

Measuring the Accuracy of Commercial Automated Speech Recognition Systems During Conversational Speech

Michael Broughton

Human Systems Integration group
Command and Control Division
DSTO

Michael.Broughton@dsto.defence.gov.au

Abstract

This paper presents research investigating recognition accuracy of two commercially available automated speech recognition (ASR) systems during spontaneous or conversational speech. This study is not aimed at finding the better ASR system, but more to use two systems to validate the technique used here to measure recognition accuracy during conversational speech. Evaluations of ASR systems are typically reported where participants dictate into the system from a prepared script. However, during spontaneous speech that will ideally result during interactions with Virtual Conversational Characters (VCC)s, the recognition accuracy of commercial ASRs is rarely advertised. As expected, there is a significant reduction in recognition accuracy during spontaneous speech, due to significant variations in the delivery of spontaneous and prepared speech.

1 Introduction

Speech recognition is a major component when human users interact naturally with virtual conversational characters (VCC). Although previous research has found a significantly higher rate of disfluencies in human-human dialogs than human-machine database queries (Oviatt, 1995), it is the belief of the author that as interfaces become more anthropomorphised, user's interactive dialogue will become less structured and more conversational when compared to isolated word, command driven systems.

Spontaneous or conversational speech is a more natural method of communication between people than prepared speech (Liu et al., 1998), however, it provides greater difficulties for automated speech recognition (ASR) systems due to speaker disfluencies and prosody effects. Speaker disfluencies include filled pauses, repetitions, deletions or false starts, repairs and covert sentence boundaries. Structural information, such as punctuation, assists natural language processing and is currently not captured by ASRs during spontaneous speech (Stolke et al., 1998). Prosody effects include duration of pauses, final vowels and final rhymes, the pitch relative to the speaker's baseline and the energy of the utterances.

Speaker disfluencies and prosody effects result from extra cognition being required, affecting speech production.

Numerous studies have compared the recognition accuracy of speech recognisers with an evaluation technique that has the participant reading aloud words from a prepared script (e.g. Cane, 1998). This script may contain either a list of commands, a short story containing several paragraphs, or a combination of the two. However, this method of evaluation provides an idealistic environment for the recogniser, as the user can simply read aloud the provided words without having to prepare content whilst speaking.

The challenge of this current research is to design an experiment that captures conversational speech, in a controlled manner, such that the recognition accuracy of two leading commercial ASR systems can be ascertained when conversational speech is used. The results of this study will provide anticipated recognition accuracy during spontaneous or conversational dialogues with VCCs.

2 Method

Twelve participants were randomly selected from a group of Australian born, novice speech recognition users. Two commercial ASR software applications were utilised throughout the experiment, Lernout and Houspie Dragon Naturally Speaking Version 5.0 (DNS) and IBM ViaVoice Version 8.0 (IVV). At the date of conducting this experiment, December 2001, these two ASR software applications were the latest release for their respective companies in Australia.

We are investigating speech production factor (TASK). The experiment followed a 2 x 3 factorial design with SOFTWARE x TASK as the two factors respectively, where the three levels of the TASK factor were Reading, Picture and Story. The Picture and Story tasks were designed to capture spontaneous speech. Order effects were controlled through a latin-square design on task order. Half of the participants trained IVV before DNS and vice-versa for the second half of participants. A single Toshiba Tecra 8200 laptop computer was used throughout the experiment for speech recognition and data collection. Specifically, this machine has an Intel Pentium3 850MHz processor and configured with 512MB of RAM. Windows2000 was installed as the operating system and both ASR applications were also installed on this machine. A head-worn, unidirectional, noise cancel-

ling, Shure VR250BT analogue microphone was used throughout the experiment. A Sony MZ-R55 Walkman MiniDisk (MD) recorder was used to store participant utterances relevant to the experiment. The MiniDisk volume level was set such that the input signal levels matched the output signal levels when the MiniDisk recorder was in record mode.

Each participant enrolled on both ASR systems to build an individual, speaker dependent acoustic model for each ASR system. The specified minimum training was carried out on both systems and was approximately ten minutes for DNS and twelve minutes for IVV. The participants then performed a series of three measured tasks, referred to as the Reading, Story and Picture tasks.

