Speech and Speech Recognition during Dictation Corrections

Keith Vertanen

Department of Physics, University of Cambridge
Cavendish Laboratory, Madingley Road, Cambridge, CB3 0HE, UK
kv227@cam.ac.uk

Abstract

A natural way to correct errors made while dictating to a computer is to respeak portions of the original sentence. But often spoken corrections are themselves misrecognized, costing the user time and testing their patience. To better understand how users behave while correcting, I created a simulated dictation interface and fooled users into believing they were correcting errors by respeaking. I found that users not only hyperarticulate during corrections, but they do so preemptively before any misrecognition. Depending on the recognizer, hyperarticulation was found to cause relatively minor changes in error rate. The correction of isolated words or phrases was more troublesome, causing substantial recognition problems for an HTK recognizer. Dragon Naturally Speaking, on the other hand, performed slightly better on hyperarticulated speech and only degraded slightly on isolated corrections.

Index Terms: speech recognition, error correction, dictation, hyperarticulation, correcting by respeaking

1. Introduction

When using a large vocabulary continuous speech recognizer to dictate text, correcting errors dominates task time [1, 2]. While users show a strong preference for correcting errors by voice, this strategy usually proves inefficient and frustrating [3].

So why are spoken corrections problematic? A possible reason is that while correcting dictation errors, people adopt a more hyperarticulated speaking style. During error resolution, users have been shown to slow their speaking rate, add pauses, and pronounce words more carefully [4, 5]. This causes a mismatch between the user's hyperarticulate speech and the naturally-read speech typical of a recognizer's training data. Increased recognition errors have been shown for utterances repeated after errors in spoken dialog systems [6, 7] and for hyperarticulated isolated words [8]. Instructing users to always "speak naturally" reduces but does not eliminate user's tendency to hyperarticulate [9, 7].

So how *do* novice users speak to a dictation interface? Do they hyperarticulate? If so, what effect does this have on recognition accuracy? To investigate these questions, I had people use a simulated dictation interface and correct errors by respeaking. Using three state-of-the-art commercial and research speech recognizers, experiments were conducted to see how user's speaking style impacted recognition accuracy.

2. Data collection

In this study, 24 volunteer users were recruited to use a dictation-style interface. Users were native North American English speakers with no prior experience dictating to a computer. Users were gender balanced and aged from 24 to 59 (average 31).

Users were told they would be using a speech recognizer and correcting any errors by respeaking parts of the text. As we will see, this was not true as the "dictation" interface was merely recording their voices. Each trial took place in a quiet environment and used the same laptop and Plantronics DSP 400 USB microphone. Users first completed the standard voice enrollment session in Dragon Naturally Speaking v8.1. The enrollment text reminded users multiple times to speak naturally – like "newscasters read the news". For the experiments reported here, the enrollment data was not used and served merely to familiarize users with dictating.

In part one of the experiment, users read 42 sentences chosen from the WSJ1 Hub 2 test set. A quarter of the sentences were hard by design, including out of vocabulary (OOV) words at the WSJ 20K and 64K levels (6.9% OOV 20K, 3.1% OOV 64K). The order of the sentences, aside from two initial practice sentences, was randomized for each user. Users were told no recognition would occur in part one and to read the sentences "naturally".

In part two, users were instructed to read the sentences again but that this time the computer would try and recognize their speech. After a "successful" recognition, a happy tone would play and the next sentence would appear. After a "failed" recognition, a sad tone would play and the recognition hypothesis would appear below the original sentence with word errors displayed in red (figure 1). A word, phrase or the entire original sentence would then be highlighted and the user would respeak the highlighted text. Users repeated speaking the highlighted text until recognition was "successful". If a user made a genuine speaking mistake such as omitting a word, they were instructed to rerecord that utterance.

In actuality, no recognition took place in part two. The user's audio was recorded but had no influence on the "success" or "failure" of recognition. Each sentence had a predetermined set of "errors" which every user experienced. These simulated errors were based in part on actual recognition results obtained on the sentences by the author. The number of simulated errors per sentence was varied from zero to five with an average of 3.5 corrections per sentence. For any sentence with one or more errors, users were required to repeat a single word (8 sentences), a portion of the sentence (16 sentences), or the entire sentence (8 sentences). A total of 5.9 hours of audio was collected.

