

First evaluation of Polish LVCSR acoustic models obtained from the JURISDIC database

Marcin Szymański,^{*,**} Jerzy Ogórkiewicz,^{*}
Marek Lange,^{*} Katarzyna Klessa,^{***}
Stefan Grocholewski,^{**} and Grażyna Demenko^{**}

^{*}Laboratory of Speech and Language Technology,
Adam Mickiewicz University Foundation, Poznań

^{**}Institute of Computing Science, Poznań University of Technology

^{***}Institute of Linguistics, Adam Mickiewicz University, Poznań

Marcin.Szymanski@cs.put.poznan.pl

sova_jo@sylaba.poznan.pl

marek.lange@gmail.com

klessa@amu.edu.pl

stefan.grocholewski@cs.put.poznan.pl

lin@amu.edu.pl

ABSTRACT

This paper presents the results of the pilot survey of the acoustic models obtained from the Polish Speech Database for taking dictation of legal texts, created for the needs of the first LVCSR system for Polish (JURISDIC). Additionally, background information about the design of the database is presented along with the description of the applied methods of the corpus construction and current statistics of the database contents.

1. Introduction

A review of the results of ASR systems developed for various languages shows that while creating such a system for highly inflexional languages like Polish certain assumptions concerning acoustic-phonetic database structure need to be modified (as compared to e.g. English) to provide adequate material for both acoustic and language modeling. Acoustic models should be derived from large corpora, involving many speakers selected to represent a typical distribution in age, sex or geographic area so that they represent an average for a particular language, and guarantee good performance in most recording scenarios. Polish is not regarded to be very diverse in terms of dialects however certain regional varieties need to be taken into account.

The acoustic models whose results are provided further in the paper were prepared with a view of their future incorporation within the ASR system for Polish constructed in the framework of the JURISDIC project [1]. As the speech corpus a part of the JURISDIC database for the needs of taking dictation of legal texts was used.

2. Corpus information

The JURISDIC database is still under construction, however it is already possible to retrieve data recorded, pre-validated and annotated so far (see the Database Statistics section for more details).

2.1. Text corpus structure

2.1.1. *Semi-spontaneous speech. Corpus A*

- Sub-corpus 2A. Spontaneous Dictation (legal, police, court vocabulary). This sub-corpus contains formal speech (dictation on various application topics). Typical tasks are: dictation of any kind of legal texts (areas: judicial, disciplinary, criminal, divorce) in court, police reports (different topics, e.g. a description of a theft, burglary using common vocabulary, etc.). The number of the recorded topics varies between speakers.
- Sub-corpus 2A. Spontaneous Dictation (common topics). This sub-corpus contains informal speech (dictation on various common topics). Typical tasks are: a description of a birthday, giving directions, giving an excuse, a description of holidays, etc. The speaker is requested to be speak in a neutral style following instructions such as: Imagine that you are calling your friend/father/boss and telling them something/excusing yourself/deciding on something, etc. The number of the recorded topics varies between speakers.
- Sub-corpus 3A. Elicited Dictation (Answering questions). The aim of sub-corpus 3A is to obtain some semantically important, frequent items such as birth dates, relative dates, times of day, city names, proper names, age, money amounts, currencies, sequences of digits and numbers, telephone numbers, mathematical operations as well as answers like yes/no/maybe, etc. and education, profession, etc. (27 categories).

2.1.2. *Read Speech. Corpus B.*

- Sub-corpus 1B. Grammatically and Phonetically Controlled Structure. Syntactically complex sentences. By 'syntactically complex' we mean: a) variable concatenation of phrases, b) variable phrase length. By 'phonetically controlled' we mean: adequate coverage of triphones, triphones in the final position of a word/phrase. For selection of the phonetically rich sentences (from above 3000 sentences) the following constraints are set: each speaker produces 60 complex sentences, each sentence is read by 15–20 speakers.
- Sub-corpus 2B. Phonetically controlled structure. Syntactically simple sentences. We expect that 90 short sentences will be provided by each speaker with the explicit intention of obtaining an adequate coverage for the chosen consonant clusters, short bigrams and triphones both in the accented and unaccented position. The whole 2B Corpus should contain approximately 4000 sentences. Each sentence should be read by 20 speakers. The main aim of the Corpus B was to obtain: a) CVC triphones in context of sonorants in a chosen accented/unaccented position. The number of accented positions depends on a particular word's frequency, e.g. for triphone. b) CVC triphones in context of voiced consonants in a chosen accented/unaccented position. The number of accented positions depends on a particular word's frequency. The whole sub-database has approximately 800 sentences with controlled consonant clusters. c)

Examples of short bigrams in utterance initial position. The whole sub-database consists of approximately 2000 sentences with the controlled bigrams (e.g. two conjunctions, conjunction and preposition, etc.) in initial position and in the middle of a phrase for the most frequent bigrams. The short (one- or two-syllable) words are most difficult to recognize for ASR systems. d) Examples of consonant clusters: the whole sub-database consists of approximately 800 sentences with controlled consonant clusters. Special attention was given to CCCC and CCCCC clusters like: pstf, mpsf: gļupstwo, skapstwo (Eng. nonsense (or trifle), avarice).