The Reading task involved the participant reading out aloud a short story of approximately 300 words. The Story task required the participant to silently read and comprehend a short story of approximately 80 words, and then to provide a verbal summary, in their own words, to the experimenter. This was repeated with several other short stories until the participant had provided more than 300 words. The stories for both the reading and story task were from a prepared collection of structurally equivalent short stories (Dixon, Hertzog & Hultsch, 1989). Similarly, the Picture task required the participant to examine a series of pictures and verbally describe each in turn to the experimenter until 300 words had been collected. During the Picture and Story tasks, the participants were instructed to speak normally to the experimenter, and were not advised to adjust speech production for the ASR system. Total contact time with the participant ranged from 50 to 60 minutes. The Story and Picture tasks were designed to elicit conversational speech from the participant, while the Reading task was constrained to the presented script.

At a later time, the experimenter manually played back the stored audio files from the MiniDisk into the computer containing the participant's trained user profile for both ASR systems. This generated text output files for each task from both ASR systems. The collected utterances from all three tasks were manually transcribed into text files.

To determine the recognition accuracy, a software utility named SCLITE (NIST, 1999) was used. This word accuracy scoring program compares two input files and produces a report containing recognition accuracy figures. The report provides figures for total number and percentage of insertion, deletion and substitution errors, as well as overall percentage word accuracy. The two input files are the reference file, which is the transcription of the original utterances, and the hypothesis file, which is the output from the ASR systems. SCLITE allows for alternate words in the reference file to be grouped together, allowing correct recognition of words that are not recognised exactly as spoken, but are valid alternatives in the context of the sentence. For example, "she has" is a valid alternate to "she's". In this case, the reference file would contain {she has/she's} grouped together in the transcription file, allowing either option in the hypotheses file to be treated as a valid recognition. Filled pauses were also coded as alternate word groups and included a null and an approximate utterance spelling, e.g. 'uhm', 'er' and 'and'. Stutters were also transcribed as the individual letter or

null. With a filled pause or any other disfluency, a null was used, as it was preferable the ASR system produce no output in these situations.

The word count of all files was standardised to 308 words with a six word tolerance. As required by the SCLITE program alignment tags were inserted into all files for comparison between the hypothesis and recognised files. Alignment tags allow the input files to be divided into smaller sections and the alignment accuracy increases with the number of alignment tags used. Typically alignment tags were placed no more than 20 words apart during this evaluation, providing each transcribed file with at least 15 alignment tags.

3 Results

The data was analysed with reference to the percentage word accuracy (PWA) figures resulting from the NIST word accuracy scoring utility SCLITE. Mean PWA results are shown in Figure 1. PWA includes substitution, insertion and deletion errors in the figure.

An analysis of variance (ANOVA) was carried out on the resultant PWA data providing the following results. For all results reported, there was no significant variation from sphericity on the Maunchly test, therefore sphericity can be assumed and standard F tests are reported. There was no significant effect from task order ($F < 1$) and no significant order interaction effect for either task or software ($F_s < 1$). With no order effects, task order has been removed from the following analysis.

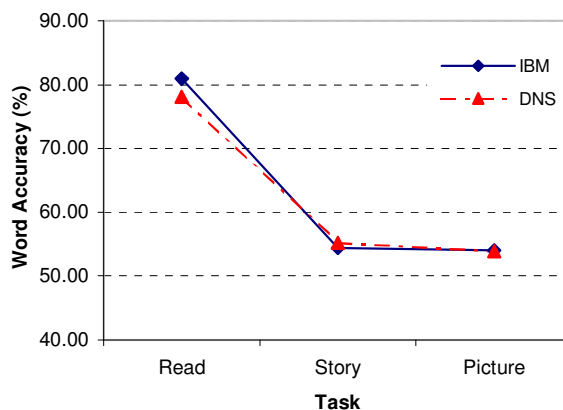


Figure 1: Mean percentage word accuracy across tasks for two commercial ASR systems.

