Polish Language Resources for Speech Technology: JURISDIC LVCSR corpora

Grażyna Demenko, Stefan Grocholewski, Katarzyna Klessa, Zbigniew Rau

The Institute of Linguistics, Adam Mickiewicz University, Poznań,
Poznań University of Technology
Adam Mickiewicz University Foundation, Poznań
Polish Platform for Homeland Security
lin@amu.edu.pl, stefan.grocholewski@cs.put.poznan.pl, klessa@amu.edu.pl, zbigniew.rau@ppbw.pl

Abstract

The paper discusses the problems of re-usability and extensibility for LVCSR (Large Vocabulary Continuous Speech Recognition) corpora exemplifying the discussion with the results from the Polish JURISDIC database. Special attention is given to the Polish language-specific features. We also look at the characteristics of the speakers whose voices are collected in the databases especially as regards balancing the corpus in terms of spoken language differences depending on regional and also professional differentiation. Besides, the corpus provides resources for fundamental research based on contrasting read, semi-spontaneous and spontaneous speech for rhythm modeling, pronunciation differences, segmental and suprasegmental variation of speech. The article presents recordings and annotation statistics including selected linguistic information as well as the information about the non-speech events and other phenomena occurring in the collected data considering their use for the needs of speech technology, primarily, for automatic speech recognition systems [1]

Keywords: LVCSR, corpus design, annotation, evaluation of databases

1. Introduction

Large text and speech corpora are indispensable to meet the challenges of various contemporary speech technology applications. The size and structure of the corpora depends heavily on their future application, however the current speech technology issues require substantial corpora to answer the questions of naturalness, diversification and changeability of languages. The usefulness and importance of creating the resources has already been confirmed by research and in practice thanks to several existing large language corpora (Vermobil (Maier et al., 1994) or SpeeCon). One of the largest existing corpora is the English 'Fisher' Corpus delivering 2300 hours of transcribed conversational telephone speech, and 75000 vocabulary items that covers training data (eg. Fisher et al., 2000). An excellent overview of the available resources is offered by ELDA (Evaluations and Language resources Distribution Agency). Since designing and creating large corpora is a very expensive and time-consuming venture it is strongly advisable to explore the collected data as effectively as it is possible - i.e. to extract all the information potentially contained in the material. In this paper we will argue that one of the key issues is high quality linguistic information (starting from the recording scenarios, through the prompting instructions for speakers up to easily searchable structure of the annotation database). For these reasons formal evaluation procedures should be applicable not only to recording statistics but also to the linguistic structures.

Obviously, it is not possible to handle all the factors influencing the phonetic-acoustic structure of the corpus, nevertheless an attempt to consider them should enhance the overall corpus quality not only for the purposes of its dedicated application. In the present database the following factors were controlled: speaker sex, age, and dialect, his/her linguistic competence, speaking style, the degree of speech formality.

The preliminary results of acoustic modeling obtained for the present data are encouraging with respect to implementing accent information and utterance speech rate data. Parallel to the ongoing internal evaluation and validation of the JURISDIC database the external validation procedures has also been started with ELDA.

2. JURISDIC LVCSR corpora

The JURISDIC speech database is a large continuous speech database for speech recognition, possibly the largest one for the Polish language presently available, for more details cf. (eg. Demenko, et al. 2008). The first planned application of the database is the Polish speech dictation system for the use of police and legal authorities and attorney's offices. The database was designed according to the functional requirements of a dictation system for the purposes of courts, police and lawyer's offices. The applied specifications were selected to meet international standards for language resources with necessary adjustments resulting from the language specificity, the future applications and the substantial size of the database. The general assumptions for the structure of database take into account text structure: semantic structure, syntactic factors, grammatical and acoustic-phonetic factors and speaking style: semi-spontaneous, controlled spontaneous dictation, elicited dictation (answering speech). The specification is based on the general language features and also on peculiarities of Polish on the different linguistics as well as phonetics levels.

