From Proc. Int. Conf. on Spoken Language Processing, vol. 4, page 65–68, Beijing, China, 2000.

Automatic Stuttering Recognition using Hidden Markov Models

E. Nöth, H. Niemann, T. Haderlein, M. Decher, U. Eysholdt*, F. Rosanowski*, T. Wittenberg**

Lehrstuhl für Mustererkennung (Informatik 5), Universität Erlangen–Nürnberg,
Martensstraße 3, 91058 Erlangen, Germany
* Abteilung für Phoniatrie und Pädaudiologie Bohlenplatz 21, 91054 Erlangen, Germany
email: noeth@informatik.uni-erlangen.de

ABSTRACT

This paper describes the combination of the work of speech therapists and speech recognition systems. Our long term goal is to evaluate the degree of stuttering during therapy and to use the automatic analysis of stuttered speech as a screening method, e.g. the search for potential stutterers at an early age. The approach is to have a patient read a standard text aloud and then automatically count the unfluent parts and classify them. The text to be read by the patients is automatically transformed into a formal grammar that considers potential dysfluencies caused by stuttering.

Recordings from stutterers were compared to recordings of nonstutterers. Word and phoneme accuracies of the stuttered text in relation to the number of detected dysfluencies showed correlation coefficients of up to 0.99. Recordings from stutterers contained much more pauses in a wider time range than from nonstutterers, especially in the interval up to 200 milliseconds (factor 10), and between 200 and 500 milliseconds (factor 2). The sum of the durations of all detected pauses and the number of repetitions were set into relation. The results seem reasonable for a distinction between stutterers with many repetitions/short pauses and stutterers with few repetitions/long pauses.

1. INTRODUCTION

The analysis of speech disorders is based upon an examination of the patient's speaking ability with qualitative and quantitative registration of the symptom's type and frequency. One part of a therapy session can be to let the patient read a "phonetically balanced" standard text aloud and to record this on audio or video tape. The speech therapist then manually counts and classifies the observed stuttering symptoms in this recording. This protocol can be used to show the patient's improvements during therapy. Getting this protocol, however, is time consuming and subjective, i.e. it reflects the bias of one single therapist. So it might be desirable that some parts of the therapist's work can be done by an automatic system. Speech recognition systems can do the statistic analysis (i.e. counting and classification) of typical repetitions, pauses and phoneme durations. Such an approach can support the human experts by doing tedious routine work and thus allowing more time for the therapeutic session between patient and therapist.

2. STUTTERING

Stuttering is a very complex speech disorder with individual symptoms for every single patient. A very common phenomenon is the fast repetition of phonemes, syllable parts, syllables, words or word sequences. These repetitions occur more often at the beginning of linguistic units such as words or sentences. Many people equate them to stuttering per se, even though there are many other symptoms (dysfluencies)[1] like:

- unusual lengthening
- blocking of the vocal cords
- filled pauses
- speaking without pauses between words[2]
- uncontrolled breathing.

There are some measurable factors of stuttering that can be used to classify the degree of stuttering:

- frequency of dysfluent portions in the speech; a typical stutterer has about 10 dysfluencies, a normal speaker has about 2 disfluencies per 100 words
- duration of the dysfluencies; typical values for stutterers are in the range of 1 second
- speaking rate; stutterers typically speak about 25% fewer words than normal speakers in the same time.

For a detailed description of stutter symptoms and measurable factors for stuttering see for instance [2], [3], and [4].

[1] We use the medical term dysfluency for abnormal speech disorders instead of the usual term disfluency, which describes disorders that can be observed during normal speech.

[2] In the medical sense this is considered to be a dysfluency, even though it is a faster speaking rate.

Our goal is to use automatic speech recognition to find measurable factors. This can be used to classify the strength of the audible stuttering symptoms or the progress achieved during therapy. It is not the intention to classify between stutterers and nonstutterers. A standard test is to have the patient read a phonetically balanced text aloud (in our experiments we used the fable "Northwind and Sun") while recording it on audio or video tape. Afterwards the therapist is able to subjectively classify the progress of the therapy. As the strength of the stuttering symptoms varies differently for the individual patient depending on the communication situation (i.e. spontaneous vs. read speech), other tests are usually performed as well.

3. A GRAMMAR FOR POTENTIAL STUTTER PHENOMENA

We wanted to use speech recognition technology to calculate these factors mentioned above. Therefore, we started with the data from the read story and constructed possible phoneme strings if a stutterer reads the story.

The system can detect several classes of repetitions (from phonemes up to phrases), stretched phonemes or words, unwanted interrupts and filled pauses.

