automatic speech recognition,
hospital information systems,
language models

Bogumiła HNATKOWSKA[**], Jerzy SAS[*]

APPLICATION OF AUTOMATIC SPEECH RECOGNITION TO MEDICAL REPORTS SPOKEN IN POLISH

The paper presents an attempt to automatic speech recognition of Polish spoken medical texts. The attempt resulted in experimental system that can be used as a tool for practical applications. The system uses a typical recognition method based on Hidden Markov Model and domain-specific language model. Implemented software made it possible to conduct many experiments aimed on evaluation of the assumed approach usefulness. Obtained experiment results are presented and analyzed. The system architecture and the way in which it can be integrated with hospital information systems is also exposed.

1. INTRODUCTION

Automatic speech recognition (ASR) is a very convenient technique of data entering to information systems. The role of this technique is especially important in the case of medical information systems, where ASR helps to enter typical elements of medical records expressed in natural language ([4]). ASR can be applied to transcribe voice notes on disease analysis or observations but the most frequent application seems to be in diagnostic image interpretation and reporting, where the physician can freely dictate the report by voice while observing the image and controlling the image viewer with his hands.

There are a few commercial modules for medical speech recognition, which can be integrated with hospital information systems (HIS) or radiology information systems (RIS). The available products support many languages as English, French, German and Japanese. Unfortunately, to our best knowledge none of the systems supports Polish. Although the methods and techniques used in speech recognition are language independent, they require the language model prepared individually for each language. Polish language is exceptionally difficult in automatic speech recognition due to its highly inflectional nature and great number of words forms. It leads to significant difficulties in building the language model for Polish and in result the lack of ASR tools for this language. The works related to ASR in Polish reported in literature concern mainly recognition of isolated word commands or simple utterances used in telephony or in software or device control ([2]). Grammar based approaches have been also considered ([1]). There are however very few reports related to ASR applied to natural or almost-natural Polish language. By almost-natural language we mean here the language with relatively limited dictionary (of the order of 10,000 words) where any word sequence can be uttered, but some typical phrases appear very frequently. For this kind of applications statistical language model based on n-grams ([3]) seems to be more appropriate than grammar based approach.

In this article the preliminary results of ASR application to diagnostic imaging spoken reports transcription to textual form are described. The domain-dependent approach is applied, which makes the recognition sufficiently accurate, even with imprecise language model built with relatively small amount of texts in the domain-specific area. The general system architecture of medical speech ASR system focused on integration with HIS is presented. The experimental system is based on HTK package ([6]) widely used in ASR research.

The aim of works and experiments described in this article was to evaluate the practical usefulness of ASR based on Hidden Markov Models (HMM) and domain-specific language model applied to recognition of domain-specific Polish spoken medical reports. In order to make the solution applicable in clinical practice the use of inexpensive audio equipment (microphone, sound card) and typical desktop computers was assumed. In particular, we wanted to find out answers to the following questions:

- Is ASR for Polish feasible at the sufficient accuracy level with the domain specific statistical language models built using small corpora of texts available in HIS databases?
- Can ASR be achieved in real time on typical inexpensive PCs provided that the dictionary covers high fraction of words appearing in utterances typical for the domain?
- What level of accuracy can be achieved in practice?
- What is duration of training utterances necessary to achieve reasonable ASR accuracy?

The paper is organized as follows. In section 2 applied method for automatic speech recognition is briefly described. In section 3 the architecture of ASR system and the way of its integration with HIS is presented. Section 4 summarizes obtained experimental results. Section 5 concludes the paper with final remarks.

[**] Institute of Applied Informatics, Wroclaw University of Technology, Wyb. Wyspianskiego 27, 50-370 Wroclaw, POLAND, e-mail: bogumila.hnatkowska@pwr.wroc.pl

[*] Institute of Applied Informatics, Wroclaw University of Technology, Wyb. Wyspianskiego 27, 50-370 Wroclaw, POLAND, e-mail: jerzy.sas@pwr.wroc.pl

2. APPLIED ASR METHOD

In our approach we followed domain-dependent and speaker-dependent approach. It is assumed that the domains of spoken notes as well as the speaker identity are explicitly given. The domain of the note is determined by the kind of information contained in it. The examples of note domains in case of medical diagnostic imaging can be related to modalities (e.g. tomography, computed radiography, USG etc.) or can be determined by anatomic partitioning of body or by the group of diseases being diagnosed. For each domain, individual language model is created using the corpus of texts from the domain.

