DEVELOPMENT OF LARGE VOCABULARY CONTINUOUS SPEECH RECOGNITION USING PHONETICALLY STRUCTURED SPEECH CORPUS

G. Demenko[a], *M. Szymański*[a], *R. Cecko*[a], *M. Lange*[a], *K. Klessa*[b] & *M. Owsianny*[a]

[a]Poznań Supercomputing and Networking Center, Polish Academy of Sciences, Poznań, Poland;
[b]The Institute of Linguistics, Adam Mickiewicz University, Poznań, Poland
grazyna.demenko@speechlabs.pl; marcin.szymanski@speechlabs.pl; marek.lange@speechlabs.pl; mariusz.owsianny@speechlabs.pl; cecko@man.poznan.pl; klessa@amu.edu.pl

ABSTRACT

This paper presents the results of acoustic modeling used in a Large Vocabulary Continuous Speech Recognition (LVCSR) system designed with the use of a phonetically controlled large vocabulary corpus. Evaluation experiments showed that relatively good speech recognition results may be obtained with adequate training material, taking into account: a) the presence of lexical stress; b) speech styles (a variety of segmental and prosodic structures, various degree of spontaneity of speech, various pronunciation variants and dialects); c) the influence of the sound level and environment noise. Moreover, the article includes information about the speech corpus structure and also an outline of the design of the speech recognition system.

Keywords: segmental and suprasegmental phonetics, very large speech corpora, acoustic models, speech recognition

1. INTRODUCTION

A review of the results of automatic speech recognition (ASR) systems built for various languages shows that while creating such a system for highly inflectional languages like Polish, additionally characterized by a comparably flexible word order, certain assumptions concerning the acoustic-phonetic database structure need to be modified (as compared to e.g. English) in order to provide adequate material for both acoustic and language modeling (cf. [2]).

Due to the fact that the phonetic-acoustic structure of Polish is comparably well specified, we expect that a speech corpus characterized by an adequate phonetic structure (triphone and prosodic coverage) and acoustic structure will ensure obtaining representative acoustic models and high recognition results based on the acoustic models.

Acoustic models for ASR need to be based on large corpora, involving many speakers selected to represent a typical distribution of age, sex and geographic area so that they represent an average for a particular language, e.g. according to Moore [5], a 1000-hour database allows for building a system with a word error rate of ca. 12% when language modeling is applied, and over 30% word error rate with no language modeling. He also estimates that at least 100,000 hours of speech is needed to train an ASR system with an accuracy comparable to that of a human listener.

Lexical stress patterns annotated in various pronunciation dictionaries are expected to be effective indicators of pitch accents in speech [1]. These observation should be used to augment a standard ASR model to improve recognition performance. In particular, it would be of high importance for languages with fixed place of lexical accent (in Polish lexical stress falls on the penultimate syllable). It was shown in [10] that the inclusion of stressed vowel models for Polish ASR yields approximately 6% reduction of word-error rate: an inventory of 39 Polish phones was used, and as an addition six units representing stressed vowels (as opposed to their unstressed equivalents) were included. The latter modification was made on dictionary level only, i.e., no acoustical analysis of stress is performed either on the training set or during the recognition.

In this paper, we report on the evaluation of the influence of the speaker gender, speaking style and recording sound level on the recognition accuracy obtained with various acoustic models. The structure of the remaining parts of the paper is as follows: Section 2 is a brief introduction to the speech database and training data, in Section 3 the experimental results are reported, Section 4 describes the general design of the present LVCSR System and summary of evaluation data, Section 5 includes a short discussion and conclusions.

