AUTOMATIC RECOGNITION OF DUTCH DYSARTHRIC SPEECH A PILOT STUDY

Eric Sanders[1], Marina Ruiter[2], Lilian Beijer[2], Helmer Strik[1]

[1]A[2]RT, Department of Language and Speech
University of Nijmegen, the Netherlands
{Sanders, Strik}@lands.let.kun.nl
[2]Sint Maartenskliniek, Nijmegen, the Netherlands
{M.Ruiter, L.Beijer}@maartenskliniek.nl

ABSTRACT

This paper describes a feasibility study into automatic recognition of Dutch dysarthric speech. Recognition experiments with speaker independent and speaker dependent models are compared, for tasks with different perplexities. The results show that speaker dependent speech recognition for dysarthric speakers is very well possible, even for higher perplexity tasks.

1. INTRODUCTION

Dysarthria is a speech disorder resulting from dysfunction of the nerves and muscles that control speech. The intelligibility of dysarthric speech is usually low, especially for unfamiliar listeners. Speech therapy can help improve intelligibility to some extent, but generally communication remains difficult.

The quality of communication can be enhanced by using Automatic Speech Recognition (ASR) technology. ASR based communication aids are potentially faster and less tiring than other communication aids like e.g. pointing or scanning aids [15]. Furthermore, in case studies it has been observed that ASR systems can outperform human listeners, in the sense that the number of words recognised correctly by ASR systems is higher than that of human listeners [2, 12]. These potential advantages of ASR technology (i.e. faster, less tiring, and higher intelligibility) could explain, at least partially, why many disabled users prefer to use speech instead of other communication aids [5], and why many individuals with physical and speech disabilities are highly motivated to learn to use ASR technology [6, 8].

There have been many studies on using ASR for dysarthric speech; an excellent overview is given in [9]. Most research has focused on using ASR as an (additional) channel of communication. However, there is another purpose for which ASR can be used by dysarthric speakers: pronunciation training. It has been observed that simply using ASR systems can improve the intelligibility of dysarthric speakers [1, 4, 11]. ASR can also be employed in systems developed specifically for pronunciation training, like the ones that already have been developed for language learning [7]. Likewise, it is possible to develop pronunciation training systems for dysarthric speakers, as is the aim of the STARDUST project [16].

So far, in most studies off-the-shelf ASR systems have been used. For instance, many papers in the literature are about case studies in which standard, commercial ASR software is used (i.e. the ASR software packages for dictation and command-and-control that are widely available nowadays). However, off-the-shelf ASR systems are probably not optimal for recognising dysarthric speech because:

(1) The pronunciation of dysarthric speakers often deviates from that of non-dysarthric speakers in several respects: rate of speech is lower, segments are pronounced differently, pronunciation is less consistent, and for longer stretches of speech pronunciation can be even more varying due to fatigue.

(2) The lack of suitable training material. For Dutch no such material existed when we began our experiments.

Although standard ASR systems are probably not optimally suited for recognising dysarthric speech, there are very few studies in which attempts have been made to develop ASR systems specifically for dysarthric speakers. This paper describes our attempt in a collaborative project between the Sint Maartenskliniek in Nijmegen [17] and the Dept. of Language & Speech. As the Sint Maartenskliniek has many dysarthric patients, the staff of this hospital is keen to know to what extent ASR can be used to their benefit. We did not have a specific application in mind. Instead, the goal of this pilot study was to conduct a feasibility study with an emphasis on the technical performance of ASR. The kind of questions we intended to answer were:

- How well can dysarthric speech be recognised by a Continuous Speech Recogniser (CSR) trained on non-dysarthric speech?
- Will the recognition results improve if we train the CSR on (a limited amount of) speech of dysarthric speakers?
- To what level of complexity are automatic recognition tasks of dysarthric speech feasible with current ASR technology?

In order to answer these questions, we conducted a series of experiments for which we used read speech of two dysarthric speakers and two non-dysarthric (reference) speakers.