In this section we describe the generation of a pronunciation graph for the utterance to be tested. A detailed description of the grammar is given in [1]. A pronunciation graph is a directed graph, in which phonemes are nodes, and edges lead to potential successor phonemes. The pronunciation graph contains all potential realizations of the test utterance which can possibly occur due to stutter phenomena.

In the current phase of the project we assume that the patient reads a given text. For the words of the text we assume that we have a pronunciation lexicon in which syllable boundaries are marked. The punctuation of the text is used to mark potential "hot spots", i.e. locations where an increased number of stutter phenomena can be observed. Hot spots are positions like syntactic phrase boundaries and beginnings of sentences. If a patient stutters within n words after a hot spot he often repeats all the words starting from the hot spot. By default the program constructing the pronunciation graph sets n to three.

We start with a concatenation of the standard pronunciation of the spoken word sequence as a linear pronunciation graph. For each word we add edges in the graph:

1. After each phoneme a silent or filled pause can occur; thus for each phoneme we create a silent and a filled pause node, add an edge to these nodes and an edge from there to the node of the successor phoneme.
2. Each phoneme can be repeated, i.e. we insert an edge to the predecessor phoneme.
3. After each phoneme the syllable can be restarted, i.e. we insert an edge to the beginning of the current syllable.
4. After each phoneme the word can be restarted, i.e. we insert an edge to the beginning of the current word.
5. If a hot spot is within n words to the left of the phoneme, an edge to the hot spot is inserted to allow restarting the current phrase.

4. DATABASE

The database consists of 37 patients with stutter symptoms in many variations that either read all or the beginning of the fable "Northwind and Sun". Table1 shows some statistics about the database.

	male	female	total
number of speakers	27	10	37
number of recordings	49	20	69
short text	13	4	17
complete text	36	16	52
age of oldest speaker	45	31	45
age of youngest speaker	12	10	10
average age	23	18	22
duration of shortest compl. text	0:35	0:36	0:35
duration of longest compl. text	3:09	5:41	5:41
average duration of short text	0:24	0:22	0:23
average duration of long text	1:35	1:38	1:36

Table 1: Characteristics of the stutter database with the "Northwind and Sun" text. The duration is given in minutes.

To compare the results for stutterers and nonstutterers another database was used where 16 nonstutterers read the whole text of "Northwind and Sun" once.

5. EXPERIMENTS AND RESULTS

The experiments concentrated on two topics:

- How good can an automatic system find dysfluencies in stuttered speech?
- Which computable features are suitable for measuring the intensity of stuttering?

To answer the first question a set of 16 recordings of the stuttered "Northwind and Sun" text was searched for dysfluencies by hand and then compared to the results of the automatic system. The same was done for the 16 signals from the nonstutterers. Fig. 1 shows the average number of dysfluencies per word and the average detected number of dysfluencies per word. Although the system tended to overestimate the reading errors, the correlation coefficient for the results was 0.99. Fig. 2 illustrates the relation between the number of dysfluencies and the phoneme error rate[3] of the

[3]We calculated the "phoneme accuracy" between the ideal phoneme string of the read story without stuttering phenomena and the best path through the pronunciation graph with potential stuttering phenomena. The displayed Y-values are 100% - phoneme accuracy.

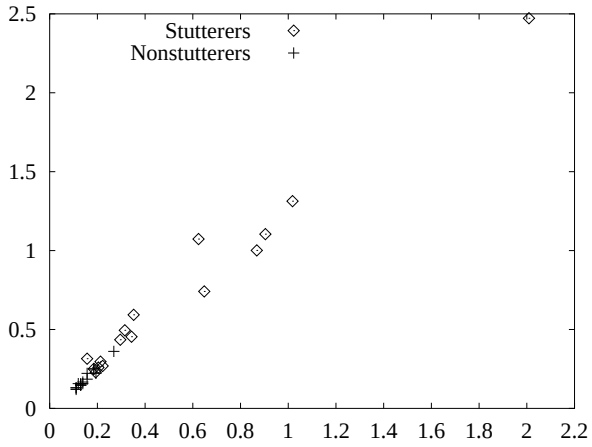

Figure 1: Number of dysfluencies (X–axis) and dysfluency hypotheses (Y–axis) per word.