Figure 1: User asked to respeak "but the get together wasn't".

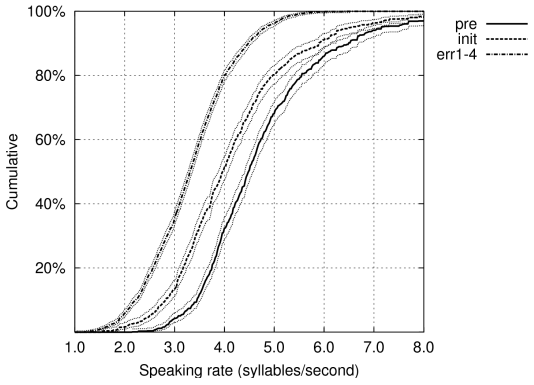

Figure 2: Cumulative distribution of speaking rate (dotted lines are two-sigma error bars).

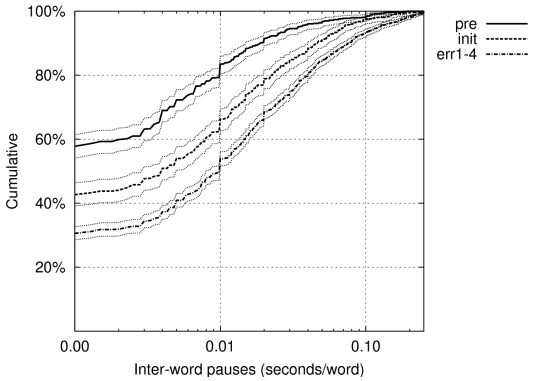

Figure 3: Cumulative distribution of inter-word pausing (dotted lines are two-sigma error bars).

3. Speech changes during corrections

To measure changes in user's speech during the experiment, word- and phone-segmentations of the utterances were obtained by forced alignment using the correct transcription and the HTK recognizer. Pitch, intensity, and formant frequencies for each utterance were calculated using praat [10]. As in [7], the degree of hyperarticulation in each utterance was judged on a three-point scale (0 = normal, 1 = somewhat hyperarticulate, 2 = strongly hyperarticulate). Judging was done by the author in random order and without knowledge of which utterance instance was being scored.

In the analysis which follows, utterances are identified as follows:

- `pre` - complete sentence, collected before any simulated errors (part one of experiment)
- `init` - complete sentence, collected before any errors on that particular sentence (part two of experiment)
- `err1-4` - error corrections by respeaking a word, phrase, or sentence (`err1` is the first correction, `err2` is the second, etc)

For purposes of comparison, just the audio sections corresponding to the words in `err1-4` were analyzed in `pre` and `init`.

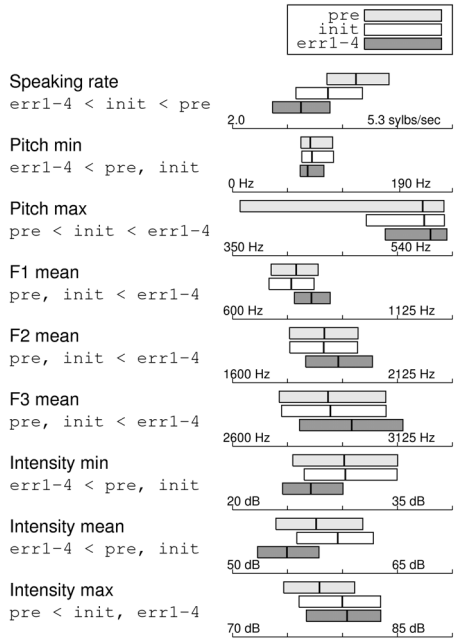

Figure 4: Differences between `pre`, `init`, and `err1-4` (boxes show lower and upper quartiles, the median is the central line).