Typical ASR method based on HMM was applied in the experimental system ([3], [6], [7]). Three-level hierarchical approach was used ([3]). On the lowest level, simple Markov models are created and trained for individual Polish phonemes. Uniform HMM topology for each phoneme is assumed. It consists of five states including initial and terminal ones.

The state transition probabilities as well as the parameters of observations emission probability density functions for all phoneme HMMs are estimated using Baum-Welch procedure and the set of correctly transcribed training utterances. The recognized utterances consist of words coming from finite dictionary $D = \{w_1, w_2, ...w_N\}$. For each admissible word its phonetic translation is created and word HMM is built by concatenating HMMs for subsequent phonemes. The phonematic transcription is achieved using translation rules set for Polish language described in [5]. The original rules set appeared to be incomplete, so it was complemented by a number of new rules. Because, despite of this extension, some orthographical symbols in certain contexts were still not translated it was assumed that for each orthographical symbol the default phoneme is defined. If the symbol appearing in certain context cannot be transcribed using applied rules set, it is transcribed as its default phoneme. Experiments with words appearing in explored medical text corpora proved that the transcription obtained in this way is correct in the case of about 99.8% of words.

Finally, the compound HMM of the whole utterance is built by connecting word HMMs. The probabilities of transition from the terminal state of a word HMM to initial state of another word HMM are taken from domain-specific bi-gram language model. Bi-gram language model consists of the set of conditional probabilities $p(w_i|w_j)$ of word w_i appearance provided that the preceding word was w_j. In order to uniformly handle the words beginning (ending) utterances, the artificial "blank" word representing beginning (ending) of the utterance is added to the dictionary. The conditional probabilities in the language model are estimated using the domain specific corpora extracted from HIS database. The structure of the HMM used to recognize complete utterances is presented on Fig. 1.

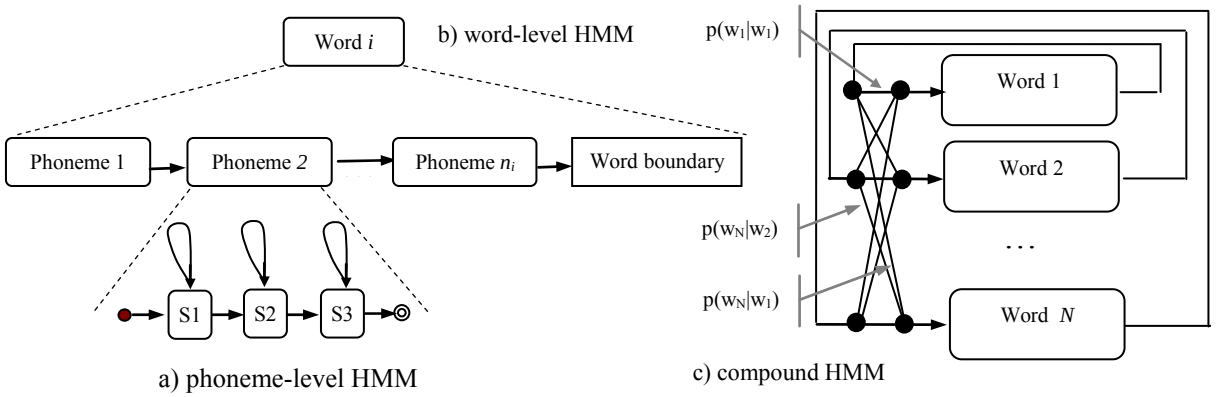

Fig. 1. The structure of HMM used in the isolated utterance recognition

The recognition with HMM consists in finding the most probable path between states of HMM for the observed sequence of feature vectors (observations) extracted from the acoustic signal recorded while speaking the utterance. Let (o_1, o_2, ..., o_t) be the sequence of observations for the utterance being recognized. Let the compound HMM consists of M states {s_1, s_2, ...s_M} and for each state the probability density function determining the probability distribution of observations emitted in this state is defined. Let S_W denotes the subset of states $\{s_{w_1}, s_{w_2}, ..., s_{w_N}\}$ being the terminal states of HMMs created for words in the dictionary. The recognition with HMM consists in finding such word sequence W^* which maximizes its conditional probability given the observed sequence:

$$W^* = \underset{w_{i_1} w_{i_2}, ..., w_{i_k} \in D^+}{\arg\max} P(w_{i_1} w_{i_2}, ..., w_{i_k} \mid o_1, o_2, ..., o_t) \qquad (1)$$

where D^+ denotes the set of all nonempty sequences of words from the dictionary D.