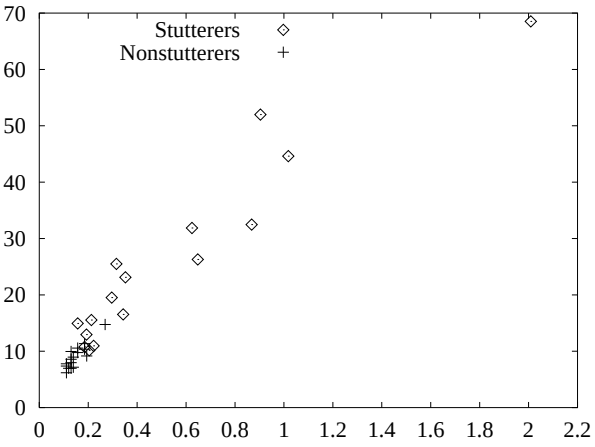

Figure 2: Number of dysfluencies (X–axis) and "phoneme error rate" in % (Y–axis).

system. Again, the number of dysfluencies is normalized to dysfluencies per word. The correlation coefficient for this set of data was 0.95.

To answer the question about measuring stuttering severity, we first analyzed the duration of phonemes in the records. For these experiments we used all the recordings with the full standard text. This attempt to find a factor for severity, however, failed because stutterers can stretch or shorten phonemes. This is illustrated in Fig. 3. It shows the fricative durations for two stuttering persons. While one distribution has a very high peak and is restricted to a short time range, the other has several maxima over a much wider stretch of time. This means that in the average no significant differences between stutterers and nonstutterers can be found. In Fig. 4 this is shown for vowel durations.

We also considered the durations of fluently spoken words in the stutterer's records and compared them to the nonstutterers but found only minor differences. A further problem is that approaches measuring only durations of spoken sequences in a record neglect the type of stuttering where the patients can speak fluently between unwanted breaks. Since those blockings can last for several seconds, we looked at the durations of pauses. The experiments showed that not only the recordings of the blocking-type stutterers had significantly more pauses than nonstutterers, but the recordings of stutterers in general. Especially short breaks in the range up to 200 ms occurred about ten times more often than in the signals of nonstutterers. In the range of 200 to 500 ms the factor was about two. Fig. 5 shows the absolute number of occurrences of pauses of different durations in the time range of up to 1.5 seconds.

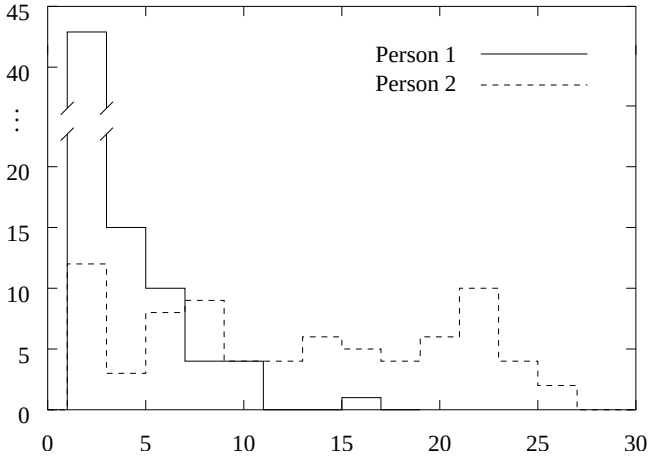

Figure 3: Distribution of fricative durations: duration in 10 ms frames (X–axis) and number of occurrences for two stutterers (Y–axis).

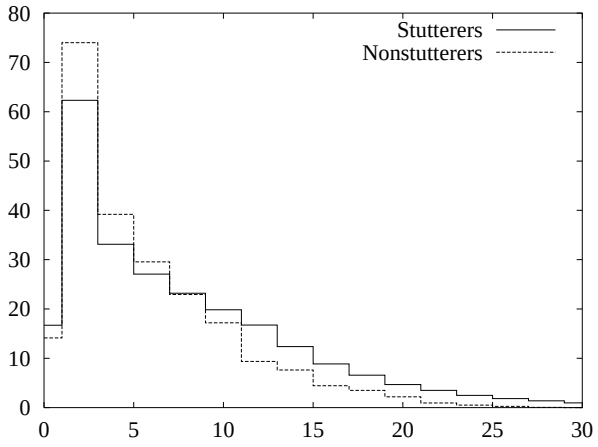

Figure 4: Vowel durations: duration in multiples of 10 ms frames (X–axis) and absolute number of occurrences (Y–axis).

