Acoustic and Biomedical Engineering

Development of Large Vocabulary Continuous Speech Recognition for Polish

G. Demenko[a,*], M. Szymański[a], R. Cecko[a], E. Kuśmierek[a], M. Lange[a], K. Wegner[b], K. Klessa[c] and M. Owsianny[a]

[a]Laboratory of Integrated Speech and Language Processing Systems
Poznań Supercomputing and Networking Center
The Institute of Bioorganic Chemistry of the Polish Academy of Sciences, Poznań, Poland

[b]Faculty of Electronics and Telecommunications, Poznań University of Technology, Poznań, Poland

[c]The Institute of Linguistics, Department of Phonetics, Adam Mickiewicz University, Poznań, Poland

In this study, the results of acoustic modeling used in a large vocabulary continuous speech recognition system are presented. The acoustic models have been developed with the use of a phonetically controlled large corpus of contemporary spoken Polish. 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 degrees of spontaneity of speech (spontaneous vs. read speech), pronunciation variants and dialects); (c) the influence of the sound level and background noises. The present large vocabulary continuous speech recognition evaluation results were obtained with Sclite assessment software. Moreover, the article delivers information about the speech corpus structure and contents and also a brief outline of the design and architecture of the automatic speech recognition system.

PACS: 43.72.−p, 43.72.+q

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 (or Arabic, Russian), 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. [1]).

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 [2], a 1000 h 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 h of speech is needed to train an ASR system with an accuracy comparable to that of a human listener. Lexical stress patterns labeled in various pronunciation dictionaries are expected to be effective indicators of pitch accents in speech [3]. These observations should be used to augment a standard ASR model to improve recognition performance. In particular, it would be of high importance for languages with a fixed place of lexical accent (the default for Polish is the penultimate syllable). It was shown in [4] that including 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 6 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. It is known that the amounts of text data required for the development of language models to be very large and representative in terms of both vocabulary and structures (e.g. [5]) to enable the extraction of plausible statistics informing on the frequency of single words and of words in context, and also to determine the lexical, syntactic, and finally — semantic patterns. A number of speech recognition systems quite successfully use language models based only on statistical word level n-grams (cf. Sect. 3 below). Yet, for inflectional languages this methodology might appear insufficient, and at least some level of linguistic knowledge may appear necessary to be formalized and implemented. This is especially the case when dealing with a language with a comparably flexible word order. In such cases the impact of an n-gram statistical language model might not be satisfactory.

In the present paper, we report on the applied methodology and on the results obtained in the experiments with various setups for the training sets used by acoustic models, namely, of the influence of the speaker gender,

[*] corresponding author; e-mail: grazyna.demenko@speechlabs.pl

speaking style and recording sound level on the recognition accuracy (Sect. 2). Then, the details concerning the text corpus and language model are presented (Sect. 3). Section 4 introduces the general design and implementation of the present large vocabulary continuous speech recognition (LVCSR) system. The results of the prototype system and their evaluation are presented in Sect. 5. The next section discusses the experimental results, while Sect. 7 provides conclusions and a brief summary of the future work and the necessary improvements.

2. Methods used to construct and evaluate the acoustic models

2.1. Speech corpus

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 acoustic database contains recordings of speech delivered in quiet office environments by over 2000 speakers (a total of over 1155 h of speech) from 16 regions of Poland. All data were first annotated manually according to SpeeCon guidelines [6]. The SpeeCon guidelines assume orthographic, word-level transcription with only several non-speech events markers for speaker and background noises. Then, for the purposes of acoustic modeling, the files were subject to automatic, phone-level segmentation using Salian [7]. The conversion of the orthographic SpeeCon annotation labels into phonetic transcriptions was made with the use of a large lexical relation database *Speechlabs.ASR* [8], the central lexical resource for the present ASR system project, providing the information on above 3 million word forms (orthographic and phonetic transcriptions for two most popular regional pronunciation variants, accentuation and syllabization, part-of-speech tags, word inflection, and special unit categorisation, e.g. proper names, abbreviations, words of foreign origin). The Jurisdict database (cf. also [9]) is composed of the three main types of recordings: (a) *read* speech (texts designed specifically for the coverage phonetic and syntactic structures as well as original legal texts provided by the future end-users — for lexical coverage); (b) *semi-spontaneous* speech (controlled dictation); (c) *spontaneous* recordings from court trials.

For the needs of the acoustic modeling experiments described below over 568 h of speech produced by 1488 speakers were selected from the Jurisdict database (namely, its read and semi-spontaneous speech subcorpora).

2.2. Training tools

The acoustic speech models were trained using HTK [10]. 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 (most importantly, the manner and the place articulation) served for state and triphone clustering.

