speech, disorders,
recognition, HMM

Marek WIŚNIEWSKI[1], Wiesława KUNISZYK-JÓŹKOWIAK[2]

AUTOMATIC DETECTION OF STUTTERING IN A SPEECH

In the work authors applied speech recognition techniques to find disfluent events. The recognition system based on the Hidden Markov Model Toolkit was built and tested. The set of context dependent HMM models was trained and used to locate speech disturbances. Authors were not concentrated on specific disfluency type but tried to find any extraneous sounds in a speech signal. Patients read prepared sentences, the system recognized them and then results were compared to manual transcriptions. It allowed the system to be more robust and enabled to find all disfluencies types appearing at word boundaries. Such system can by utilized in many ways, for example like a "preprocessor"that finds strange sounds in a speech to be analyzed or classified by other algorithms later, to evaluate or track therapy process of stuttering people, to evaluate speech fluency by ńormalśpeakers, etc.

1. INTRODUCTION

Stuttering is a serious problem concerning about 1% of population [1]. It causes problems in an interpersonal communication and should be subjected to the treatment that includes diagnosis and therapy. The common approach is employing a speech therapist who "manuallyąssesses stuttering severity and applies proper treatment. Severity assessment is a tedious work and consists of identification, classification and counting disfluencies appearing in a patient speech. Furthermore, during the therapy process, these steps must be repeated many times in order to judge the progress. It's a very time consuming and not always objective.

Due to the above there are reliable counting and classification systems of speech disorders very desirable.

There are many studies that concentrate on computer diagnosis of stuttered speech. They utilize number of parameterization ways and different kinds of classification algorithms. Tests usually are performed under different conditions and concern selected aspects of stuttering. Achieved results are generally promising (usually above eighty percent of proper detections) but are difficult to compare because of different circumstances in which they are obtained.

2. REVIEW OF AUTOMATIC STUTTERIG RECOGNITION STUDIES

One of the most often used method of judgement of the speech fluency are Artificial Neural Networks (ANNs). ANNs are trained on a set of parameterized samples that contains some

[1]Institute of Computer Science, Maria Curie-Skłodowska University, ul. Akademicka 9, 20-031 Lublin, Poland, e-mail: marek.wisniewski@poczta.umcs.lublin.pl

[2]Józef Piłsudski University of Physical Education in Warsaw, Faculty of Physical Education and Sport in Biała Podlaska, ul. Akademicka 2, 21-500 Biała Podlaska, Poland

kinds of disturbances and afterwards they are used as classifiers. ANNs are given unknown speech fragments and mark them as fluent or not. In general one ANN is trained to recognize one disturbance type (fricatives prolongations, plosives repetitions, blocades, etc.).

Authors of the [10], [11] applied ANNs to classify speech utterances as fluent or not. They analyzed a set of eighty 4-seconds length utterances where half of them were fluent and the rest included one or more disfluencies of one type - blockades before words starting with stop consonants. One of the reason the authors had chosen that type of disfluency was that it was well recognized by ANNs in their previous researches. Utterances were recorded at various stages of patient's therapy and in two situations: reading a story and describing an illustration. Each recording (sample frequency 22050 Hz, resolution 16 bits) was parameterized to the sequence of 171 vectors with 21 elements (parameterizing steps: FFT analysis on 512 points, filtering over 21 digital 1/3-octave filters with center frequencies between 100 and 10000 Hz, weighing with an A-weighting filter).

First the authors applied Kohonen neural network (with 21 inputs and 25 outputs) in order to map vectors parameters into scalars (each 21 elements vector was mapped to one of the 25 numbers). Next, they applied Multi-Layer Perceptron (MLP) neural network for classification. They taught and tested several MLPs with one or two hidden layers and with different number of neurons. Next, the best MLP network was selected (number of neurons: 171 inputs, 53 in hidden layers, 1 output) and there was classification performed against additional 30 disfluent utterances not used during the teaching process. The best final result achieved was about 77% of correctly classified samples.