2.1. Speakers' profiles

Altogether the database contains above two thousand voices. As for the speaker characteristics in terms of the

[1]This project is supported by The Polish Scientific Committee (Project ID: R00 035 02 – "Technologies for processing and distributing verbal information in internal security systems")

professional or social status: they were selected taking into consideration the first application of the system the most numerous group were policemen and students of police schools all over the country (above thousand voices, which consists of more than a half of all recorded speakers). The remaining speakers were judges and attorneys (a hundred speakers) and also university students and office workers. The speakers were adults aged of 20-50. It was intended to cover equally for male and female voices. The current distribution of speakers' age is presented on the figures below distinguishing the numbers of female and male speakers. The most numerous age group for both females and males is between the age of 20 and 30. The total number of male speakers was higher by approximately 200 than females (which may be explained - and justified - by the type of profession of the recorded speakers being at the same time the first group of end-users).

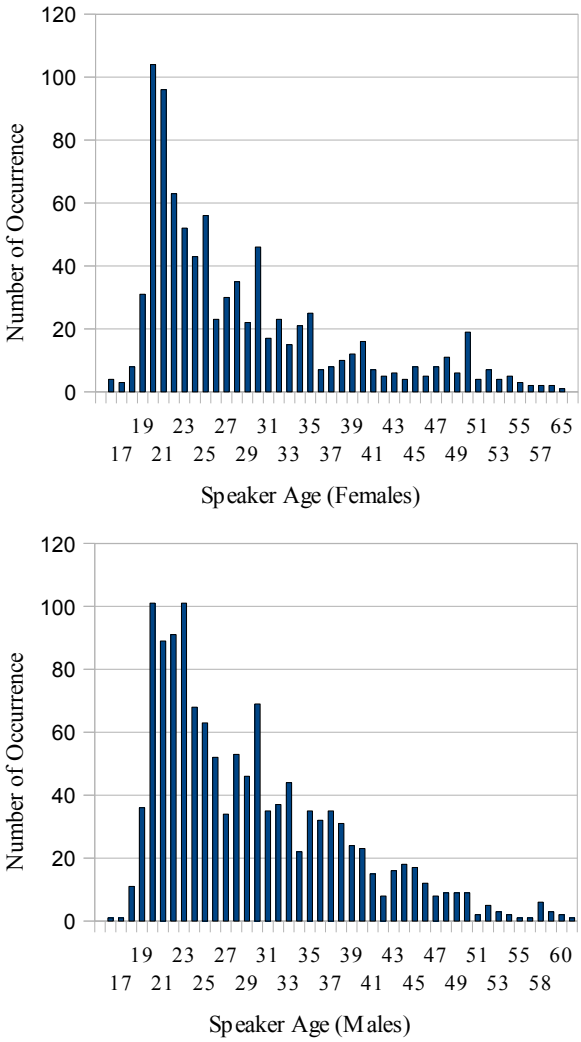

Figure 1: The distibution of speaker age in JURISDIC (top: Females, bottom: Males)

The speakers were informed about the task in advance by providing them with a short pilot specification. During the recording sessions, trained technicians supervised the reading style, the speech rate and the quality of pronunciation. In case of serious mispronunciations or easily audible differences of speech rate, the recordings of the faulty utterances were repeated until an acceptable version was obtained. Speakers with speech defects (like stuttering or lisping) were not recorded.

The recorded speakers come from 17 regions of Poland. The choice of the regions was motivated by the will to cover all potential varieties of Polish anticipating the possible differentiation among the future end-users of the designed ASR system. Although Poland is regarded as comparatively homogeneous in terms of dialectal diversification we decided to gather material from all over the country with a view to train the acoustic models with as rich material as possible to minimize future errors. The most numerous region was Wielkopolska (both for the reason that the research laboratory is situated in Poznań, Wielkopolska, and also because of the close cooperation between the Wielkopolska Police Head Quarter and the laboratory).

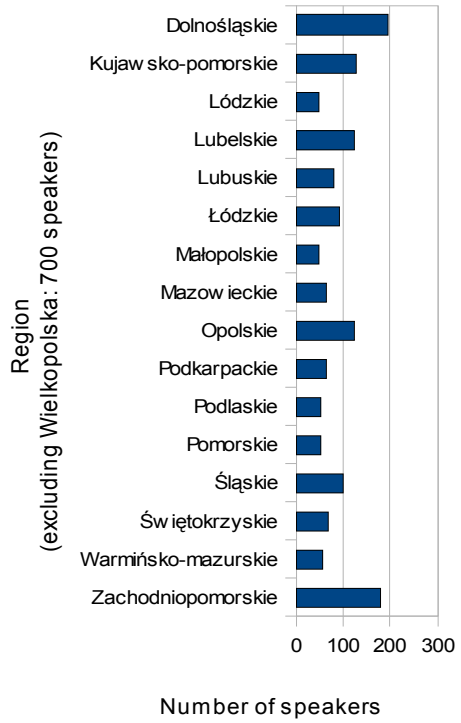

Figure 2: The number of speakers recorded in 17 regions of Poland (the most numerous region Wielkopolska with 700 voices was not included in the chart[2])