Because words are represented by separated HMMs in the compound model, the optimization problem given by (1) is equivalent to finding such state sequence $S^* = \{s_{i_1}^*, s_{i_2}^*, ..., s_{i_t}^*\}$ in compound HMM which ends in state $s_{i_t} \in S_W$ and which maximizes its conditional probability:

$$S^* = \underset{\substack{s_{i_1} s_{i_2},...,s_{i_t} \in (S_{HMM})^t \\ s_{i_t} \in S_W}}{\arg\max} P(s_{i_1} s_{i_2},...,s_{i_t} \mid o_1, o_2,...,o_t) \quad (2)$$

where $(S_{HMM})^t$ denotes the set of all t-element sequences of states of the recognizing HMM. By applying the formula for conditional probability:

$$P(s_{i_1} s_{i_2},...,s_{i_t} \mid o_1, o_2,...,o_t) = P((s_{i_1} s_{i_2},...,s_{i_t}) \wedge (o_1, o_2,...,o_t)) / P((o_1, o_2,...,o_t)) \quad (3)$$

and by rejecting the term in the denominator, which is constant for all state sequences, the optimization problem in (2) can be further simplified:

$$S^* = \underset{\substack{s_{i_1} s_{i_2},...,s_{i_t} \in (S_{HMM})^t \\ s_{i_t} \in S_W}}{\arg\max} P((s_{i_1} s_{i_2},...,s_{i_t}) \wedge (o_1, o_2,...,o_t)) \quad (4)$$

The optimization task (4) – and in result also the complete speech recognition problem – can be efficiently solved with widely used Viterbi algorithm ([3]).

The speech recognition algorithm described above can be applied to the sequences of observations corresponding to isolated complete utterances (sentences). The segmentation of formally infinite stream of observations extracted from continuous stream of audio samples provided by the acoustic channel is achieved as a kind of audio pre-processing. The sentence boundaries are detected by finding the silence intervals of the length exceeding the minimal duration established experimentally.

3. ASR SYSTEM ARCHITECTURE FOR INTEGRATION WITH HIS

The components of speech recognition system support two main processes: training and recognizing. The training process consists of the following stages: (a) acquiring the raw text corpora for defined speech domains, (b) training text purification, (c) building domain-specific dictionaries, (d) complementing dictionary words with its phonematic transcription, (e) building domain-specific bi-gram language model, (f) selection of utterances for phoneme HMMs training, (g) training phoneme HMMs, and (h) building compound HMMs for all domains. The recognition is a much simpler process. It consists of: (a) segmentation of acoustic sample stream into fragments corresponding to sentences, (b) observation sequence extraction, and (c) recognition with HMM.

In the case of ASR application in hospital information systems, it is assumed that text domains are defined in HIS and texts used as corpora for domain-specific model building are contained in HIS database. For the sake of usage convenience, the corpora is extracted from HIS database automatically by a process that periodically queries HIS database and updates all domain-specific language models. The speech recognition module can work in on-line or off-line mode. In on line mode the recognizer transcribes the audio stream in real time and sends the text transcription directly to the connected editing control in HIS user interface. In off-line mode the recognition module performs batch processing of voice files stored in given location. The voice files can be downloaded from portable audio recording devices used by physicians when dictating voice notes. The architecture of ASR subsystem integrated with HIS is shown on Fig. 2.