3.1. Duration and pausing

The speaking rate in syllables per second was calculated for each utterance, removing starting and ending silence using the forced alignment. As shown in figure 2, even before errors occurred on a particular sentence, the `init` utterances were slowed in comparison to the `pre` utterances. A further speaking rate reduction is seen on the `err1-4` error corrections. Speaking rate was not found to differ significantly among the error instances `err1`, `err2`, `err3` and `err4`. The forced alignments also show increasing amounts of inter-word silence in the `init` and `err1-4` utterances (figure 3).

3.2. Pitch, intensity and formant frequencies

Similar to duration and pausing, cumulative distributions were found for the min, max and mean of intensity, the min, max, mean and slope of pitch, and the mean of formants F1-F5. The differences reported in figure 4 were significant in the sense that the two-sigma error bars of the cumulative distributions were non-overlapping. For brevity, the distributions are summarized here by their quartiles. It was found that during corrections, users tended to expand their pitch range, lower their intensity and increase formant frequencies F1-F3.

3.3. User strategies

Aside from the initial Dragon instructional text, no intervention or advice was given to the user on how to speak during the experiment. Users employed a wide variety of strategies in their efforts for correct recognition. Some users consistently hyperarticulated corrections while others spoke normally throughout. Others explored different strategies, changing between normal, hyperarticulated and hypoarticulated speaking styles. A number also tried isolated speech, inserting long pauses between every word. The

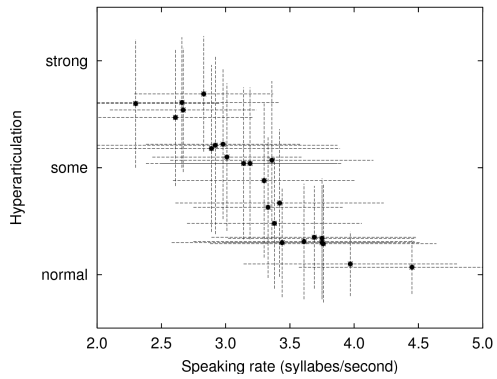

Figure 5: Speaking rate and judged hyperarticulation on `err1-4` (dotted lines are one-sigma error bars).

spectrum of user behaviors and the correlation between judged hyperarticulation and speaking rate is shown in figure 5.

4. Recognition experiments

Speech recognition performance during corrections was investigated using using three different speech recognizers: Microsoft's Speech SDK v5.1, Nuance's Dragon Naturally Speaking v8.1, and Cambridge's HTK v3.3. All recognizers were used in a speaker-independent fashion. Microsoft was set to maximum accuracy and accessed via the SAPI API. Dragon was set to default accuracy and accessed via the C++ SDK.

Training of HTK followed the recipe available from [11]. Audio was parametrized into 12 Mel Frequency Cepstral Coefficients plus the 0th cepstral, deltas and delta deltas, normalized using cepstral mean subtraction. 39 phones were used from the CMU dictionary with each phone having three output states and a left-to-right topology with self-loops. Monophones were bootstrapped from TIMIT and cross-word triphones trained on WSJ (SI-284 training set, 66 hours). The system used 16 Gaussians per output state and 32 Gaussians for silence states, all with diagonal covariance matrices. Output states were tied using a phonetic decision tree. This yielded a gender-independent model with about 9.3 million parameters. Bigram and trigram language models were trained on English Gigaword and used a vocabulary of the top 60K words from the corpus. For recognition, the bigram was used to generate a word lattice which was expanded and rescored using the trigram.

Table 1 shows word error rate (WER) and real-time factor (on a 2.8GHz Pentium 4) for each recognizer on the San Jose Mercury sentences from the WSJ Hub 2 test set (207 sentences).

Recognizer	WER	Real-time factor
Microsoft	$32.1\% \pm 0.8\%$	0.55
Dragon	$23.9\% \pm 0.7\%$	0.49
HTK	$20.5\% \pm 0.7\%$	23.00

Table 1: Recognizer performance on WSJ Hub 2 test set.

	`pre`	`init`
Microsoft	$31.3\% \pm 0.5\%$	$31.1\% \pm 0.5\%$
Dragon	$22.5\% \pm 0.4\%$	$18.5\% \pm 0.4\%$
HTK	$19.9\% \pm 0.4\%$	$18.9\% \pm 0.4\%$

Table 2: WER on `pre` and `init` utterances.