- Sub-corpus 3B. Special lexical phrases (words)
The sub-corpus with more than 400 short one- or two-word includes special words like modulants, greetings, jargon/vulgar expressions. It was constructed manually based on dictionaries and other resources for Polish. At least 7 items are provided by one speaker.

2.1.3. C. Read Speech. Semantically Controlled Structure

- Sub-corpus 1C. General purpose words and phrases. Within this group utterances are divided into: general words/phrases and general-purpose commands. The general-purpose words/phrases include 33 categories, among them: isolated digits, numerals, measures, letters, special keyboard characters, special legal acronyms, emails, web addresses. No instructions are given to speakers as to how to spell these items.
- Sub-corpus 2C. Application-specific short texts for users' needs. Texts extracted from original police reports and professional legal documents (up to 100 sentences).

2.2. Database Statistics

The corpus used for the present experiments contained over 116 h of speech, produced by 321 speakers which is a part of recordings of 1030 speakers already included in the JURISDIC Annotation Database, ca. 900 of which have been already annotated.

The database preparation is still in progress and is planned to close this year; the assumed variable part of the database will include speech delivered by approximately 2000 speakers. The recordings included in the corpora (will) come from: a) the court (speech by a judge), b) the legal/notary's/prosecutor's office (speech by a lawyer), c) the police station (speech by a police officer), approximately 500 voices, d) office/high school teachers/university: approximately 600 voices. The distribution of sex and age is approximately 50:50. Although Polish is not very diverse as far as dialects are concerned, the recordings have been done in 16 main districts of Poland. The session recorded for each speaker consists of approximately 20–40 minutes of semi-spontaneous speech and, depending on the speech tempo, approximately 30 min of read speech (about 170 shorter and longer sentences). The speakers are asked to read a text in a dictation style.

2.2.1. Phonetic coverage

The overall statistics of triphone coverage within the whole read speech phonetic coverage text corpus is as follows: within-word triphones: 10593, triphones containing an accented vowel: 8492, unaccented triphones 10650, triphones in phrase final position: 4495. Triphone lists serving as reference for the purpose of manual preparation of the read speech text corpus were created based on 2 million words randomly selected from a corpus of texts including about 10 million words altogether.

2.3. Recording procedure

The recordings for the JURISDIC database purposes are stereo recordings acquired in an office environment from two microphone positions: a ‘close distance’ and ‘medium distance’ position using a headset microphone and a ‘table’ microphone. (Sennheiser ME-3 for ‘close distance’ position, and AKG C-1000S – for the ‘middle distance’ position).

To enable easy management of great number of speakers data and the recorded utterances together with their description a special tool was created using JAVA as the programming language: the *QuestionRecorder* program (cf. [1]).

2.4. Annotation

In the first stage the recordings are labeled by a group of 30 trained students of the Institute of Linguistics in Poznań whose work is supervised (and corrected if necessary) by a phonetician. The second step will be a thorough verification of the label files by a team of phoneticians accompanied by the automatic parsing of the files in order to syn-chronize the files contents with the lexicons available for the labelers in the annotation system. The first synchronization with the lexicon has been already performed.

The annotation specification follows the guidelines of SPEECON [2] adjusted to the Polish language. The transcription is orthographic, case sensitive. Proper names and foreign words as well as any application words or abbreviated forms received their language-specific spelling rules. The labelers may check spelling on the current basis via the ‘search lexicon’ option available in the annotation system. The conventions of labeling mispronunciations and noises and the planned post-hoc assessment method are in agreement with the abovementioned SPEECON document.

For the purpose of the annotation of the recorded speech data new software was designed based on the Client-Server architecture using MSDE 2000, and Windows 2003 Server Client applications were programmed in C#. The tool was called *PPBW Annotation Database Manager* (cf. [1] and the JURISDIC project’s website) and is in charge of all the stages of the annotation procedure connected with sound and label files, text files, speaker information, lexicons search, and multi-user management. The program enables the import of the recordings produced with *QuestionRecorder* and the respective text files to the Annotation Database (after the database annotation is completed it will be possible to export all the files again to the required final database format, the Export options are ready to use, and may be applied at the intermediate stages of the project as well).

2.5. Recording quality assessment

The recordings are pre-validated on a current basis by an expert phonetician with the help of a special tool: “Recording Checker” designed specifically for the recording control procedure in the present project. The most important characteristics of the program are as follows: volume measure module and distortion detector; session completeness control module; subjective assessment module (reading style, pronunciation, possible noises, reverberation, wrong microphone setup); session(s) assessment reports, plus a comfortable interface for listening to the recordings; easy navigation between recording sessions.