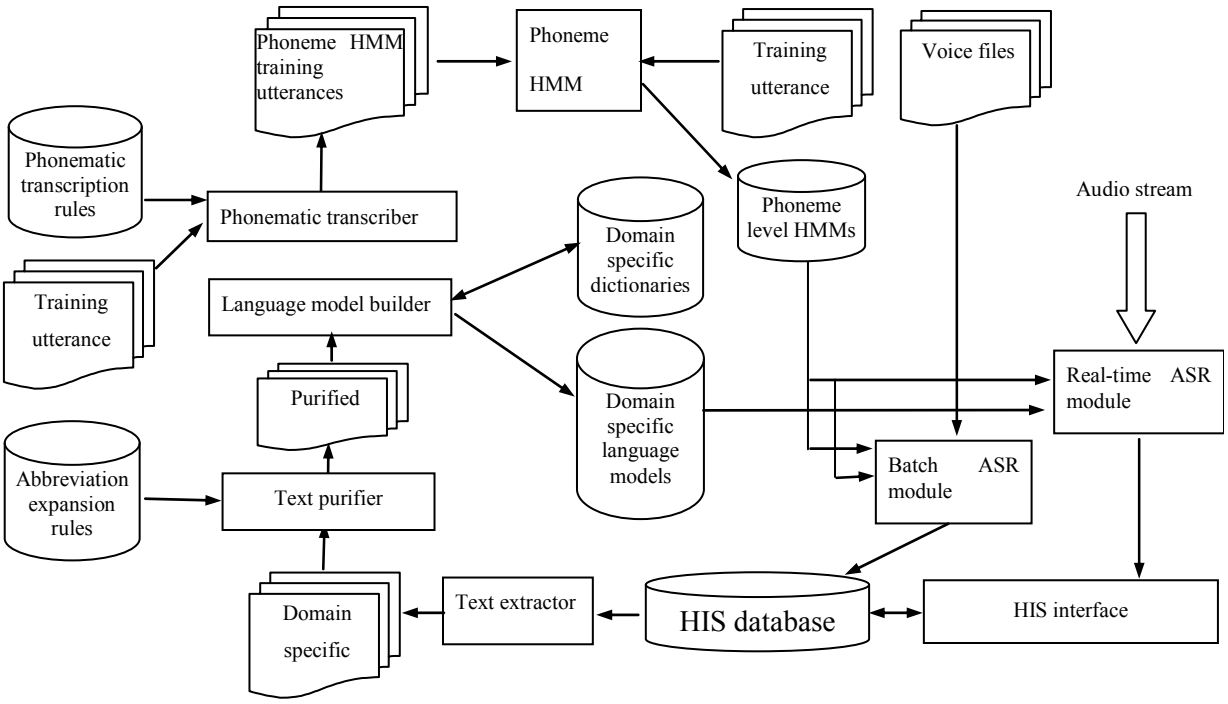

Fig. 2. The architecture of ASR module focused on integration with hospital information system

4. EXPERIMENTS

4.1. ESTIMATION OF DICTIONARY SIZE AND ITS INFLUENCE ON ASR ACCURACY

ASR application to medical spoken notes recognition has a number of specific features making the recognition problem easier that in the case of general spoken natural language. The spoken notes belong to a known domain in which the set of used words and their forms is relatively small. Additionally, there exist typical phrases frequently used in utterances. The typical examples in the area of diagnostic imaging are phrases denoting no changes or deviations or describing frequently appearing pathological changes. Taking these advantages into account, the restricted domain-specific language model can be constructed and used for further ASR for notes related to the specific domain.

The conducted experiments with HMM based speech recognizer revealed that speech recognition time increases significantly when the dictionary grows. The Viterbi algorithm analysis shows that the recognition time is approximately proportional to the square of HMM compound states number, which in turn grows proportionally to the dictionary size. For practical success of ASR application, the recognition speed may be crucial. It is because fast speech recognition makes it able to run in real-time mode and to allow immediate manual correction of recognition errors. For this reason it is desirable to reduce the dictionary size to the limits ensuring acceptable recognition time. Restricting the dictionary leads to degradation of recognition accuracy, because out-of-dictionary words will never be recognized correctly. If the probability that the word is out-of-dictionary is p_e then the word recognition correctness will never be higher than $1 - p_e$. The appropriate balance must be then kept between the recognition accuracy and speed when determining the domain-specific dictionary size.

In order to evaluate the size of necessary domain-specific dictionary we examined five domains of diagnostic imaging reports related to: computed radiography, tomography, mammography, endoscopy and USG examinations. The corpus of authentic written reports stored in RIS (a subsystem of HIS) was explored for each domain. Domain-specific dictionaries consisting of all used words were collected and examined individually. For each word w_i its relative frequency f_i was calculated as:

$$f_i = \frac{c_i}{\sum_{j=1}^{N} c_j}, \tag{5}$$

where c_i is the number of appearances of the word w_i in the corpus. The words were then sorted by their relative frequencies in decreasing order giving the sequence $F = ((w_{i_1}, f_{i_1}),(w_{i_2}, f_{i_2}),...,(w_{i_N}, f_{i_N}))$. For any relative frequency level $r \in <0,1>$ such shortest leading subsequence of $n(r)$ words $w_{i_1},...w_{i_{n(r)}}$ in F can be found that:

