How does automatic speech recognition handle dysarthric speech?

Elisabet Rosengren, Parimala Raghavendra, and Sheri Hunnicutt
Department of Speech Communication and Music Acoustics, KTH,
Box 70014, S-100 44 Stockholm, Sweden

ABSTRACT

This paper describes how the speech of a severely dysarthric speaker is recognized by two speech recognition systems, namely the Infovox RA system and the Prototype Swedish DragonDictate (PSDD) system. The results indicated that the PSDD system adapted to the speech of the severely dysarthric subject but with lower values than for a normal speaker. On the Infovox RA system, the dysarthric subject had a mean recognition score of 74% while the normal subject scored 97%. The results are discussed in terms of what effect the speech characteristics of the subject had on the speech recognizer.

INTRODUCTION

Automatic speech recognition (ASR) is a viable interface even for individuals with speech impairments such as dysarthria with which to access computers (Lariviere, MacKinnon, & Risebrough, 1993). However, it is not known how various types of recognition systems work with individuals with different degrees and types of dysarthria. This information would guide professionals in selecting a suitable recognizer for an individual and in using ASR with a wider population for a variety of applications.

AIM

The aim of the ongoing project is to investigate how two speech recognition systems handle dysarthric speech of mild, moderate and severe degree. One system is Infovox RA, a whole-word pattern-matching recognition system based on dynamic time warping (Carlson, Granström, & Hunnicutt, 1988) and the other is the prototype Swedish DragonDictate (PSDD) system, a phoneme-based system (Bamberg, 1990). An area of interest is also to study how well PSDD adapts to dysarthric speech in a short period of time. This paper will focus on the results of one normal speaker and one subject with severely dysarthric speech, who are two of six subjects tested so far.

METHOD

Subjects

Ditte is a 35-year-old female with spastic and athetoid cerebral palsy and severe dysarthria. She has recently completed her studies at the university and would like to find employment. She uses a wheelchair for mobility. Her motoric problems make it difficult for her to control and co-ordinate breathing, phonation and articulation. Her articulation is severely affected with imprecise consonants and distortion of certain vowels. For example, she has difficulties with /s/ and /r/, and especially with consonant clusters containing those sounds. However, most phonemes are not confused with each other. Her speech is very slow and laboured. She often has difficulties in initiating words, resulting in the production of an additional sound. Her speech is difficult to understand for unfamiliar listeners. Eskil is an 18-year-old male with normal speech, who is in his final year of high school.

Two speech recognition systems

Infovox RA 201/PC is a speaker dependent, discrete system with a vocabulary up to 150 words (Carlson, Granström, & Hunnicutt, 1988). The device stores a pattern for each vocabulary item. The PSDD is adapted from the English DragonDictate (Bamberg, 1990) and is a speaker-adaptive, discrete system with an active vocabulary of 7600 words and a 60,000-word on-line dictionary.

Procedure

The subjects worked with PSDD first and then with Infovox RA on separate days. One of the investigators demonstrated how the PSDD worked. The English DragonDictate system is said to require at least 8 hours of use by a normal speaker for the system to adapt fully. Since our goal was to investigate how well the system adapted in a brief period of time, the subjects trained only 45 command words that were considered essential (e.g., begin-spell-mode, words for the Swedish alphabet list), each word 3 times. Two pieces of text were selected and divided into sections for training and tests. In order to determine how the system adapted to a fixed vocabulary, the subjects dictated parts of the text three times. Finally the subjects dictated a free text. The PSDD statistics option was used to check the performance of the subjects and was reset for every section.

The subjects trained the Infovox RA system by reading a vocabulary of 43 single words (e.g., words for the Swedish alphabet, editorial commands, environmental commands) three times. Of the command words, 37 were the same for PSDD and Infovox RA systems. After training, the same words were said three times each to test for recognition. The correct and incorrect recognitions were noted.

RESULTS AND DISCUSSION

From the PSDD statistics, number of correct words (top choice of the recognizer) and number of words that were on the choice list (ranked between 2 and 9) were added to arrive at the correct recognition score. The normal speaker completed the training and testing for the PSDD system in less than 1 hour of active time, whereas Ditte needed about 8 hours of active time to complete the procedure. The command words were not recognized after the initial training and hence needed to be retrained.

Eskil (normal speaker)

As seen in the figure, Eskil's correct recognition score on PSDD increased by 9% from 66% to 75%. There was also a 30% increase in number of words that were recognized immediately from baseline to the test, demonstrating the adaptation process. The figure also demonstrates adaptation to the same vocabulary taking place during dictation of free text. On the whole, it appeared that the normal subject did not have any problems in using the PSDD system. His rate of speech increased for the free text, from 4 words/minute the first time to 16 words/minute the third time. Infovox: Eskil's correct recognition scores were 95% the first time, 98% the second and third time. One word was not recognized correctly all three times.