The paper is structured as follows. In the next section, the speech material and the speech recogniser are described. In section 3, the experiments and the results are presented. A discussion can be found in section 4 and conclusions in 5.

2. EXPERIMENTAL SETUP

2.1. Speakers

From two Dutch dysarthric speakers, who will be referred to as DYS1 and DYS2 in the remainder of this paper, a set of utterances was recorded. The speakers were men, known to the staff members of the department of rehabilitation medicine of the Sint Maartenskliniek, and selected mainly because they have a mild form of dysarthria that is non-progressive. Still, their speech is fairly unintelligible for unfamiliar listeners. The first speaker recently had a brainstem stroke, after which he received speech training. The second one has been a dysarthric speaker from birth, and received extensive speech training.

As reference material, we also recorded the same set of utterances for two Dutch male speakers without a speaking disorder (the first and last authors of this paper). They will be referred to as REF1 and REF2. The total duration of the speech material (i.e. of the speech plus utterance internal silences) per speaker is given in Table 1. It can be seen that the speech rate for DYS2 is lower than that of the other three speakers.

Table 1: Total duration of the speech material.

DYS 1	DYS 2	REF 1	REF 2
8.5 min.	12.8 min.	9.1 min.	7.9 min.

2.2. Speech tasks

All four speakers had to read the same list of items, consisting of four different tasks:

1. NUM: the **NUM**bers '0' - '12', spoken in isolation
2. PFU: from **P**olyphone, the 50 most **F**requent **U**tterances
3. PMS: 130 **P**lomp-**M**impen **S**entences (which are semantically unpredictable sentences) [10]
4. PRS: 10 **P**honetically **R**ich **S**entences

The PRS and NUM items were read three times: near the beginning, middle, and end of the recording session. In Table 2 the number of utterances and word tokens per task are given.

Table 2: Number of utterances and words per task

	NUM	PFU	PMS	PRS
# utt.	39	50	130	30
# words	39	91	809	336

All speech material was recorded on DAT tape. The speech was transferred to a computer and the contents of the utterances were checked and corrected if necessary. If speaker noises were present in the utterances, the appropriate symbols were added to the orthographic transcription.

2.3. Polyphone material

To investigate how well dysarthric speech can be recognised by a CSR trained on non-dysarthric speech, we used speech from the Dutch Polyphone database [3]. This is a 5000-speaker corpus with 40+ recorded items per speaker. Items similar to the items of the four tasks mentioned above were selected. In this way a total of 27,834 items was obtained: 4022 connected digit strings, 3702 items with the 50 most frequent (short) utterances and 20,110 phonetically rich sentences. The Polyphone corpus contains speech recorded over the telephone, whereas the speech recorded for the DYS and REF speakers is wide band. However, previous experiments have shown that the effect of this mismatch (for the CSR described in the next section, with filter banks between 350 and 3400 Hz) is relatively small, and we expect that the effect of this mismatch is much smaller than the effect of the mismatch in speech type (i.e. dysarthric vs. non-dysarthric).

2.4. Speech recogniser

For recognition a standard phone-based CSR was used [13], which has the following characteristics. Features are extracted every 10 ms using a 16 ms Hamming window. 14 cepstral coefficients and their derivatives are computed from Mel-scaled filter banks between 350 and 3400 Hz. These features were used to train 3 state Hidden Markov Models (HMMs). 37 context independent HMMs were used, corresponding to 36 Dutch phones and 1 noise model. The CSR uses a language model consisting of a unigram and a bigram.

Results of the recognition experiments are presented as Word Error Rates (WERs), i.e. the number of substitutions, insertions and deletions divided by the total number of words in the transcriptions. In interpreting the (differences in) WERs, one should keep in mind that the number of words for some tasks is small. For instance, for the NUM and PFU tasks one additional error results in an increase in the WER by 2.6% and 1.1%, respectively.