$$\sum_{j=1}^{n(r)} f_{i_j} \geq r \qquad (6)$$

It means that selecting only $n(r)$ most frequent words from the dictionary gives the subdictionary ensuring the corpus coverage factor r. In other words, restricting the dictionary to $n(r)$ words results in the probability that the word encountered in the utterance is out-of-dictionary is not greater than $1\text{-}r$. By selecting appropriate value of r and by reducing the dictionary to $n(r)$ words the appropriate balance between speech recognition accuracy and the efficiency of the speech recognizer can be achieved.

The dictionary size $n(r)$ for various values of r and for examined domains of medical texts is shown in Table 1.

Table 1. Dictionary size for various coverage factors and various text categories

Text domain	Total number of words	Total number of word appearances	Reduced dictionary size $n(r)$ for the coverage factor r				
			$r = 0.99$	$r = 0.98$	$r = 0.97$	$r = 0.96$	$r = 0.95$
Computed radiography I	7,756	841,063	5,327	2,181	1,588	1,240	1,041
Computed radiography II[1]	24,309	1,203,055	10,684	6,243	4,218	2,735	2,283
Computer tomography	12,643	551,468	7 ,229	5,186	3,773	2,808	2,176
Mammography	6,253	952,581	1,517	1,177	983	880	704
Endoscopy	23,424	1,046,111	10,194	5,846	4,431	3,116	2,157
USG	14,317	1,598,026	5,636	2,935	1,760	1,268	1003

The experiments described in section 4.3 indicate that if beam search technique ([3], [8]) is applied then the recognition time for the dictionary of the size about 8000 is comparable to the speaking time. For dictionaries of significantly larger size the recognition time is also significantly higher, what makes the method inefficient for real-time recognition. The results presented in Table 1 indicate that for all considered categories of medical texts, the dictionary used in ASR can be reduced to the relatively small size of the order below ten thousands words, while still preserving the coverage factor at the level 0.99. It means that if the dictionary size is limited then speech recognition allows for the recognition in real time.

4.2. DEPENDENCE OF ASR ACCURACY ON TRAINING TIME

Another factor important for practical usability of ASR is the amount of efforts related to training a recognizer by an individual speaker. Long training time can seriously discourage users and made ASR impractical. In the next experiment we examined how the recognition accuracy depends on the training time. Computed radiography I (characterized in Table 1) corpus was used. The set of 745 short utterances were randomly selected from the corpus. Sample utterances were then pronounced and recorded by two speakers: professional lector and amateur speaker. The lector utterances were recorded in high quality with the audio equipment of more than 90 dB signal/noise ratio without audible background noises. The amateur recordings were performed in low quality using inexpensive headset microphone and Sound Blaster Connect external audio card, where signal/noise ration was about 60dB.

The single iteration of the experiment consisted in assessing the average recognition accuracy for fixed summed duration of training utterances. The subset of recorded utterances was randomly selected so as to obtain the total duration of training samples close to the duration assumed. Remaining recorded utterances were used as testing set. To reduce the randomness of obtained assessment the elementary experiment consisting of evaluation of ASR accuracy for randomly selected training set was repeated several times for given training duration. The experiment was performed for training duration in the interval <1,30> minutes.

The recognition accuracy was defined on word level. The accuracy was calculated according to the formula used by HTK `HResult` tool:

$$Acc = \frac{N - I - S - D}{N}, \qquad (7)$$

[1] Two corpora for computer radiology from two different diagnostic centers were examined

where *I*, *S*, *D* and *N* are numbers inserted, substituted, deleted, and actual words correspondingly. The dependence of the recognition accuracy on the training set total duration for both speakers is shown on Fig. 3. It can be observed that reasonable accuracy can be achieved even for very short training time of the order of several minutes. Practically no further observable improvements of the recognition quality can be achieved by extending the training time above 20 minutes.