Ditte (dysarthric speaker)

The PSDD recognition scores for Ditte showed a similar pattern as for the normal speaker, but with lower values (See Figure). Ditte showed an increase from 31% at baseline to 38% at the final test. Her recognition scores for the free text showed an increase from 44% the first time to 82% the third time. Her rate increased from 4 to 9 words/minute for the same text. On average, she reached 74% correct recognition with the Infovox system. A better result was achieved by selecting and retraining alternative pronunciations of unrecognized words. All words were then recognized at least once out of three. Long commands like "Sätt på TV:n" caused most problems as Ditte could not complete the command within the system's capability of two and a half seconds for each unit.

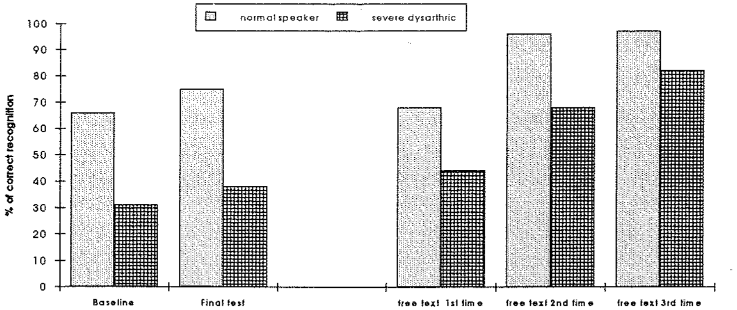

Figure 1 *: Percentage of correct recognition by PSDD system during baseline, final test and repetition of free text 3 times*

As Ditte did not pronounce sounds as clearly as a normal speaker, the risk of confusions and misrecognitions increased, resulting in lower recognition scores in general. However, there were also some specific factors in her speech that influenced the recognition. Ditte was made aware of some of these factors, which improved the recognition. Furthermore, the system was continuously adapting to Ditte's speech. Some representative examples of misrecognitions that occurred due to certain characteristics of Ditte's speech, after the brief training, are shown below:

Pronounced Recognized

1) **stava ---> talat** : The vowels were well pronounced and also recognized, but due to the distorted consonant cluster, the word was confused with another word.

2) **då ---> bror** : Difficulties with lip rounding caused problems in making a clear distinction between /o:/ and /u:/. In addition, the initial voiced stop was imprecise.

3) **kan ---> kallat:** The phonemes sounded clear, but there was a little hesitation on the /k/ and a very slow pronunciation which made the word prolonged causing it to be interpreted as a two-syllable word.

4) **Viktor ---> Bertil Kalle:** Words with a voiceless stop were often recognized as two words, because the system interpreted the long silent occlusion phase as the end of that word ("Vi-" became "Bertil"), and the final part of the same word as a new word ("-ktor" became "Kalle"). This happened even after the parameter "pause between words" was set at maximum. The phenomenon also occurred for the normal speaker.

5) Stuttering-like repetitions or hesitations when initiating a word like "**s .. så**", also led to misrecognition.

6) Involuntary sounds, e.g., loud aspirations, occurred rather frequently and were sometimes recognized as a word (e.g., "punkt").

CONCLUSION

The results for Ditte indicated that a person with severely dysarthric speech could benefit from speech recognition. It is concluded that this efficient procedure can be used to investigate whether a user would profit by PSDD. Both whole-word and phoneme-based recognition systems recognized the dysarthric speech of our subject so that decisions about selection of a system could be made based on needs of the subject. More training and adaptation time would be needed for Ditte to reach her optimal performance.

ACKNOWLEDGEMENT

This project has been supported by the Swedish Council for Social Research (SFR).

REFERENCES

Bamberg, P.G. 1990. 'Adaptable phoneme-based models for large-vocabulary speech recognition,' *Speech Characterization in Speech Technology, Proceedings of an European Speech Communication Association Workshop*, Edinburgh, 1-10.

Carlson, R., Granström, B. & Hunnicutt, S. 1988. 'Application of speech technology in aids for disabled,' in *Proceedings of Second Australian International Conference on Speech Science and Technology*, Sydney, 358-363.

Lariviere, J., MacKinnon, E. & Risebrough, N. 1993. 'Is speech recognition worth it?' *Speech and Language Technology for Disabled Persons, Proceedings of an European Speech Communication Association Workshop*, Stockholm, 95-98.