Automatic recognition of repetitions was also studied in [4]. The authors utilized Continuous Wavelet Transform (CWT) with 18 bark scales as a parameterization method and ANNs as classifiers. The recognition idea was to divide recognized utterances into short pieces (3 seconds length) and then decide whether they are fluent or not. As the test sample Polish stuttered speech recording was used of 548 seconds length. It was divided into 3 seconds length fluent and disfluent fragments. Division was made manually for disfluent fragments and automatically for the fluent. Finally there were 275 fluent and 288 disfluent fragments. Each fragment, described by the set of vector parameters (CWT), was analyzed by the Kohonen network. It reduced parameters space from 3 to 2 dimensions. Further the output from the Kohonen network was given to the input of a MLP network. The MLP network acted as a classifier - it decided whether given fragment is fluent or not. They used modified teaching algorithm for Kohonen network to obtain better stability and performed series of tests to find optimal configuration of MLP network. They used described above audio data and divided it into three portions: teaching (50%), verifying (25%) and testing (25%). The best result achieved was 88% of sensitivity and 90% of predictability.

In [3] the same authors studied some recognition methods of another stuttering type like prolongations. They used parameterization of audio files similar to the described above. First they found non-silent fragments of certain length (longer than 200 ms) in a speech signal. Such fragments were treated as potential prolongations. Next, Kohonen network was used to map sequence of vector parameters into a sequence of winning neurons for each fragment. After that the authors checked whether there was a sequence of the same neurons long enough. If found sequence was longer than the established threshold (200 ms) it was considered the prolongation. The authors performed series of tests analyzing Polish utterance of the length of over 18 minutes and containing 373 prolongations. The best result obtained was: sensitivity 92% and predictability 82%.

One can notice that results obtained using ANNs are quite reasonable but there are also some drawbacks. The main problem of ANNs is that training samples are required that must be manually prepared. Another issue is that ANNs can classify fragments of a fixed length.

ANNs work on short samples and decide whether they are fluent as a whole. If a sample is too long then it must be divided what can lead to new problems.

Another approach was presented in the work [2], where authors resigned form ANNs and utilized correlation algorithm as a classifier. The aim of the work was to find syllable repetitions automatically in a continuous speech. As in many previous works they used Continuous Wavelet Transform with bark scales to parametrize audio samples. They tested series of selected Polish utterances coming from 5 persons, containing syllables repetitions, where each of them was surrounded by 4 seconds length fluent speech. Finally all such fragments were connected into one large sample of 5 min 26 sec length and contained 106 repetitions.

The recognition process was as follows. They isolated sequence of non-silent fragments (i.e. words or phrases) using proper threshold and applied a correlation algorithm to each of the two consecutive fragments. However, not all fragments pairs were taken into account - only those where lengths of fragments were in the range of 70 ms - 500 ms and only those where the second fragment was shorter no more than 100 ms than the first one. All other cases were allowed. According to the authors these restrictions arose from studies on statistical lengths of disturbed syllables. The authors performed series of tests to find the optimal wavelet transformation and the correlation threshold that indicate syllable repetition. They compared obtained results with the manual annotation using sensitivity and predictability ratios. The best obtained recognition was: sensibility 81% and predictability 83%. The authors emphasize the results were obtained by the fully automatic procedure (without a teacher) on the one long continuous speech sample.

Results obtained in the above work are quite reasonable but the recognition procedure is complicated and subjected to some circumstances.

Other approaches for evaluation of a stuttering speech are statistical methods commonly used in speech recognition systems and based on Hidden Markov Models (HMMs) [5]. HMMs can be utilized in many ways - from a simple pattern recognition methods to very complicated systems in which many kinds of speech pathology aspect can be modeled.

The authors of the work [12] used HMMs to locate speech disorders in a continuous speech. The assumption was to build the system that gives phonetic transcription of the input signal and then it finds disfluencies using, for example, regular expressions. Such the way could give the information about the kind of a disturbance and about the most often disturbed phonemes . In that work there were phoneme repetitions recognized. From tested utterances phonetic transcriptions were obtained and checked if there are sequences compounded from plosives (stop consonants) and silence. Such sequences were treated as repetitions. Additionally, disfluent phonemes were identified. The authors tested 79 utterances where 20 of them included plosives repetitions. Obtained ratio of sensitivity was 89% and predictability 94%.