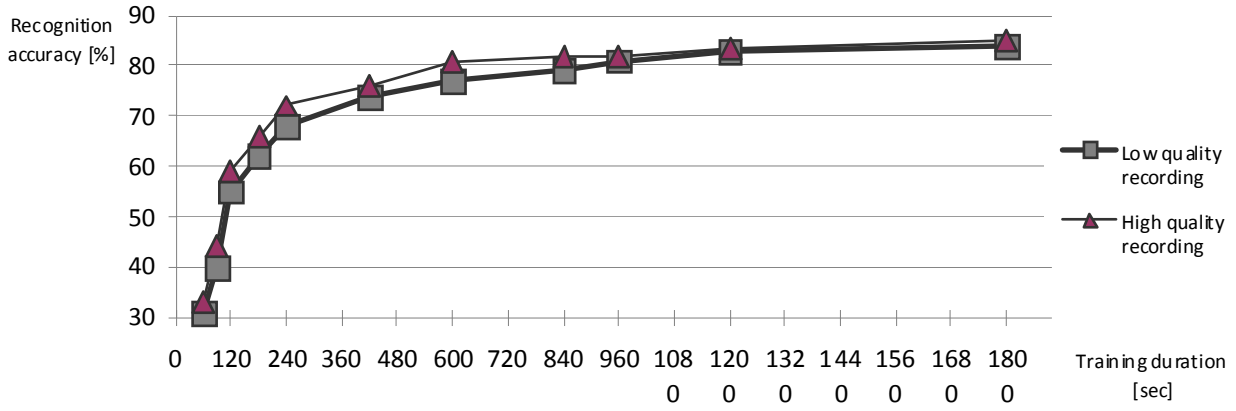

Fig. 3. Speech recognition correctness as a function of training time

4.3. BALANCING ASR SPEED AND ACCURACY

Experiments described in section 4.1. demonstrated high repeatability of utterances from domain-specific corpora of medical texts. In result, domain-specific dictionaries for all tested domains of medical texts can be reduced to less than 10,000 words by devoting only 1% of ASR recognition accuracy. Unfortunately, experiments conducted on low-cost desktop computers (equipped with 2.5 GHz Core 2 Duo Pentium processor and with 2GB RAM) exhibited that basic Viterbi algorithm used as a recognition method is not able to work in real time with dictionaries of that size.

