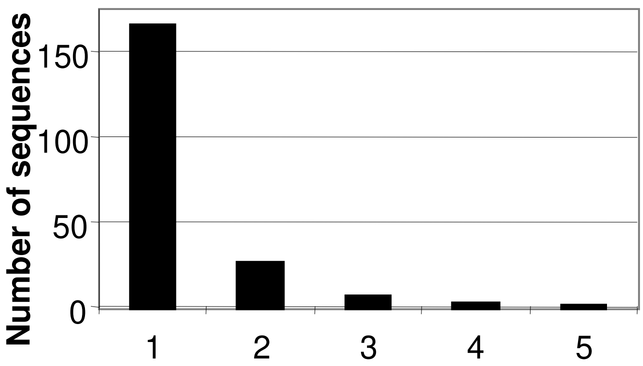

Figure 1. Number of repetitive sequences in the database.

The 452 utterances were closely analyzed with the purpose of examining if and how the users changed their manner of speaking as they repeated something they had already said. Furthermore, we wanted to see whether there were any significant differences between the adults and children who were using the spoken dialogue system.

2.2. Selecting features

As this study was initiated, it was our hypothesis that people adapt their manner of speaking to distinguish instances of repetition from original input to a spoken dialogue system. We therefore sought a way of determining whether these hypothesized features were perceivable. A group of 36 students were asked to listen to twenty pairs of original input/repetition, all of which were extracted and played randomly. The students were then asked to estimate which of the two utterances was the original input to the system and which was the repetition. 16 out of 20 utterances were correctly judged by 82% of the subjects, which indicates that it is usually possible to distinguish a repetition from its original input. In the pairs of original input/repetition that were correctly judged, utterance duration was increased in all cases. The remaining four utterances were incorrectly judged by an equally large group, 80% of the subjects. These incorrectly judged utterances were different from the correctly judged ones in that the repetitions were shorter than the original input. It thus appeared the subjects responded to duration as an important cue in distinguishing a repetition from its original input. Utterance duration and speech rate seemed to be relevant features.

Seven pairs of utterances in the above mentioned test were correctly judged by more than 90% of the subjects. A detailed analysis of these utterances indicated that other features, apart from duration, might be equally important for distinguishing repetitions from original input. These features included a movement towards clearer articulation, increased loudness, inserted pauses and focus shifting. Figure 2 below shows a typical exchange in which an adult male user repeats his original input twice. In the first repetition, loudness is increased. The second time the utterance is repeated it is hyperarticulated and contains inserted pauses between the words. As can be seen in Figure 2, the utterance duration is increased in the first as well as the second repetition.

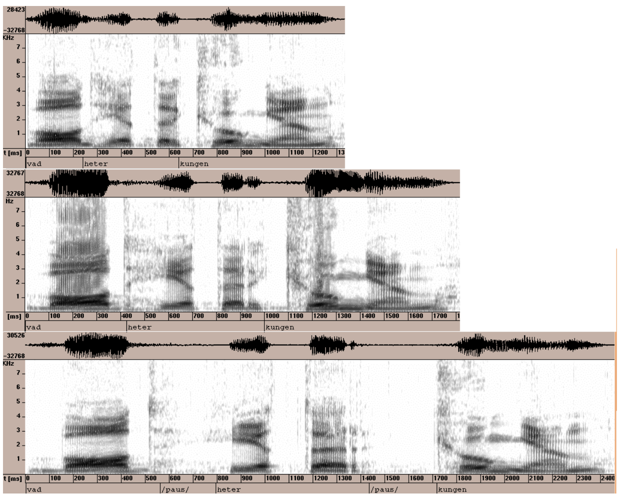

Figure 2. The utterance “Vad heter kungen?” (“What is the name of the king?”) as original input (top) and repeated twice by the same user

2.3. Coding and labeling of data

Because of the noisy, public environment in which the spoken dialogue system was set up, it was necessary to install a push-to-talk mechanism [1]. As a result, short silences at the beginning and end of the sound files were frequent. These silences were removed from the sound files after which they were automatically measured. This enabled a comparison between the duration of the original spoken input and its corresponding repetition. Moreover, the number of syllables per second was measured and inserted pauses were marked. In order to examine the variation in articulation in the spoken input to the system, all original utterances as well as repetitions were labeled with respect to their respective degree of articulation. Articulation was labeled as either reduced, normal or hyperarticulated. These labels were subjectively assessed by the present authors. The following two features were assessed in the same way: perceived loudness (high, normal, low) and shifting of focus (yes/no).