3. EXPERIMENTS

3.1. Speaker independent models

First we wanted to know how well dysarthric speech can be recognised with HMMs trained on non-dysarthric speech from a standard corpus. To this end speaker independent (SI) HMMs were trained on 27,834 items from the Dutch speech corpus Polyphone (see section 2.3). In order to make it easier to compare results between speakers, the lexicon and language model for each task were based on orthographic transcriptions of all four speakers together. Recognition tests were carried out for the four tasks described in section 2.2, separately for each of the four speakers. The WERs are shown in Table 3. As expected, the WER for the two reference speakers is much lower than the WER for the two dysarthric speakers. DYS2 has a higher WER than DYS1 in the short utterances. This is due to a lower speaking rate (cf. Table 1), which causes many insertions (e.g. often two digits are recognized instead of one).

***Table 3*:** WERs for speaker independent recognition

	DYS1	DYS2	REF1	REF2
NUM (SI)	15.4	41.0	0.0	0.0
PFU (SI)	19.8	22.0	1.1	1.1
PMS (SI)	30.3	15.2	2.1	1.7
PRS (SI)	7.4	4.5	1.2	0.0

A similar experiment was conducted in the ENABL project [14]. For each of 5 male and 5 female dysarthric speakers the same 10 utterances were recorded (with on average 7.5 words per utterance). A CSR was trained with non-dysarthric speech. The WER for dysarthric speech was 23% for one male speaker

(with a word intelligibility of 100%), and varied from 67% to 171% for the other 9 speakers. It is, however, difficult to compare the results in [14] to our own results, because details about the (the perplexity of) the tasks are not given in [14], and because we do not know how the severity of the dysarthria of their 10 subjects compares to that of our two subjects.

3.2. Speaker dependent models

Next, we wanted to test how much the performance of the CSR could be enhanced by using speaker dependent (SD) models. Given that the available amount of speech for the four subjects was small, we decided to use a jackknife procedure. For each speaker the utterances were randomly split up in five (almost) equal parts. Five recognition experiments were carried out. In each recognition experiment 1/5 of the data was used for test, and the remaining 4/5 was used to train the HMMs. The HMMs were trained on 4/5 of the data of all tasks together. Reported test results are the average of the results of the five individual tests (thus based on the whole set of utterances).

Training was done in the following way. The SI HMMs (cf. section 3.1) were employed to obtain an initial segmentation of the training speech. Starting from this segmentation, HMMs with increasing complexity were trained: first 1 Gaussian per state, by each split the maximum number of Gaussians was doubled until after 6 splits the maximum number of Gaussians per state was 64 ($=2^6$). Recognition experiments were carried out for all these sets of HMMs to find the optimum HMM resolution for the small amount of training data.

Testing was done for the four tasks separately. The lexicon and language model were the same as those used in the SI experiments. For brevity, we first present in Table 4 the WERs computed over the whole test set for the model sets with increasing number of splits (e.g. SD3 means 3 split SD models).

Table 4: WERs for the jackknife experiments for all tasks together per model set (rows) and per speaker

	DYS1	DYS2	REF1	REF2
ALL (SD0)	14.3	7.5	3.4	3.6
ALL (SD1)	12.0	4.1	2.2	2.4
ALL (SD2)	9.5	2.9	1.8	2.8
ALL (SD3)	9.7	2.4	2.6	3.0
ALL (SD4)	10.3	3.0	3.5	3.3
ALL (SD5)	11.7	3.8	4.0	3.9
ALL (SD6)	15.1	5.3	4.2	4.4

In Table 4 we can see that the WERs (going from top to bottom) first decrease and then increase again. In previous experiments, in which a large amount of training material was available, the optimum was usually found for 6 or 7 split models [13]. Given the small amount of (training) material, it comes as no surprise that the high-resolution models perform less well than HMMs with a lower resolution; the high-resolution models are probably undertrained. On average, the best results were obtained for 2 or 3 split models (4 or 8 Gaussians per state). Below we will only present the results of the 2 split models (SD2), because they perform best on average, and all important tendencies can be seen in the results of the 2 split models.