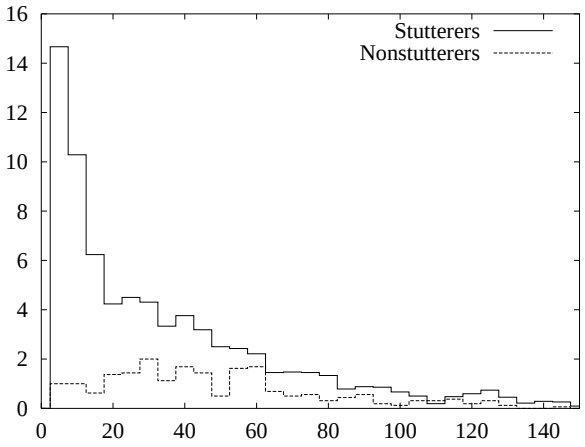

Figure 5: Duration of pauses in 10 ms frames (X–axis) and number of occurrences in one record (Y–axis).

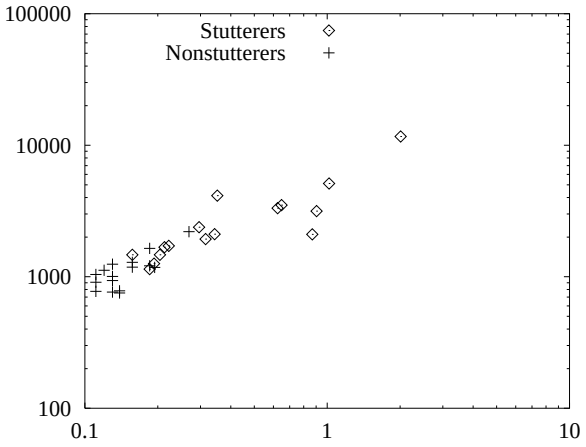

Figure 6: Number of dysfluencies in one word (X–axis) and sum of the duration of all pauses in one record (Y–axis). A logarithmic scale is used.

First results for a rather good distinction of low and high degree of stuttering and simultaneously of the repetition type and the blocking type of stuttering can be achieved by the combination of the number of dysfluencies per word and the total duration of all pauses in a record. This is illustrated in Fig. 6 for these features calculated for the signals from 16 stutterers and from the 16 nonstutterers. High values on the X–axis indicate a high degree of stuttering. Values above the regression line can be attributed to patients who tend to blocking. If the speaker tends to repetitions only, his value is below the average. Because of the small amount of data, we could not train an automatic classifier.

This distinction is still crude and needs more experiments in the future, especially with stutterers belonging clearly to either the repetition or the blocking type. These experiments will allow us to optimize both the automatic recognition of the different kinds of stuttering and the extraction of suitable factors for the qualitative analysis from the best found path in the pronunciation graph.

6. OUTLOOK AND SUMMARY

We presented an innovative approach to assess the degree of stuttering using speech recognition technology. We are currently setting up a clinical field test to collect more data with the system. At the same time we want to expand the capabilities of the system. We plan to look at the following extensions:

1. Training of the recognizer with data from stutterers. So far we used a standard speech recognizer which was trained on about 10 hours of spontaneous speech from a very different application. A retraining with the test data for which we have the stutter analysis from the human experts, can greatly increase the results for the forced alignment.
2. Detailed statistics with user friendly presentation. If the parsing tree for a record is available, the output of the system can be much more precise. Facts like "the patient repeated the first word of the second sentence twice" can be produced automatically and be printed as protocol.
3. Extention of the method to free speech. For this long-term goal, a knowledge based system for the grammar of the German language has to be developed in order to distinguish between pauses and duplications of words or word parts which are grammatically correct from those appearing as a stuttering symptom. This extension, however, is desirable because many stutterers can read fluently and so the input data for the analysis is already not correct. Free speech, however, gives us a more objective image of the speech disorder. It is also suitable for the therapy of children who cannot read properly yet.
4. Looking at other knowledge resources like prosodic information. This is especially interesting, if laryngograph recordings are available. We are looking into collecting some laryngograph recordings during the upcoming field test.

7. REFERENCES

1. T. Haderlein. Quantitative Stotteranalyse: Erstellung eines Diagnoseprogramms zur Feststellung von Unflüssigkeiten in gesprochener Rede. Studienarbeit, Lehrstuhl für Mustererkennung (Informatik 5), Universität Erlangen–Nürnberg, 2000.
2. G.D. Riley and J. Riley. Physician's Screening Procedure for Children who may Stutter. *Journal of Fluency Disorders*, 14:57–66, 1989.
3. C. Van Riper. *The Nature of Stuttering*. Prentice-Hall, Englewood Cliffs, 1971.
4. J. Wendler, W. Seidner, G. Kittel, and U. Eysholdt. *Lehrbuch der Phoniatrie und Pädaudiologie*. Georg Thieme Verlag, Stuttgart, New York, 1996.