2.3. The number of Gaussian mixtures

The subject of the first experiment was to investigate the dependence of the accuracy and speed of speech recognition on the number of Gaussian mixtures in each state. For each tested acoustic model three setups 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 expert supervision.

Table I below 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 — the recognition time was significantly longer with the 24-mixture model.

TABLE I

Acoustic modeling results for different mixtures number (% acc — mean percentage of correctly recognized minus inserted words, std dev. — standard deviation across speakers, % r. time — recognition time percentage, 100% = 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 II

Acoustic modelling results for gender dependent models (% acc — mean percentage of correctly recognized minus inserted words, std dev. — standard deviation across speakers).

	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)

2.4. Training sets — speakers' gender

In the second experiment, the influence of the speaker's sex was tested in terms of using male or female voices at the stage of model construction as opposed to using a mixed male and female model.

For the need of this 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 II, 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 as compared to F and M models.

2.5. Training sets — spontaneous and read speech

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 h of speech produced by 1488 speakers was used. For testing, the test set from the Gaussian mixtures experiment was used (cf. 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.

2.6. Training sets — sound levels

The subject of the sound levels experiment was to test whether 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 0 dB and −13 dB.

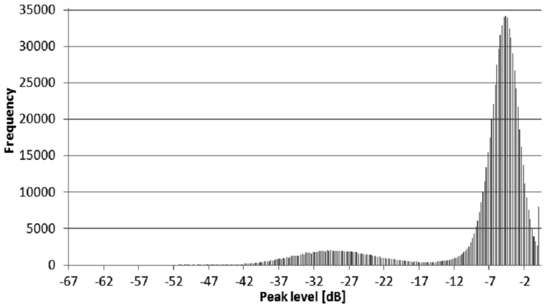

Fig. 1. Distribution of peak level in original dataset.

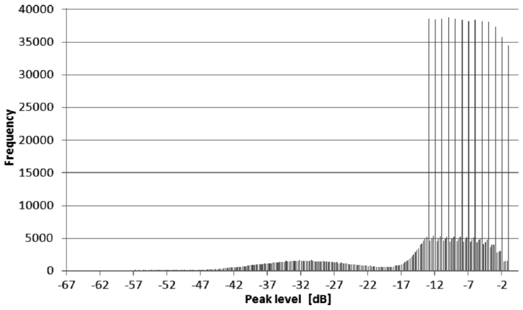

Fig. 2. Distribution of peak level in preprocessed dataset.

TABLE III

Acoustic modeling results for different training data sets (% acc — mean percentage of correctly recognized minus inserted words, std dev. — standard deviation across speakers, % r. time — recognition time percentage, 100% = the real recognition time).

	Test file volume	Model on preprocessed data	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

Figures 1 and 2 depict the distribution of peak levels in the original and preprocessed datasets. In Table III the word level recognition rate of original and preprocessed training set models are presented. The test set used in the experiment covered 147 utterances from 20 speakers.

The obtained results may suggest that the model trained on a 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. However, additional experiments are required, to determine the extent to which these phenomena are caused by low recording levels themselves, as opposed to possibly inadequate level of insensitivity of the signal parameterization stage, as the latter is still under tuning.

3. Methods used to construct and evaluate the language model

3.1. Text corpus, text processing and model construction

The corpus used for language modeling contained over 4 GB of text contained in judiciary documents (court protocols, witness testimonies and statements, briefs, reports, legal contracts etc.). Additionally, a database of legal advices (a formal on-line legal service) was included and also a set of newspaper and journal articles as well as transcripts of the Polish parliament speeches.

The language model used for rescoring of the decoder results was a word based statistical model. More precisely, a 3rd order n-gram model was built with the Kneser–Ney discounting and the Katz back-off used for smoothing.

In order to obtain a form suitable for statistical analysis text preprocessing was necessary. Besides, the standard conditioning such as sentence start and end labeling, converting to uppercase and dealing with punctuation marks, further preprocessing was necessary to correct misspellings, expand abbreviations and replace numerals. The latter part of preprocessing is language dependent and quite challenging to perform automatically for a highly inflected language such as Polish.

The main difficulty is the selection of a correct inflected form of a word based on the context in which the word occurs in the text. The preprocessing was performed semi-automatically with some manual corrections. The language model was then built with SRILM toolkit [11] based on a dictionary of 370 thousand words. The dictionary was created based on the text corpus, by selection of the top frequent words, mostly unigrams. The largest model we have experimented with, contained over 30 million bigrams and over 36 million trigrams. We have adopted an open vocabulary approach.