***Table 5*:** WERs per task for the jackknife experiments

	DYS1	DYS2	REF1	REF2
NUM (SD2)	2.6	0.0	0.0	0.0
PFU (SD2)	9.9	5.5	1.1	2.2
PMS (SD2)	12.2	3.3	2.2	3.6
PRS (SD2)	3.6	1.5	1.2	1.2

In Table 5 the results of the jackknife experiments are given per task. When compared to Table 3, it can be observed that for the dysarthric speakers the WERs with SD HMMs are much better than the WERs with the SI HMMs. Relative improvements in WER vary from 50% to 100%. For the reference speakers some of the WERs are slightly higher. This is probably because the positive effect of speaker dependent training material is counterbalanced by the negative effect of less training material. The two reference speakers are recognised roughly equally well, while the values for DYS2 are clearly better than those for DYS1. In fact, for three of the four tasks the WERs for DYS2 are similar to those of the two reference speakers.

3.3. No language model and a large lexicon

The main aim of the current pilot project was to study to what extent speech recognition of dysarthric speech is possible. Given the high WERs mentioned in [14], and the fact that dysarthric speech deviates a lot from non-dysarthric speech, we decided to start with the four low perplexity tasks mentioned above (with a small lexicon and a highly constrained language model). The test-set perplexities of these four tasks are shown in Table 6.

***Table 6*:** Test set perplexity for the different tasks

NUM	PFU	PMS	PRS
13	15	8	2

Given the encouraging recognition results presented above, we decided to study what the recognition results were for tasks with higher perplexity. The perplexity was gradually increased by extending the size of the lexicon and the amount of training material for the LM, and recognition results were evaluated. Here we present the results of the experiments in which for each task no language model and a large lexicon consisting of all the 516 words of the four tasks was used. The results are shown in Table 7.

***Table 7*:** WERs for the experiments without a language model and with a large lexicon

	DYS1	DYS2	REF1	REF2
NUM (SI)	64.1	100	12.8	18.0
NUM (SD2)	28.2	20.5	0.0	10.3
PFU (SI)	89.0	84.6	37.4	33.0
PFU (SD2)	50.6	20.1	18.7	24.2
PMS (SI)	87.2	81.6	51.2	60.8
PMS (SD2)	63.8	36.5	40.4	37.3
PRS (SI)	86.4	77.4	57.1	64.9
PRS (SD2)	52.2	33.0	45.5	41.4

As expected, recognition performance drops drastically. In all cases, the results for the SD2 HMMs are better than those for the SI HMMs. The results for DYS2 for the SD2 HMMs are remarkable: for the tasks PMS and PRS the WERs for DYS2 are lower than those of the reference speakers, for the PFU task they are about equal, while for the NUM task the WERs for DYS2 are higher than those of the reference speakers and are more similar to those of DYS1.

4. DISCUSSION

The differences between the results for DYS1 and DYS2 are interesting. For speaker independent models, the WERs for the dysarthric speakers are much higher than those of the reference speakers. However, if we compare the WERs for the speaker dependent models we notice that for DYS2 they are much lower than for DYS1 and are similar to those for the reference speakers.

Most likely this is due to the fact that the speech rate of DYS2 is lower than that of the other three speakers, and thus the amount of inter-word coarticulation will probably be less for DYS2. This is reflected in the WERs in Table 7: for DYS2 the relative differences in WERS between the tasks (isolated numbers and utterances of different lengths) is smaller than the relative differences for the other three speakers.

In the near future we intend to explore whether the performance of the CSR can be enhanced by, e.g., speaker adaptation, lexicon adaptation, and tuning the internal parameters of the CSR.