Another concept was presented by Noth and collegues [8], [7]. They utilized HMMs mainly in order to evaluate speech fluency and took many factors into account like ratio of stuttered fragments to all articulated words (typically about 10% for a disfluent and 2% for a fluent speech), speaking rate (number of word articulated per slot time - for a disfluent speech it is about 25% lower than for a normal speaking person), the average length of stuttered and fluent words, the average length of silence pauses and filled pauses (pauses when some sound is articulated like Polish inclusion of "y"[X-SAMPA: I] or ę"[X-SAMPA: E]). To perform tests they used a story to be read by stuttering persons. They prepared directed graphs (grammar) of uttered sentences and included there all potential articulations variants. Tests were performed on groups of 37 persons of different age and sex suffering on different disturbances. The people read the story and the system found the best transcription under the defined constraints. After that obtained transcriptions were compared to manual annotations. According to the authors the correlation between automatic and manual recognition was 99%, what is the excellent result.

Comparing the studies described above, HMMs seems to be a very flexible way of a disturbed speech modeling. It allows a signal analyzing of any length and using any parameterizations. Application of some constrains during the recognition process can greatly narrow false alternatives.

3. STUTTERING RECOGNITION CONCEPT

The idea was to recognize the input signal and compare results with manual annotations. Stuttering persons articulated earlier prepared sentences and the system recognizes them. Next, differences between automatic and manual transcription were found. For example, a person had spoken the Polish sentence: "I one też czują oddziaływanie czarodziejskiego domku". The system recognized it as: "I one też czują o oddziaływanie cza czarodziejskiego domku". Additional words 'o' and 'cza' were interpreted as disfluencies.

In order to perform tests the system using HMMs was build. It was based on context dependent models (triphones) with the adaptation to each recognized speaker.

4. BUILDING THE SYSTEM

The system was built using the Hidden Markov Model Toolkit (HTK) [6]. For the models training and testing a lot of audio recordings were prepared. In order to automatize the work there were many PERL scripts written.

For the models training Polish part of the GlobalPhone corpus was used [9]. It includes over 24 hours recordings done with the frequency of 16000 Hz and the amplitude resolution of 16 bits. Recordings came from 100 persons of different sex and age.

4.1. AUDIO SIGNAL PARAMETRIZATION

Input signals were parameterized in the following way. Frames had 32 ms length (512 samples) and were taken with the step of 10 ms. Frames were treated by pre-emphasis (filter parameter 0.97) and Hamming filters. Next, Fourier analysis was performed and the results were filtered by 26 triangular filters (filters had equal width on the Mel scale and spread over the whole frequency range up to the value of 8000 Hz). After that MFCC ceofficients and their first and second derivatives were calculated. Finally each frame feature vector consisted of 39 elements: 13 MFCC coefficients, 13 of the first and 13 of second derivatives (i.e. delta and acceleration parameters).

4.2. MODELS TRAINING

The most important part of the system were phoneme models. Below successive steps of models preparation are briefly presented. The whole process included two phases and in the first one context independent models (monophones) were trained. All training steps were done automatically using authors scripts and HTK tools (initializing: HCompV, training: HERest, realigning and recognition: HVite)

Monophones training steps:

- prototype model definition - left to right (LR) with three emitting states,
- prototype model initializing - mean and variance values computed from all the training data,
- creating full set of models (for 36 Polish phonemes) and additional silence (sil) model,

- three training steps of all models,
- fixing the sil model (adding extra transitions) and creating one state, short pause (sp) model,
- two training steps of all models,
- realigning the training data,
- ten training steps of all models.

At this stage there were 36 monophone models, one sil model (with three emitting states) and one sp model (one emitting state).
The second phase included context dependent models (triphones) preparation. They were derived from monophones as follows:

- creating triphone models that had appeared at least three times in the training data - at this stage there were 16443 models,
- two training steps of all models,
- creating the full triphone model set - it included all theoretical triphones - also those not seen in the training data. The number of models was 49286 at this stage.
- tying states of triphone models - the number of models was reduced to 1919 (called 'physical' or 'counting' models)
- two training steps of all models,
- increasing number of Gaussian mixture components for each state up to 15; it was done gradually and after each action two training steps were conducted.