3.2. Language model evaluation

The language model influence on the speech recognition accuracy was evaluated by comparing results obtained with various weights assigned to language probability for computing the overall hypothesis probability. Specifically, for language weight set to 0, language probability was not taken into account at all. Our experiments showed that the language model based rescoring improved speech recognition accuracy by only a few percent on the average. The best results were obtained with the language model weight set to around 0.6 and acoustic model weight set to 0.4. We consider a few percent improvement to be a rather disappointing result. The conclusion we drew was that a simple statistical word based model was not sufficient for a highly inflected language.

There are several directions which we plan to pursue. The recognition results analysis conducted with Sclite tool [12] showed that it was a common case that a correct word was recognized but an incorrect inflected form of this word was selected. Based on this fact, we expect that a model with word stems and endings used as modeling units, should improve accuracy. An improvement on the order of a few percent due to stem and ending based model, was reported in [13] for Slovenian language, which also is a highly inflected language. In addition to accuracy improvement, such an approach can also reduce model size since there are many common sub-word units in inflected languages. A more demanding but also a more promising approach is to use a parts-of-speech (POS) based model. A grammatically tagged corpus is needed as well as a morphological lexicon to build such a model. Inflectional nature of a language again makes the tagging process difficult due to a high degree of ambiguity. Most likely a combined approach will be needed, i.e., a simple statistical n-gram model accompanied by a sub-word or a POS model, in order to achieve better rescoring results.

4. ASR system implementation

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).

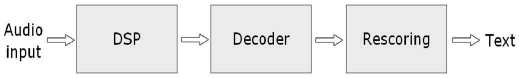

Fig. 3. The architecture of the LVCSR system.

The audio data in the form of PCM signal is passed to digital signal processing (DSP) module (Fig. 3). 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 a linear discriminant analysis (LDA) 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 built upon modified Viterbi algorithm [14] 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 result. The Lattice elements are attributed with appropriate probabilities. All hypotheses are evaluated using linguistic model in the Rescoring module. Hypothesis with best probability is returned as a recognized text.

5. Results of the prototype ASR system

The evaluation tests have been carried out using the Sclite tool (performing text alignment through a minimization of a Levenshtein distance), with recordings from 97 speakers (7 749 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 (as much as 10%) with the total error percent of 15.4% and 87.3% correct. The percent of errors caused by deletions was 2.7% and those from insertions accounted for 2.8%. The percent of word accuracy was 84.6%. 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 of: *language model* (LM) and *speaker adaptation*.

TABLE IV

The influence of using language model rescoring (% acc — mean percentage of correctly recognized minus inserted words, std dev. — standard deviation across speakers, % r. time — recognition time percentage, 100% = the real recognition time).

		Unigram LM	Trigram 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 V

Adaptation results for 13 speakers (% acc — mean percentage of correctly recognized minus inserted words, std dev. — standard deviation across speakers, % r. time — recognition time percentage, 100% = the real recognition time).

		without adapt.	with adapt.
word	% acc (std dev.)	84.6 (4.9)	88.6 (3.6)
sentence	% acc (std dev.)	46.2 (7.3)	53.6 (7.1)
	% r. time	290.4	131.3

Table IV presents the test results showing the influence of the applied language model on the recognition accuracy and recognition time given in percents.

Table V shows the influence of speaker's adaptation on the recognition accuracy tested on 13 speakers. The text of 257 sentences (duration of *ca.* 25 min) including utterances taken from the set covering triphones and the fragments of the police reports has been used as an adaptating set. The supervised way of adaptation with maximum likelihood linear regression (MLLR) was applied [10]. The results in Table V show that the adaptation improves speech recognition by about 4% (26% reduction of error) and reach the level of acc% equal to 88.6%. As the baseline speaker-independent model in this experiment was trained on a single microphone, recognition accuracy was poor (*ca.* 57%) on sentences recorded using microphones with acoustic characteristics different than the main microphone. After the adaptation process, the results of speech recognition for different microphones are similar (standard deviation with respect to different microphones is 1.6%), as for the "unseen" microphones the adaptation boosts the accuracy by 30% (67% reduction of error). In the tests, the significant relationship between the recognition time and the utterance quality (in terms of speaker performance) has been observed (the better the speaker the shorter the recognition time). The speaker voice adaption shortens also the recognition time. In all cases the 8-mixture Gaussian acoustic model was used.