5. CONCLUSIONS

In our experiments we find WERs ranging from 4.5% tot 41.0% for dysarthric speech that is recognised with HMMs trained on non-dysarthric speech. Large relative improvements of 50% to 100% in these WERs were found when the HMMs were trained on a limited amount of speaker specific material. The resulting WERs show that ASR of dysarthric speech is certainly possible for these low-perplexity tasks. For the high-perplexity tasks (in which no language model was used) the WERs for DYS2, who had a lower speech rate than the other speakers, were on average similar to the WERs of the reference speakers. These results are encouraging, and indicate that also for higher perplexity tasks ASR of dysarthric speech is within reach, especially when the user speaks at a relatively slow pace.

6. ACKNOWLEDGEMENT

This study was sponsored by, and carried out in cooperation with the Sint Maartenskliniek in Nijmegen. We would like to thank (in alphabetical order) Sander Geurts, Petri Holtus and Jacques van Limbeek of the Sint Maartenskliniek; and Lou Boves and Toni Rietveld of the University of Nijmegen, for useful discussions about this project. Furthermore, we would like to thank the two dysarthric speakers, for their willingness to participate in this experiment.

7. REFERENCES

[1] Arnold D. (1998) Speech Recognition Application and Limitations for Motor Impaired, Learning Disabled and Speech Impaired Operators. Proceedings of CSUN-98, Northridge, 1998.

[2] Carlson, G. S., Bernstein, J. (1987) Speech recognition of impaired speech. Proceedings of RESNA 10th Annual Conference, pp. 103-105.

[3] Damhuis, M., Boogaart, T., Veld, C. in 't, Versteijlen, M., Schelvis, W., Bos, L., Boves, L. (1994) Creation and Analysis of the Dutch Polyphone Corpus. Proceedings of ICSLP '94 Yokohama, Japan.

[4] Donegan M. (2000) Voice Recognition Technology in Education - Factors for Success, ACE Centre/Becta publication, 2000. ISBN 1 903303 00

[5] Ferrier, L. J. (1991) Clinical study of a dysarthric adult using a Touch Talker with words strategy. Augmentative and Alternative Communication, 7(4), pp. 266-274.

[6] Fried-Oken, M. (1985) Voice recognition device as a computer interface for motor and speech impaired people. Archives of Physical Medicine and Rehabilitation, 66, pp. 678-681.

[7] Neri, A., Cucchiarini, C., Strik, H. (2001) Effective feedback on L2 pronunciation in ASR-based CALL. Proceedings of the workshop on CALL, AI-ED 2001, San Antonio, Texas, pp. 40-48.

[8] Noyes, J. M., Frankish, C. R. (1992). Speech recognition technology for individuals with disabilities. Augmentative and Alternative Communication, 8, pp. 297-303.

[9] Patel, R. (2000) Identifying information bearing prosodic parameters in severely dysarthric speech. Doctoral Dissertation, University of Toronto.

[10] Plomp, R., and Mimpen, A.M. (1979) Improving the Reliability of Testing the Speech Reception Threshold for Sentences. Audiology 18, pp. 43-52.

[11] Schmitt, D. G., Tobias, J. (1986). Enhanced Communication for the severely disabled dysarthric individual using voice recognition and speech synthesis. Proceedings of RESNA 9th Annual Conference, pp. 304-306.

[12] Stevens, G., Bernstein, J. (1985). Intelligibility and machine recognition of deaf speech. Proceedings of RESNA 8th Annual Conference, pp. 308-310.

[13] Strik, H., Russel, A. J.M., Heuvel, H. van den, Cucchiarini C., Boves, L. (1997), A spoken dialog system for the Dutch public transport information service. Int. Journal of Speech Technology, Vol. 2, No. 2, pp. 119-129.

[14] Talbot, N. (2000) Improving the speech recognition in the ENABL project. KTH TMH-QPSR 1/2000, pp. 31-38.

[15] Treviranus, J., Shein, F., Haataja, S., Parnes, P., Milner, M. (1991). Speech recognition to enhance computer access for children and young adults who are functionally non-speaking. Proceedings of RESNA 14th Annual Conference, pp. 308-310.

[16] http://www.dcs.shef.ac.uk/~pdg/stardust/

[17] http://www.maartenskliniek.nl/