2. SPEECH DATABASE & TRAINING DATA

The study material was selected from the Jurisdict database designed specifically for the present ASR system whose target end-users are judges, lawyers, policemen and other public officers. The database contains recordings of speech delivered in quiet office environments by over 2000 speakers (a total of over 1155 hours of speech) from 16 regions of Poland. Most of the speech material (855 hours) was recorded in a two-channel mode using two types of microphones (a close-talk microphone and a table microphone). The remaining recordings are one-channel, a part of them have been recorded in the environment of higher noise level (large courtroom, noisy office). All data were annotated manually according to SpeeCon guidelines [3] by a team of trained labelers. The SpeeCon guidelines assume orthographic, word-level transcription with only several non-speech-event markers for speaker and background noises. For purposes of acoustic modeling, the files were then subject to automatic, phone-level segmentation using Salian [9]. The structure of the database is based on three major sub-corpora described below (more details in [4]).

Police & Office sub-corpus - complex recording scenarios composed of read and semi-spontaneous speech, approximately 350 utterances (sentences and phrases) and up to 0,5 hour of speech per speaker. Semi-spontaneous speech: elicited dictation on various topics (everyday life, professional life). Read speech: Syntactically controlled sentences (variable concatenation of phrases, variable phrase length; phonetically controlled utterances (triphones, special lexical phrases, bigrams, modulants, greetings, commands; application-specific texts and phrases).

Lawyer sub-corpus - recording scenarios of 80-100 utterances each, approximately 20 minutes of read speech per speaker (the text materials acquired from original legal texts).

Court sub-corpus - original recordings from court trials, 33 speakers, various duration of recordings per speaker, up to a total of 15 hours, spontaneous speech.

For the purposes of acoustic modeling experiments over 568 hours of speech produced by 1488 speakers were selected from the Police & Office and Lawyer sub-corpora of the database. So far, only close-talk microphone channel recordings from the two-channel data set have been used (characterized by lower background noise level and higher speaker noise).

3. ACOUSTIC MODELS

3.1. Training tools

The acoustic speech models were trained using HTK [11]. The standard training procedure for triphone Continuous Density Hidden Markov Model was generally used, consisting of running the training tools offered by HTK, namely: HInit, HRest, HERest and HHEd. A list of approximately 60 contextual questions formulated on the basis of phoneme articulation features served for state/triphone clustering.

3.2. Gaussian mixtures experiment

The subject of the experiment was to investigate the dependency of the accuracy and speed of speech recognition on the number of Gaussian mixtures in each state. For each tested acoustic model three set-ups for the number of mixtures were used: 24, 8 and 4 mixtures (24, 8 and 4 are average figures, the actual number per state depended on the number of training frames). The test set contained 147 utterances produced by 20 speakers. The test utterances were recorded by speakers themselves without supervision.

Table 1: Acoustic modeling results for different number of mixtures (% acc – mean percentage of correctly recognized minus inserted words, std dev. – standard deviation across speakers, % r. time – recognition time percentage where 100% is the real recognition time).

	4 mix	8 mix	24 mix
% acc (std dev.)	68,4 (13.3)	70.3 (12.9)	72.9 (12.1)
% r. time (std dev.)	201 (69.8)	248 (76.8)	638 (198.1)

Table 1 presents the word level recognition rates and recognition speed rate obtained with different number of mixtures. The best recognition rate was acquired for the 24-mixture model. The time figures suggest a significant trade-off between recognition rate and the required processing.

3.3. Speaker's gender experiment

For the need of the experiment the recordings of 646 male voices (M) and 646 female voices (F) were selected from the speech corpus. Then, two additional reference sets were obtained: one (F+M) was created by merging M and F. In order to preserve training corpus size equal to sex-specific models, another set (FM2) was prepared by randomly selecting half of the recordings from each F and M. As it can be presumed based on the results shown in Table 2, the models created using recordings of females perform better for female

voices, and analogously, "male" models are better with male voices. The results of the mixed FM2 model is slightly worse in comparison.