3. RESULTS

3.1. Duration, inserted pauses and speech rate

On average, the original utterances were 1361 msec, while the repetitions were 1565 msec, or 15% longer. The adults' utterances were on average 18% longer for repetition while the duration of the children's repetitions increased by 7%. These figures are comparable to those reported in recent studies on error resolution [4, 5]. Even though a majority of users spoke slower as repetition occurred, this was not always the case. Some speakers did not increase the duration of their utterance, and some even spoke faster. This means that the average numbers above are misleading in the sense that they include both increases and decreases in utterance duration. A more accurate picture shows that the average lengthening of duration in the repetitions is over 40% and the corresponding shortening is 15% on average. Figure 3 below illustrates this.

In more than half of all cases, the second repetition was even longer than the first repetition of the same utterance. The second repetition was shorter than the first one in only 14% of all instances. The second repetitions were also distinguished by the fact than they frequently contained inserted pauses between words. Such pauses were found in 29% of the second repetitions, compared to 7,5% in the first repetition and 2,5% in the original utterances.

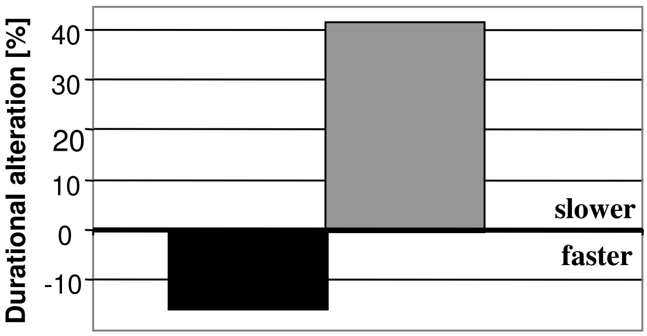

Figure 3. The average durational changes in repetitions when they are compared to the original utterances.

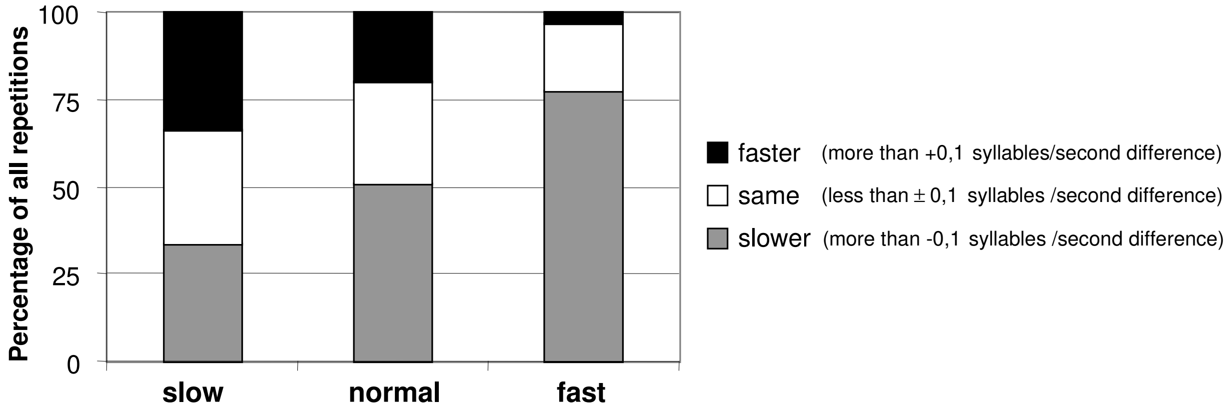

Figure 4. The change in speech rate in the repetitions when they are compared to the original utterances. The utterances are grouped according to syllables per second, where slow is up to three, normal is three to five and fast is more than five syllables per second.