6. Discussion

The evaluation results suggest that we are already close to the expected system accuracy [2] even when no language modeling was implemented. However, the quality and impact of the 3-gram language model is still not satisfying. Thus, the implementation of a language model using grammatical information and also detection and clustering of proper names and other special lexical units appears as the necessary step on the way to achieve a really significant progress. As the first step to achieve this goal, the text corpus used to build the language model was transformed into a relation database, and then the contents of the lexical database *Speechlabs.ASR* [7] was used to provide part-of-speech and inflection tags for each word of the speech corpus. After that, an important task emerged, namely, disambiguation of the assigned tags (Polish is characterized by a high degree of ambiguity across and within word inflection paradigms, cf. also [15]). As a starting point, the input categories for the proper names clustering have been established, relating to the first names, last names and selected groups of toponyms; a number of sub-categories is also needed based on inflectional groups distinguished within the main categories (again, based on the classification proposed in *Speechlabs.ASR*).

The statistical insignificance of the differences in recognition of spontaneous and read speech in the above acoustic modeling experiments is the consequence of the specific speaking style (exclusively dictated speech of a rather formal style). Thus, it confirms the validity of the use of read, linguistically prepared text in LVCSR corpus design, since it ensures an appropriate triphone representation and enables controlling the phonetic structure of the utterance. Since the times of recognition in the present experiments exceed wave files duration a few fold in cases of the largest acoustic models, further improvements of the present prototype ASR system should include both an optimized decoder and a well tuned heuristic pruning. In particular, cross-word triphones appear to pose a challenge in terms of performance (thus, so far, only word-internal triphones have been modeled during decoding). Currently, a feature space optimization experiment is being conducted, in which different parameters such as the influence of voicing and the span of neighboring frames analyzed for each observation are investigated. Based on linguistic assumptions and preliminary laboratory tests, it is expected to additionally enhance the system's performance.

7. Conclusions

We presented the methods of development, as well as the results and evaluation of various setups for acoustic models used a LVCSR. We present the setups for:

- the number of Gaussian mixtures (the 24-mixture model giving the best results, however with slightly longer recognition times);
- the influence of the gender of the speakers whose voices are used at the stage of models construction (using separate models for male/female voices appeared more successful than using mixed models for both genders);
- the influence of preprocessing the training set recordings (the model trained on a more level-varied training set performs not significantly worse as compared to the model trained on original files).

Moreover, we include description of the methods applied to construct the language model used in the present system as well as its preliminary evaluation. Subsequently, an outline of the system is presented, followed by the evaluation of the results. Then, the results are discussed, especially from the perspective of the on-going work and the possible further improvements.

Acknowledgments

This project is supported by the Polish Ministry of Science and Higher Education (project ID: OR00006707 — Integrated system of automatic speech-to-text conversion based on linguistic modelling designed in the environment of the analysis and legal documentation workflow for the needs of homeland security).

References

[1] G. Demenko, S. Grocholewski, K. Klessa, J. Ogórkiewicz, A. Wagner, M. Lange, D. Śledziński, N. Cylwik, in: *Proc. 6th Int. Language Resources and Evaluation Conf., Marrakech 2008*.

[2] R.K. Moore, in: *Proc. Eurospeech, Geneva 2003*, p. 2582.

[3] S. Ananthakrishnan, S. Narayanan, in: *Proc. Int. Conf. on Acoustics, Speech and Signal Processing, Los Angeles 2007*.

[4] M. Szymański, K. Klessa, M. Lange, B. Rapp, S. Grocholewski, G. Demenko, *Best Practices — Nauka w obliczu społeczeństwa cyfrowego*, Poznań 2010, p. 280.

[5] *Handbook of Standards and Resources for Spoken Language Systems*, Eds. D. Gibbon, R. Moore, R. Winski, Mouton de Gruyter, Berlin 1997.

[6] V. Fischer, F. Diehl, A. Kiessling, K. Marasek, *Specification of Databases — Specification of Annotation*, SPEECON Deliverale D214,2000.

[7] M. Szymański, S. Grocholewski, in: *Proc. 2nd Language and Technology Conf., Poznań*, 2005.

[8] K. Klessa, M. Karpiński, O. Bałdys, G. Demenko, *Speech and Language Technology*, Vol. 12/13 Polish Phonetic Association, Poznań, 2009.

[9] K. Klessa, G. Demenko, in: *Proc. Interspeech, Brighton (UK) 2009*, p. 1815.

[10] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland, *The HTK Book (Version 3.2)*, Cambridge University Engineering Department, Cambridge 2002.

[11] A. Stolcke, in: *Proc. Int. Conf. Spoken Language Processing, Denver*, 2001.

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

[13] M. Maucec, T. Rotovnik, M. Zemljak, *IJST* **6**, 245 (2003).

[14] L.R Rabiner, in: *Proc. IEEE* **77**, 257 (1989).

[15] M. Steffen-Batogowa, T. Batóg, *The Families of Polish Homophones. Dictionary of Homophones*, Vol. 1, 2, Wydawnictwo UAM, Poznań, 2009, (in Polish).