Table 2: Acoustic modeling results for gender depended models (for explanation of abbreviations cf. caption of Table 1).

	test/model	F	M	FM2
% acc (std dev.)	F	65.5 (6.6)	-	62.1 (7.1)
	M	-	63.8 (8)	60.6 (7.7)
	F+M	-	-	61.2 (7.4)

3.4. Speaking style experiment

The aim of the experiment was to check whether using the (semi)spontaneous part of the speech corpus for acoustic modeling could cause any change in dictated speech recognition. For this part of the study a sub-corpus containing over 405 hours of speech produced by 1488 speakers was used. For testing, the test set from the Gaussian mixtures experiment was used (cf. par. 3.2. above). The resulting figures for accuracy and recognition time were better for a combined model, i.e. when both read and (semi)spontaneous speech recordings were used (69.2 % of correctly recognized words for read speech, 70.3 % for the combined model). However, these differences are not statistically significant.

3.5. Sound level experiment

During preliminary system testing, it was realized that the training corpus is apparently insufficiently varied wrt. recording level (and more generally, quality). Hence, the subject of this experiment was to test if it is possible to increase the recognition rate of low audio level recordings by artificially reducing peak level in the training set utterances. The training recordings were preprocessed in order to achieve uniformly distributed peak levels between values 0dB and -13dB.

Table 3 presents the word level recognition rate of original and preprocessed train set models. The test set used in the experiment covered 147 utterances from 22 speakers. The results obtained suggest that the model trained on more level-varied training set performs not significantly worse on well adjusted volume range recordings, compared to the model trained on original files. At the same time, the preprocessed model yields better recognition rates for testing recordings with volumes lowered by 6, 12 or 18 dB. Additional experiments are required, however, to determine the extent to which these phenomena are caused by low recording levels themselves, as opposed to possibly inadequate level-insensitivity of the signal parametrization, as the latter is still under tuning.

Table 3: Acoustic modeling results for different training data sets (for explanation of abbreviations cf. caption of Table 1).

	test file volume	model on preprocessed	model on original data
% acc (std dev.)	original	69 (12,9)	69,5 (12,9)
% r. time		257	241
% acc (std dev.)	-6 dB	67,5 (13,5)	64,6 (14,4)
% r. time		261	269
% acc (std dev.)	-12 dB	60,4 (15,4)	44,5 (20,1)
% r. time		308	373
% acc (std dev.)	-18 dB	39,1 (19,9)	11,7 (13)
% r. time		389	348

4. ASR SYSTEM DESIGN & EVALUATION

The present LVCSR system for Polish was developed based on Microsoft .NET Framework 4.0 platform with the intense use of Task Parallel Library (TPL). The system can work in offline mode (the speech signal is taken from a file), or in online mode (the speech signal is taken directly from an audio device).

Figure 1: The architecture of the LVCSR system.

The PCM audio signal is passed to DSP (Digital Signal Processing) module (Fig. 1). The DSP analyzes audio data, and performs Voice Activity Detection. The signal is divided into separate observations (25 ms window with 10 ms stepping). For each observation an LDA (Linear Discriminant Analysis) transformation is computed over Mel-frequency cepstral coefficients and Filterbank parameters, giving a short feature vector as a result. The observation vectors are passed to a recognition based decoder. The decoder is build upon modified Viterbi algorithm [6] and works over a recognition network being a word-loop of ca. 320-thousand dictionary entries with imposed unigram probabilities. The decoder produces a hypotheses Lattice as a result. The Lattice elements are attributed with appropriate probabilities. All hypotheses are evaluated using N-Grams linguistic model in the Rescoring module. Hypothesis with best probability is returned as a recognized text.

The evaluation tests have been carried out using the Sclite tool [7], with recordings from 97 speakers (7749 sentences and 157 426 words). The analysis of errors both in word and sentence recognition showed that the highest percentage of errors is connected with word substitution (20.5%).

The percent of errors caused by deletions and insertions was 9.5% and 5.0%, respectively. Hence, the word accuracy was 65.0% and the correctness 70%. For Polish, this fact is of high importance, due to the variability of inflectional word endings and the resulting ambiguities.

Preliminary tests of the system were carried out to investigate the influence of the use a) language model (LM), and b) speaker adaptation.

Table 4: The influence of a language model in use (for explanation of abbreviations cf. caption of Table 1).