The speech rate of the utterances was computed in terms of number of syllables per second. As can be seen in Figure 4, those users whose original input to the system was normal or fast tended to speak slower during repetition. The users who spoke slowly in the original utterance, on the other hand, may already have adapted themselves to the supposed demands of the spoken dialogue system. Figure 4 shows that the users of this group do not act in a uniform manner.

3.2. Articulation, loudness and focus

About 40% of the adults' repetitions were labeled as more clearly articulated than the original input, as can be seen in Figure 5. The corresponding figure for the children was 28%. However, a small number of utterances became less clearly articulated as they were repeated. Adults and children appear to behave in a similar way in this respect.

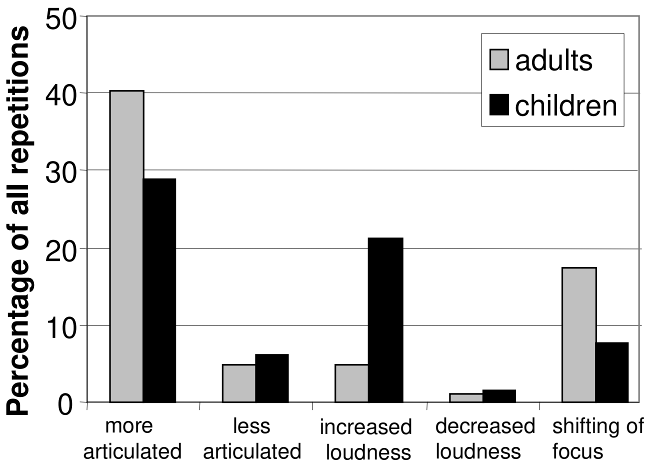

Figure 5. Distinguishing features in repetition

Previous studies have shown that while it is common for people to speak louder during human-human error resolution, this is not the case in human-computer error resolution [4]. In the present study, however, 21% of the children's repetitions were labeled as increased in loudness. The same cannot be shown for adult users, where increased loudness occurred in only 5% of the repetitions. Figure 5 also shows that focus shifting in the repeated utterances occurred in 17% of the adults' utterances and in 7% of the children's

4. DISCUSSION

Most people adapt their manner of speaking to meet the demands of a spoken dialogue system. One third of the repetitions in the current database, however, were not labeled as different from the original input to the system. Little or no adaptation took place, which could be explained by the fact that the users were sometimes unsure of whether their original input had been correctly processed by the system.

Users often move from conversational to clear speech during repetition. This partly explains some of the distinguishing features described in this paper. The increase in average utterance duration is one of those features, and hyperarticulation is another. Inserted pauses were much more frequent in the repetitions than in the original utterances to the system, and they became increasingly frequent the longer the repetitive sequence lasted. It appeared that the users of the system believed that they could resolve errors by means of modifying their articulation. This is one way of indicating a contrast between the original input and repetition.

In the present study, some differences between the strategies used by adults and children were observed. Focus shifting in the repetitions occurred, but primarily among the adult users of our system. Children, on the other hand, tended not to increase the duration of their repetitions, but rather to speak louder. This difference in adult and children strategies could have a number of explanations. It could be argued that while adults believe the system did not 'understand' them the first time, children think the system did not 'hear' what they were saying.