There was an overall significant task effect ($F(2,22) = 90.612$, $MSE = 0.006$, $p < 0.001$). Specifically there was a significant effect between the Reading task and the two spontaneous speech tasks, Story and Picture tasks ($F(1,11) = 247.543$, $MSE = 0.009$, $p < 0.001$), however there was no significant effect between the two spontaneous speech tasks ($F < 1$) and these contrasts did not interact with the software factor. There was no significant effect between the two ASR software applications ($F < 1$), or an interaction between task and software. In a SOFTWARE x TASK x GENDER analysis, gender provided no effect and did not interact with the other factors.

4 Discussion

The performance figures achieved from the experiment are lower than have been published by the respective companies for a number of reasons and should be treated as worst case. Novice speech users were deliberately chosen and only allowed to do the minimal training specified by the respective ASR products. Therefore common techniques used to increase recognition accuracy, such as longer enrolment, retraining of commonly misrecognised words, and document analysis for language modelling, have not been utilised. These techniques were not included, as the casual ASR system user typically undertakes minimal enrolment and uses the system as soon as possible. In the experiment, the conversational tasks, Story and Picture, were designed to induce 'thinking on your feet' conversations, reflective of humans describing certain facts. This dialogue, when compared to reading aloud from prepared material, contains disfluencies, prosody effects and can be grammatically incorrect, all of which degrade the performance of ASR systems.

Techniques to improve the recognition accuracy of spontaneous speech are ongoing (e.g. Liu et al., 1998; Stolke et al., 1998) and it would be anticipated research developments flow through to commercially available products. For instance, Dragon NaturallySpeaking 6, which was released after the present study, contains an additional feature claiming to avoid unwanted insertion errors, by eliminating fillers and noises between dictation (Scansoft, 2002). Using the current collected data, a follow-up study to investigate this claim is proposed.

Also of future interest is the ability of ASR systems to correctly recognise key words in an utterance. As most VCCs use some form of Natural Language Processing (NLP), keywords and facts must be correctly recognised. Without customisation this is typically not guaranteed, even when the ASR system is returning word accuracy rates above 95%. Techniques to increase accuracy of keywords needs to be investigated.

5 Conclusions

This study has demonstrated there is currently a significant degradation in recognition accuracy of commercial ASR systems when conversational or spontaneous speech is used. It is anticipated that conversational speech will be more prevalent at the interface when VCCs are being used. The user needs to be aware that filled pauses and other disfluencies have a significant effect on recognition accuracy. In scenarios that provide a conversational setting, the user needs to prepare the speech train before making any utterances, 'think before you speak', to optimise system performance.

Acknowledgements

The author would like to thank Dr Glen Smith and Mr John Hansen for their help with the statistical analysis of collected data, Dr Ahmad Hashemi-Sakhtsari for the loan of hardware utilised throughout the experiment and Mr Jason Littlefield for reference to the NIST SCLITE software.

References

- Cane, J. 1998. Comparing Dragon NaturallySpeaking and IBM ViaVoice Gold, ENW International, <http://www.enw-ltd.com/vv-dragonComparison.htm>, accessed August 2002.
- Dixon, R., Hertzog, C. and Hultsch, D. 1989. A Manual of Twenty-five Three-tiered Structurally Equivalent Texts for Use in Aging Research, *CRGCA Technical Report No. 2*, April, 1989.
- Liu, D., Nguyen, L., Matsoukas, S., Davenport, J., Kubala, F. and Schwartz R. 1998. Improvements in Spontaneous Speech Recognition. *Proceedings of the Broadcast News Transcription and Understanding Workshop*, February 8-11, 1998, Lansdowne, Virginia.
- NIST 1999. NIST Broadcast News Evaluation, http://www.nist.gov/speech/tests/bnr/bnews_99/bnews_99.htm, accessed August 2002.
- Oviatt, S. 1995. Predicting spoken disfluencies during human-computer interaction, *Computer Speech and Language*, 1995, vol. 9, no. 1, pp. 19-35.
- Scansoft. 2002. Dragon NaturallySpeaking – What's New, <http://www.scansoft.com/naturallyspeaking/whatsnew>, last accessed October 2002.
- Stolcke, A., Shriberg, E., Bates, R., Ostendorf, M., Hakani, D., Plauche, M., Tur, G. & Lu, Y. 1998. Automatic Detection of Sentence Boundaries and Disfluencies based on Recognized Words. *Proceedings International Conference on Spoken Language Processing*, vol. 5, pp. 2247-2250, Sydney, Australia.