3. Preliminary acoustic modeling and experiments

The present results were obtained using only the close-talk microphone recordings of 116 h of speech, produced by 321 speakers.

The database was divided into three sets: utterances with a low (bottom quarter), moderate (interquartile range) and high speech rate (top quarter), respectively. Speech rate was determined based on automatic alignment and estimated as the inverse mean duration of a vowel. The moderate rate set was randomly reduced to half its initial size to assure that all the sets contained the same number of sentences. (The slow, moderate and fast sets contained the same number of utterances. However, it was later found that the low speech rate part contained ca. 30% shorter sentences, resulting in a lower number of training examples.) The speakers were split into 5 cross-validation sets in such a way that the number of speakers, the number of slow sentences and the number of fast sentences were approximately equal across different cross-validation sets. This guaranteed that the same speaker was never used for both training and testing (even with a different speech rate).

A list of words was generated from orthographic annotations, from which a dictionary was automatically generated using grapheme-to-phoneme transcription rules. The dictionary contained over 32000 different words. The recognition grammar allowed any sequence of words from the dictionary, ie. no higher level language model was used at this stage.

The stochastic acoustic speech models for Polish were trained using HTK [5]. Prior to the modeling, the corpus was segmented by forced alignment using models based on a different 5h database. The standard training procedure including HInit, HRest, HERest and HHed for triphone CDHMM was generally used: a list of ca. 60 contextual ‘questions’ served for state clustering; the average number of Gaussian mixtures in each state was set to 12.

Table 1 below presents the word level recognition rates of 5-fold cross-validation tests on different speech rates (one group of speakers for testing and the other four for training) on different speech rates. Apart from 3 ‘homogenous’ speech rate sets (*fast*, *moderate* and *slow*), two speech rate-independent acoustic models were trained: *big* denotes a set combining all sentences, *diffr* denotes a speech rate-independent corpus, which was reduced to the same size as fast, moderate or slow. It can be observed that:

- the high speech rate test sentences were consistently harder to recognize than the moderate ones, even for high speech-rate models,
- the ‘big’ speech-independent corpus yielded better or similar results compared to speech dependent sets,
- save the above rules, the best recognition rates were obtained for models of the same speech rate class as the test set.

At this preliminary stage we focus on the problem of speech rate, as it is one of the sources of recognition rate degradation, because of poor acoustic matching or even reduction of some phones in case of fast speech [6]. The results suggest it is not worth maintaining separate rate-dependent speech models, however more thorough experiments still should be performed.

For speech of rate estimation we used automatic forced alignment. Although certain authors report performing the manual correction of an automatic segmentation

Table 1. Acoustic modeling results for different train and test speech-rate classes. ‘mu’ denotes mean percentage of correctly recognized words, ‘sig’ denotes a standard deviation across cross-validation folds

test set	model	big	diffr	fast	moder.	slow
fast	mu	55,67	52,31	53,99	52,54	39,39
	sig	1,45	1,30	1,12	2,36	1,87
moder.	mu	61,52	59,19	56,26	61,64	52,88
	sig	1,08	1,07	0,86	3,05	1,40
slow	mu	59,39	56,75	40,09	51,90	59,51
	sig	1,03	0,95	1,16	0,58	1,45
cumulative	mu	58,62	55,80	51,39	55,56	49,01
	sig	0,65	0,29	0,59	1,69	0,41

prior to calculating the speech rate [7], others claim that Viterbi alignment can provide a robust estimation, if a transcription is known [8]. Also, the differences between the results for different train and test sets suggest that the accuracy of the speech rate estimation was reasonable in our case. We consider, however, to apply a more sophisticated approach in the future, as compared to e.g. Wang et al. [9].

4. Discussion

The preliminary results are encouraging, but still below what would be acceptable for practical applications.

For the evaluation results reported in Section 3 above only 321 speakers and 116 h were available. The current plans assume including the rest of the corpus of over 2000 speakers, resulting in ca. 1000 h of speech. (It will be possible to provide the general statistics for the database after the annotation of the variable part of the database. The evaluation process by an independent center (e.g. ELRA [10]) should estimate the quality and usefulness of the database for building ASR system for Polish.) This should boost the recognition rate, however according to Moore [11] 1000 h-database allows for building a system with a word error rate of ca. 12% when a 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.

Acoustic and language modeling of Polish as a language with complex flexion rules and a comparably flexible word-order might encounter more problems than it is the case for English (cf. for example a comparison of ASR results for English, Spanish and Chinese in [12], for more results and evaluation information cf. e.g. to the TC-STAR website [13] and [14, 15, 16]). A 32000 word dictionary was used for first experiments, in the final system it is planned to use a dictionary of at least 150 000 words.