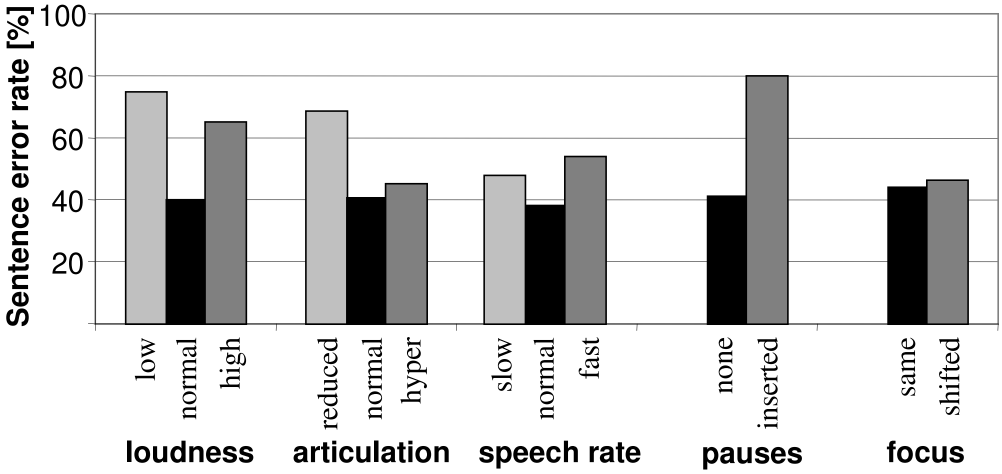

Figure 6. Sentence recognition error rates grouped by the linguistic features as labeled in the database.

To examine some possible implications of what has been discussed in this paper, all 452 utterances were analyzed in an experimental speech recognition test. The recognition lexicon was constructed by adding all words that occurred in these utterances to the lexicon that was used in the actual spoken dialogue system. In this preliminary experiment, the total sentence error rate was 44%. The adult error rate was 37% while the error rate for children was 65%. Figure 6 indicates that computer-directed speech should be as neutral or unaffected as possible to be correctly recognized. From the point of view of speech recognition, fast and reduced speech is more difficult to handle than slow and hyperarticulated speech. The lowest recognition rates in the present test were observed in utterances with inserted pauses, as can be seen in Figure 6. The explanation for this is that the recognizer used in this experiment had an insufficient model for silent segments within an utterance. Research has shown that the difficulty for speech recognizers trained on continuous speech to handle words spoken in isolation will not be solved by simply adding isolated speech to the training material [8]. The solution might be to have an isolated speech recognizer run in parallel, and let the dialogue manager predict which one to use depending on the situation.

5. CONCLUSION

In this paper, we have shown that there are several acoustic and prosodic features that make repetitions distinguishable from original input to a spoken dialogue system. The repetitive utterances in the current database are longer in duration, more articulated and sometimes spoken with a louder voice than the original input. Results indicate that these variations in speaking style may in fact make recognition rates worse, so that they interfere with the users' intentions. It is an important task for developers of future systems with spoken input to make users aware of how they should speak in order to be understood. Speech recognition technology faces a difficult task if several varieties of spoken language are to be correctly handled. The different speaking styles in computer-directed speech may also have implications for dialogue management, especially during error resolution, and for improving automatic speech recognition.

ACKNOWLEDGEMENTS

The authors would like to thank all the speakers who contributed to our research by talking to August, our animated agent. We would also like to thank all the people who have been involved in the development of the August system at the Centre for Speech Technology. Rolf Carlson, David House and Mattias Heldner have been helpful in the process of writing this paper and we are grateful for their comments and suggestions.

REFERENCES:

[1] Gustafson, J., Lindberg N., Lundeberg M., Svensson, E-L. 1999. The August spoken dialogue system. Submitted for publication, Eurospeech '99.

[2] Aitchison, J. 1994. "Say, say it again Sam" The Treatment of Repetition in Linguistics. In Fischer, A. (ed.) *Repetition*. Tubingen: Gunter Narr Verlag.

[3] Swerts, M. Koiso, H., Shimojima, A. and Katagiri, Y 1998. On the different functions of repetitive utterances. In *Proceedings of ICSLP '98*

[4] Oviatt S., Levow, G, MacEachern, M. and Kuhn, K. 1996. Modeling hyperarticulate speech during human-computer error resolution. In *Proceedings ICSLP '96*

[5] Levow, G. 1998. Characterizing and recognizing spoken corrections in human-computer dialogue. In *Proceedings of COLING/ACL'98*

[6] Oviatt, S and VanGent, R. 1996. Error resolution during multimodal human-computer interaction. In *Proceedings of ICSLP' 96*

[7] Bell, L. and Gustafson, J. 1999. Interacting with an animated agent: User strategies in a spoken dialogue system. Submitted for publication, Eurospeech '99.

[8] Alleva, F. Huang, X., Hwang, M-Y and Jiang, L. 1997. Can continuous speech recognizers handle isolated speech? In *Proceedings of Eurospeech '97*