		Unigram LM	Trigr. LM
word	% acc (std dev)	68.3 (12.1)	76.3 (10)
sentence	% acc (std dev)	7.0 (6.9)	13.6 (10)
	% r. time	204.4	211.6

Table 5: Adaptation results for 3 speakers (for explanation of abbreviations cf. caption of Table 1).

		without adapt.	with MLLR
word	% acc (std dev)	81.1 (1.4)	83.8 (2.5)
sentence	% acc (std dev)	41.6 (2.8)	44.6 (3.5)
	% r. time	159.0	106.9

Table 4 presents the test results showing the influence of the applied language model on the recognition accuracy and recognition time given in %. The current language model is a word 3-gram built with SRILM toolkit [8]. It was estimated on over 4GB of automatically normalized text, mainly from legal domain (judgments, law acts, briefs, contracts etc.) plus some newspaper articles.

Table 5 shows the influence of 3 speakers' adaptation on the recognition accuracy. In the tests the significant relation between recognition time and the utterance quality has been observed (the better the speaker the shorter the recognition time). The speaker voice adaption shortens also the recognition time. In both cases the 8-mixture Gaussian acoustic model was used.

5. DISCUSSION & CONCLUSIONS

The evaluation results suggest that we are already close to the expected system accuracy [5] when no language modeling was implemented; however, the quality and impact of the 3-gram language model is still not satisfying.

The statistical insignificance of the differences in recognition of spontaneous and read speech is the consequence of the specific speaking style (exclusively dictated speech). Thus, it seems advisable to also use read linguistically prepared text in LVCSR corpus design, since it ensures an appropriate triphone representation and enables controlling the phonetic structure of the utterance.

Further improvements should include both an optimized decoder and a well tuned heuristic pruning, as the recognition times currently exceed wave files duration a few fold in case of the biggest acoustic models. In particular, cross-word triphones seem to be a challenge in terms of performance (currently, only word-internal triphones are modeled during decoding). Currently, a feature space optimization experiment is being conducted, in which different parameters are investigated such as the influence of voicing and the span of neighboring frames analyzed for each observation.

6. ACKNOWLEDGEMENTS

This project is supported by The Polish Ministry of Science and Higher Education (Project: OR00006707).

7. REFERENCES

[1] Ananthakrishnan, S., Narayanan, S. 2007. Improved speech recognition using acoustic and lexical correlates of pitch accent in a N-best rescoring framework, *Proc. of the International. Conf. on Acoustics, Speech and Signal Processing* Los Angeles.

[2] Demenko, G. 1999. Analiza cech suprasegmentalnych języka polskiego dla potrzeb technologii mowy. *Poznań: Wydawnictwo Naukowe UAM*.

[3] Fischer, V., Diehl, F., Kiessling, A., Marasek, K. 2000. Specification of Databases - Specification of annotation. *SPEECON Deliverale D214*.

[4] Klessa, K., Demenko, G. 2009. Structure and annotation of Polish LVCSR speech database. *Proc. of Interspeech* Brighton, 1815-1818.

[5] Moore, R.K. 2003. A comparison of the data requirements of automatic speech recognition systems and human listeners. *Proc. Eurospeech* Geneva.

[6] Rabiner, L.R. 1989. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proc. of the IEEE* 77(2), 257-286.

[7] Sclite tool kit on-line documentation: *http://www.itl.nist.gov/iad/mig/tools/*

[8] Stolcke, A. 2001. SRILM - An extensible language modeling toolkit. *Proc. Intl. Conf. Spoken Language Processing* Denver.

[9] Szymański, M., Grocholewski, S. 2005. Transcription-based automatic segmentation of speech. *Proc. 2nd Language and Technology Conference* Poznań.

[10] Szymański, M., Klessa, K., Lange, M., Rapp, B., Grocholewski, S., Demenko, G. 2010. Development of acoustic models for the needs of a speech recognition system using large lexical databases. *Best Practices - Nauka w obliczu społeczeństwa Cyfrowego* Poznań.

[11] Young, S., et al., 2002. *The HTK Book* (for HTK Version 3.2). Cambridge University Engineering Department.