Finally the system consisted of 1917 context dependent models with two without context, i.e. sil and sp models.
In order to reduce the number of triphone models states tying were performed. It was done using decision tree method. States representing following contexts (phonemes) were tied together (separately for left and right context): nasals (m,n,n)[1], fricatives (c, tS, tsj, dz, dzj, dZ, f, h, s, S, sj, v, z, zj, Z), vowels (a, oc5, e, eo5, i, o, u, i2) and stops (b, d, g, k, p, t). The rest contexts (l, w, r, j) were not tied. As the result phoneme contexts were reduced to 9 cases (including silence).

4.3. RECOGNITION PROCEDURE

Recognition process was made under strict grammar constrains. For each sentence there was a separate grammar constructed, e.g. for the Polish sentence "I one też czują oddziaływanie czarodziejskiego domku", the grammar file looked like:

```
$phoneme=a|d|s|j|y|u|k|e|g|a|t|e|z|s|cz|c|dz|b|z|w|r|z|c|l|h|sz|dz|f|i|n|dz|m|n|l|p|u|o|sil;
$sentence={$phoneme} <i> {$phoneme} <one> {$phoneme} <też> {$phoneme} <czują>
{$phoneme} <oddziaływanie> {$phoneme} <czarodziejskiego> {$phoneme} <domku>;
( sil $sentence sil )
```

The $phoneme variable represents phonemes and the $sentence variable represents recognition variants. Above constrains forces the recognition contains a sequence of words known in advance. Additionally at the beginning and between words one phoneme or more can appear. If such phoneme had appeared it was interpreted as a disfluency. For each recognized sentence there was a separate grammar file prepared.

[1]Authors use the custom symbols set to denote Polish phonemes. Below are mapping between that symbols and the International Phonetic Alphabet (IPA).

Phonemes	a	b	c	tS	tsj	d	dz	dzj	e	dZ	eo5	f	g	h	i	j	k	l	w	m	n	n~	o	oc5	p	r	s	S	sj	t	u	v	i2	z	zj	Z
IPA	a	b	t͡ɕ	t͡ʃ	t͡ɕ	d	d͡ʑ	d͡ʒ	ɛ	d͡ʒ	ɛ̃	f	g	x	i	j	c	l	w	m	n	ɲ	ɔ	ɔ̃	p	r	s	ʃ	ɕ	t	u	v	i	z	ʑ	ʒ

In order to obtain better performance models adaptation was done. For that purpose several recordings were used of summary duration of one minute. In the work supervised adaptation technique with maximum likelihood linear transformations was used.

5. RECOGNITION RESULTS

For the testing purpose the database consisted of 192 disturbed utterances was prepared. Each had the length of several seconds and included one or more disfluencies of different kind. The overall number of marked disfluencies was 384 and the total duration of all samples was 18 minutes.

Two parameters were used for evaluation purposes - sensitivity and predictability:

$$sensitivity = \frac{P}{N} \tag{1}$$

$$predictability = \frac{P}{P+F} \tag{2}$$

where: P - number of correctly detected disorders, N - total number of disorders, F - fragments detected as disorders by mistake.

Each tested utterance had a word level transcription with annotated disfluent fragments. For example, one of the sample had transcription "I one też czują NIEP oddziaływanie NIEP czarodziejskiego domku". Disfluencies are marked by the string 'NIEP'. The system recognized it as: "I one też czują o oddziaływanie cza czarodziejskiego domku". The two additional words 'o' and 'cza' were detected and interpreted as disfluencies.

Overall results are presented in the Table 1.

Table 1. Recognition results for all recordings.

	Total number	Correctly detected	Not detected	Incorrectly detected	sensitivity	predictability
Disfluencies	384	339	45	91	88%	79%

Obtained results could be yet better but a lot of recordings had poor quality (background noise, overloaded signal). Therefore results for good quality samples (coming from one person) are presented in the table 2.

Table 2. Recognition results of best quality recordings (28 utterances).