Moreover, a well tuned heuristic pruning is necessary, as the recognition times currently exceed wave files duration a few fold (and further lengthening of the recognition times is inevitable considering the planned final size of the dictionary).

5. Future work

The present results should be regarded as the preliminary verification of the development of acoustic models for Polish speech recognition system. They are encouraging however further experiments are indispensable to improve the obtained acoustic models and provide an outcome practically useful in the designed speech recognition system. Firstly, we intend to take the following steps:

- tuning modeling and testing parameters (eg. number of mixtures, word insertion penalty);
- including distant microphone recordings in training.

The next step will be incorporating language models into the recognition. We expect that integrating linguistic and acoustic models will improve the overall performance of our system. It is planned to submit the specifications of the corpus and the system to external validation as we understand the importance of their compatibility as referred to current standards e.g. [17, 18].

Acknowledgements. This research was supported by the Polish Ministry of Scientific Research and Information Technology, project no. R00 035 02.

BIBLIOGRAPHY

- [1] Demenko, G., Grochowski, S., Klessa, K., Wagner, A., Ogórkiewicz, J., Lange, M., Śledziński, D., Cylwik, N. Jurisdic – Polish Speech Database for taking dictation of legal texts. Accepted for LREC Conference, Marrakech, Morocco, 2008.
- [2] Fischer, V., Diehl, F., Kiessling, A., Marasek, K. 2000. Specification of Databases – Specification of annotation. SPEECON Deliverable D214.
- [3] JURISDIC project and Laboratory of Speech and Language Technology website: <http://www.speechlabs.pl>.
- [4] Demenko G., Wypych M., and Baranowska, E. (2003). Implementation of Grapheme-to-Phoneme Rules and Extended SAMPA Alphabet in Polish Text-to-Speech Synthesis. *Speech and Language Technology, Edition PTFON*, vol. 7.
- [5] S. Young and J. Odell and D. Ollason and V. Valtchev and P. Woodland. *The HTK Book (for HTK Version 2.1)*. 1997.
- [6] Nanjo H., and Tatsuya Kawahara, Language Model and Speaking Rate Adaptation for Spontaneous Presentation Speech Recognition, *IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING*, VOL. 12, NO. 4, JULY 2004.
- [7] Keikichi Hirose, Hiromichi Kowanami, Temporal rate change of dialogic speech in prosodic units as compared to read speech, *Speech Communication* 36(2002), pp 97–111, 2000.
- [8] Mirghafori, N., Fosler, E., Morgan, N., 1996. Towards robustness to fast speech in ASR. In: *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Atlanta, Vol. 1, pp 335–338.
- [9] Wang D., and Shrikanth S. Narayanan, Robust Speech Rate Estimation for Spontaneous Speech, *IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING*, VOL. 15, NO. 8, NOVEMBER 2007, pp 2190–2201.
- [10] ELRA: European Language Resources Association homepage: <http://www.elra.info/>.

- [11] Moore R. K., A comparison of the data requirements of automatic speech recognition systems and human listeners, Proc. Eurospeech, Geneva, pp 2582–2584, 2003.
- [12] Mostefa Hamon O., Choukri K., Evaluation of Automatic Speech Recognition and Speech Language Translation within TC-STAR: Results from the first evaluation campaign. in: Proceedings of LREC 2006, full paper available on-line at: <http://www.mt-archive.info/LREC-2006-Mostefa.pdf> (accessed 12 April 2008).
- [13] TC-STAR project homepage: <http://www.tc-star.org/>.
- [14] Docio-Fernandez Laura, Antonio Cardenal-Lopez, Carmen Garcia-Mateo, TC-STAR 2006 Automatic Speech Recognition Evaluation: The UVIGO System, TC_STAR Workshop on Speech-to-Speech Translation, June 19–21, 2006, Barcelona.
- [15] Henk van den Heuvel, Eric Sanders, Validation of language resources in TC-STAR, TC-STAR Workshop on Speech-to-Speech Translation, Barcelona, June 19–21, 2006.
- [16] Loof J., Ch. Gollan, S. Hahn, G. Heigold, B. Hoffmeister, Ch. Plahl, D. Rybach R. Schluter and H. Ney, The RWTH 2007 TC-STAR Evaluation System for European English and Spanish, Interspeech 2007, 2145–2149.
- [17] Trancoso, I. Speech Recognition Activities in Europe – Recent Trends, available from the Instituto de Engenharia de Sistemas e Computadores Investigação e Desenvolvimento em Lisboa website at: <http://www.inesc-id.pt/pt/indicadores/Ficheiros/3274.pdf> (accessed 13 April 2008).
- [18] Databases for the Creation of Voice Driven Teleservices <http://www.speechdat.org/SpeechDat.html>.