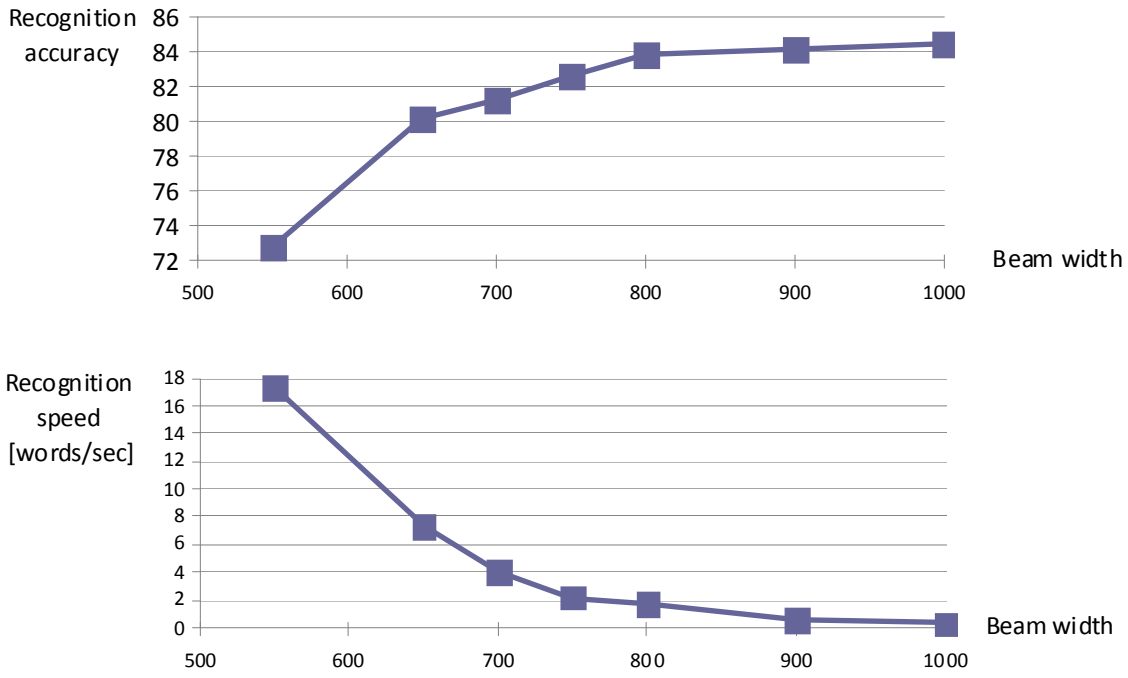

Fig. 4. ASR time and accuracy as a function of beam width

The best path search done by Viterbi algorithm can be accelerated if in each step of the forward phase of Viterbi procedure only the subset of the most promising subpaths in the HMM graph will be considered. It leads to well known "beam searc h" algorithm ([8]). The beam width in beam search is the interval for probabilities of subpaths in the compound HMM graph. If in the forward phase of Viterbi procedure the difference between the partial probability of a path and the probability

of the best path is greater than the beam width then the path is excluded from further considerations. In HTK implementation of beam search, the beam width is specified as log probability, so its typical range is approximately (100,1000). Reducing the beam width (i.e. the number of subpaths being considered) decreases the recognition time, but also lowers the recognition accuracy. The experiment has been conducted in order to evaluate the impact of reduced beam width on the ASR time and accuracy. The same set of recorded utterances as described in section 4.2. was used. Training subsets of the total duration about 20 minutes were randomly selected. For each randomly selected set, the remainder of available 745 recorded utterances were used as testing set. Every testing set was recognized multiply using various beam widths. The results of experiments obtained for various training/testing sets were finally averaged. The dependency between recognition accuracy and speed on the beam width is depicted on Fig. 4.

By appropriately setting the beam width very significant reduction of ASR time can be obtained at the expense of only minimal degradation of ASR accuracy. The optimal beam width seems to be about 750. The recognition time obtained with such beam width is almost five times shorter in relation to the recognition time of basic Viterbi algorithm. The recognition rate is about 2 words/sec, what is close to typical speaking rate.

5. CONCLUSIONS

The experiment results proved the thesis that ASR for medical texts spoken in Polish is feasible. Moreover, the sufficient accuracy level of recognition was obtained with domain specific statistical language models built with small corpora of texts that can be automatically acquired from HIS databases. The accuracy level (about 84%) could be further improved at syntax level by automatic text correction. ASR can be applied on non-expensive, commonly used hardware platforms. The transcription time is comparable to the speaking time assuming that beam search is used and the dictionary is reduced to 8000 words. The time of utterance training necessary to train a speech recognizer is acceptable (about 20 minutes) and analogous to other commercial tools. The software developed primarily for the sake of experiments described in the article can be used as a tool for practical applications for ASR in medical information systems.

BIBLIOGRAPHY

[1] KORŽINEK D., BROCKI L., Grammar Based Automatic Speech Recognition System for the Polish Language. In: Jabłoński R., Turkowski M., Szewczyk R. [eds.], Recent Advances in Mechatronics, pp. 87-91, Springer, Berlin Heidelberg, 2007.

[2] KACALAK W, MAJEWSKI W., Automatic Recognition of Voice Commands in Natural Language Given by the Operator of the Technological Device Using Artificial Neural Network, In: Kurzynski M., Puchala E., Wozniak M., Zolnierek A., Proc. of 4th Int. Conf. on Computer Recognition Systems, CORES, 05, pp. 689-696, Springer Verlag, 2005.

[3] JELINEK F., Statistical Methods for Speech recognition, The MIT Press, Cambridge, Massachusetts, 1997.

[4] GRASSO M.A., The Long-Term Adoption of Speech Recognition in Medical Applications Proc of 16th IEEE Symposium on Computer-Based Medical Systems (CBMS'03), pp. 257-262, IEEE Press, 2003.

[5] STEFFEN-BATOGOWA M., Automatization of Polish Texts Phonematic Translation (in Polish), PWN, Warsaw, 1975.

[6] YOUNG S., EVERMAN G., et al., The HTK Book, Cambridge University Engineering Department, 2005.

[7] JURAFSKY D., MARTIN J., Speech and Language Processing. An Introduction tp Natural Language Processing, Computational Linguistics and Speech Recognition, Prentice Hall, Upper Saddle-River, New Jersey, 2000.

[8] ANTONIOL G., BRUGNARA F et al., Language Model Representations for Beam-Search Decoding, Proc. of Int. Conf. on Acoustics, Speech, and Signal Processing, ICASSP-95., pp. 588—591, Detroit, 1995.