	Total number	Correctly detected	Not detected	Incorrectly detected	sensitivity	predictability
Disfluencies	55	55	0	11	100%	83%

Obtained results are showing that such approach can be very effective. Of course not only stuttered fragments are detected this way, but it can be very helpful to find suspicious parts of the utterance that could be further analyzed by other algorithms. The system gives a number of temporal information about the speech like, duration of phonemes, words, phrases and pauses, words and phoneme articulation rates, etc. what can be also utilized in a fluency evaluation process.

6. SUMMARY

The aim of the studies on the automatic diagnosis of stuttering people is to develop a complete, non-supervised and fully objective tool that can support speech diagnosis and therapy

process.

In the work there were speech recognition techniques utilized to find disfluent fragments. Obtained results are very good - sensitivity 88% and predictability 79%. Results, in the case of good quality recordings were even better - all the stuttered fragments were correctly identified whereas false recognitions were relatively small.

Such recognition concept can be very robust and in conjunctions with other methods can lead to building a very reliable system to speech fluency evaluation.

BIBLIOGRAPHY

[1] AWAD S. S., The application of digital speech processing to stuttering therapy, Instrumentation and Measurement Technology Conference, 1997. IMTC/97. Proceedings. 'Sensing, Processing, Networking'., IEEE, vol. 2, 1997, pp. 1361-1367.

[2] CODELLO I., KUNISZYK-JÓŹKOWIKAK W., SMOŁKA E., KOBUS A., Automatic disordered syllables repetition recognition in continuous speech using CWT and correlation, Proceedings of the 8th International Conference on Computer Recognition Systems CORES 2013, Springer International Publishing, pp. 867-876.

[3] CODELLO I., KUNISZYK-JÓŹKOWIKAK W., SMOŁKA E., KOBUS A., Automatic prolongation recognition in disordered speech using CWT and Kohonen network, Journal of Medical Informatics & Technologies, Vol. 2012, pp. 137-144.

[4] CODELLO I., KUNISZYK-JÓŹKOWIKAK W., SMOŁKA E., KOBUS A., Disordered sound repetition recognition in continuous speech using CWT and Kohonen network, Journal of Medical Informatics & Technologies, Vol. 17/2011, pp. 123-130.

[5] GAJECKI L., TADEUSIEWICZ R., Modeling of Polish Language for Large Vocabulary Computer Speech Recognition, Speech and Language Technology, Vol. 11, Poznań, 2008, pp. 65-70.

[6] http://htk.eng.cam.ac.uk/docs/docs.shtml.

[7] MAIER A., HADERLEIN T., EYSHOLDT U., ROSANOWSKI F., BATLINER A., SCHUSTER M., NOTH E., PEAKS - A system for the automatic evaluation of voice and speech disorders, Speech Communication 51, 2009, pp. 425-437.

[8] NOTH E., NIEMANN H., HADERLEIN T., DECHER M., EYSHOLDT U., ROSANOWSKI F., WITTENBERG T., Automatic stuttering recognition using Hidden Markov Models, Proc. Int. Conf. on Spoken Language Processing, vol. 4, Bejing, China, 2000, pp 65-68.

[9] SCHULTZ T., GlobalPhone: A Multilingual Speech and Text Database developed at Karlsruhe University, In: Proc. ICSLP Denver, CO, 2002.

[10] SZCZUROWSKA I., KUNISZYK-JÓŹKOWIKAK W., SMOŁKA E., The Application of Kohonen and Multilayer Perceptron Networks in the Speech Nonfluecy Analysis, Archives of Acoustics, vol. 31, 2006, pp 205.

[11] ŚWIETLICKA I. , KUNISZYK-JÓŹKOWIKAK W., SMOŁKA E., Artificial Neural Networks in the Disabled Speech Analysis, in Computer Recognition System 3. vol. 57/2009, Springer Berlin / Heidelberg, May 12, 2009, pp. 347-354.

[12] WIŚNIEWSKI M., KUNISZYK-JÓŹKOWIKAK W., Automatic detection and classification of phoneme repetitions using HTK toolkit, Journal of Medical Informatics & Technologies, Vol. 17/2011, Computer Systems Dep., University of Silesia, Poland, 2011, pp. 143-148.