4.1. Whole sentence experiments

Comparing the initial reading of a sentence (`pre`) to the second reading (`init`), user's utterances increased in length by 18% on average. Users tended to hyperarticulate more on `init` utterances with the judged score increasing from 0.09 to 0.58. Recognition errors however decreased for all recognizers on `init` utterances (table 2).

For eight sentences, after the two initial readings (`pre`, `init`), three full-sentence error corrections (`err1-3`) were made. As shown in table 3, measures of hyperarticulation increased in `init` and `err1-3` utterances. This did not however adversely affect recognition: error rates remained similar or decreased on repeated sentences (figure 6).

4.2. Partial sentence corrections

For 12 sentences, after the two initial readings of the full sentence (`pre`, `init`), three word- or phrase-corrections (`err1-3`) were made. Hyperarticulation increased from a judged average of 0.08 on `pre` to 0.80 on `err1-3` with a 29% reduction in speaking rate.

The error rates of the three recognizers on the corresponding fragments of the `pre` and `init` utterances were found by aligning the full sentence recognition results with the `err1-3` text. Errors increased for the word- and phrase-corrections in isolation as compared to when carried within a full sentence (figure 7). Note that no surrounding context was used during `err1-3` recognition, making the recognizer's job harder than strictly necessary. HTK degraded markedly on word- and phrase-corrections while Dragon coped reasonably well. While the details of Dragon are not known, it does suggest there are ways to improve accuracy on partial sentence corrections. Perhaps Dragon has included isolated speech in its training data, similar to [12].

	`pre`	`init`	`err1`	`err2`	`err3`
Judged score †	0.17	0.73	0.99	0.89	0.96
Syllables/sec	4.30	3.78	3.57	3.62	3.60

Table 3: Hyperarticulation on whole sentence repetitions. † 0 = normal, 1 = somewhat, 2 = strongly hyperarticulate

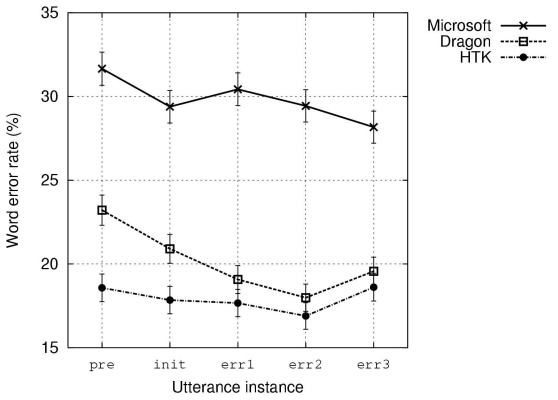

Figure 6: WER on whole sentence corrections (one-sigma error bars).

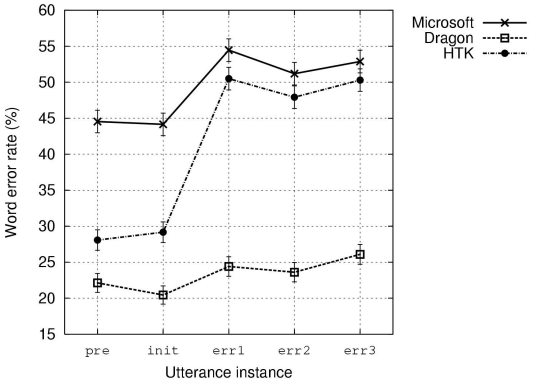

Figure 7: WER on partial sentence corrections (one-sigma error bars).

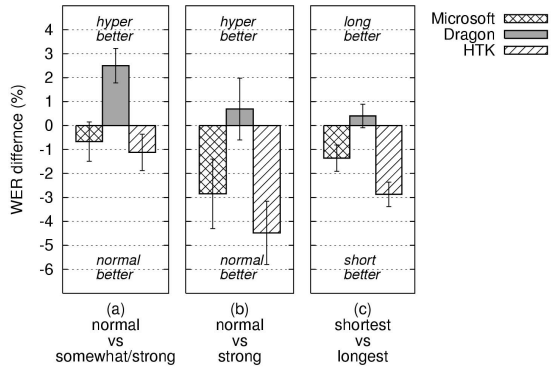

Figure 8: Difference in WER between within subject paired utterances (one-sigma error bars).

4.3. Within subject experiments

Utterances judged normally-spoken were paired within subjects to lexically identical utterances judged somewhat or strongly hyperarticulated. Microsoft and HTK did slightly worse recognizing hyperarticulated speech while Dragon did slightly better (figure 8a, 698 pairs). Similar results were found when I compared normally-spoken and strongly hyperarticulated utterances (figure 8b, 217 pairs) and when I compared the shortest and longest utterances (figure 8c, 1666 pairs).

5. Conclusions

By subjecting users to a simulated dictation interface with a high level of error, I showed that users have a strong tendency to change their speech. Compared to initial naturally read speech, speech during error corrections showed a slowed speaking rate, increased inter-word pausing, expanded pitch range, increased formant frequencies and a lowering in intensity. Human-judged levels of hyperarticulation increased both during error correction episodes and even before an error occurred on a sentence.

While I had expected a substantial increase in recognition errors on hyperarticulated speech, experiments showed otherwise. Despite increasing levels of hyperarticulation on repeated full sentences, recognition error rates remained similar or even decreased. Within subject pairings of word, phrase or sentence utterances did show some differences between normal and hyperarticulated speech. These differences were small and Dragon actually showed improved recognition on hyperarticulate speech.

The recognition of word or phrase corrections proved problematic for all recognizers tested. HTK in particular seemed ill-equipped in its standard form to handle such utterances. In the future, I plan on addressing this deficiency by providing the recognizer's language model with the surrounding text context and by tuning the insertion penalty. The audio from this experiment will be used to improve the acoustic model using speaker adaptation techniques. Adding isolated words and phrases to the model's training data may also prove helpful. This should lead to more robust recognition during user's corrections.

6. Acknowledgments

Thanks to all those who participated in the experiment. In addition, thanks to Phil Woodland and Matt Stuttle for their help with the HTK.

7. References

[1] C.-M. Karat, C. Halverson, D. Horn, and J. Karat, "Patterns of entry and correction in large vocabulary continuous speech recognition systems," in *Proceedings of CHI*, 1999, pp. 568–575.

[2] H. H. Koester, "Usage, performance, and satisfaction outcomes for experienced users of automatic speech recognition," *Journal of Rehabilitation Research and Development*, vol. 41, no. 5, pp. 739–755, September 2004.

[3] C. A. Halverson, D. B. Horn, C.-M. Karat, and J. Karat, "The beauty of errors: Patterns of error correction in desktop speech systems," in *Proceedings of INTERACT*, 1999.

[4] G.-A. Levow, "Characterizing and recognizing spoken corrections in human-computer dialogue," in *COLING-ACL*, 1998, pp. 736–742.

[5] S. Oviatt, "Modeling hyperarticulate speech during human-computer error resolution," in *Proceedings of ICSLP*, 1996.

[6] L. Bell and J. Gustafson, "Repetition and its phonetic realizations: Investigating a Swedish database of spontaneous computer directed speech," in *Proceedings of ICPhS*, 1999.

[7] E. Shriberg, E. Wade, and P. Price, "Human-machine problem solving using spoken language systems (SLS): Factors affecting performance and user satisfaction," in *Proceedings of the DARPA Speech and Natural Language Workshop*, 1992, pp. 49–54.

[8] H. Soltau and A. Waibel, "On the influence of hyperarticulated speech on recognition performance," in *Proceedings of ICSLP*, 1998.

[9] E. Wade, E. Shriberg, and P. Price, "User behaviors affecting speech recognition," in *Proceedings of ICSLP*, 1992, pp. 995–998.

[10] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer," http://www.praat.org/, 2006.

[11] K. Vertanen, "HTK Wall Street Journal Training Recipe," http://www.inference.phy.cam.ac.uk/kv227/htk/, 2006.

[12] F. Alleva, X. Huang, M.-Y. Hwang, and L. Jiang, "Can continuous speech recognizers handle isolated speech?" in *Proceedings of Eurospeech*, 1997, pp. 911–914.