2.2. Text corpora

The text scenarios for read speech recordings belong to two main categories. The first category are texts for the coverage of phonetic and syntactics structures. These were constructed especially for the purpose by a team of phoneticians (Demenko, et al. 2008), we also used one subset from the Polish speech synthesis Boss corpus as a supplement for missing consonant clusters (eg. Demenko et al. 2007).

Another important category of the read text sources were original texts obtained from the police in the form of police reports from police interrogations, and also legal texts provided by legal offices (expert opinions,

[2]Note: the voices obtained in Wielkopolska were obtained not only from the native habitants but also from student residents from outside the region

judgements, court decisions etc.). The use of the latter, again, was aimed at supplying dedicated vocabulary and higher level text structures semantically covering the language of the system's end-user.

The JURISDIC text corpora provide the following coverage of triphones in the scenarios of read text created within the framework of the present project: within-word triphones: 10593, triphones containing an accented vowel: 8492, unaccented triphones 10650, phrase-final triphones: 4495. The number of triphones is substantial, however so far, the general triphone statistics for Polish were not investigated in a form enabling an informative comparison or permitting to provide the actual percentage of the total number of triphones covered by the present data. Another subset of texts was obtained as a result of annotation of the acoustic data giving a collection of transcriptions of the spontaneous and semi-spontaneous speech.

2.3. Annotation information

The starting point for annotation specification applied for the present corpus were SpeeCon annotation guidelines (Fischer et al., 2000) based on orthographic, word-level transcription. In the first step, annotators (a team of students of The Faculty of Modern Languages and Literature in Poznań, above thirty people during the whole period of the annotation process) manually validated the agreement of the recorded text with the input orthographic transcription by inserting necessary adjustments, special events markers, and time boundaries. The first-step annotations were hand-validated (if necessary) by two expert phoneticians and four experienced labelers for whom the inter-labeler agreement was monitored, especially as concerned the number and types of special events and time boundary insertion and spelling errors (Klessa, Bachan, 2008). The inter-labeler agreement concerning the time boundaries was high (above 90%), the agreement for the special events labels depended on the type of label and was best for the unintelligible speech markers (above 80%) and filled pause labels (approx. 70%). It was lower for speaker noise labels and mispronunciation markers because of a greater variation observed for one of the labelers, after excluding the results for that labeler, the agreement was up to 70%. For the purposes of acoustic modeling, the speech data were labeled on the phone level via an integrated automatic segmentation tool Salian (Szymański, Grocholewski, 2005) according to rules used also in Polphone (Demenko et al., 2003). Annotation Database Manager supports also fully manual validation of the automatic labeling (however so far, the phone level segmentation was performed only automatically with Salian, the hand-validation is planned in the near future).

The annotated database is composed of three major sub-corpora including:

- "Police&Office" - read and semi-spontaneous speech including three main types of text: type A – (semi-)spontaneous speech: elicited dictation of short descriptions, isolated phrases, numbers or letter sequences; type B – read speech for phonetic coverage and syntactically controlled structures; type C – read speech: semantically controlled structures, application-specific vocabulary.
- "Court"- spontaneous speech of judges recorded during real court trials in courtrooms
- "Lawyer" - read speech covering specialized language recorded in lawyers' offices.

2.4. Recording conditions and acoustic database

The acoustic data were obtained with an especially designed software using two microphones (a close-distance headset microphone and a table microphone), giving two-channel recordings. The recording sessions took place in a typical office environment (silent, but non sound-proof rooms) similar to the planned end user environment. More technical details of the recording setup are given in Demenko et al. 2008.

The "Police&Office" corpus, which is the fundamental and the richest one, delivers the total number of 478579 utterances from 1369 speakers (784 hours of speech). Its largest subsets contain utterances for grammatic and phonetic coverage, and dictation task texts. An important percentage belongs also to read sentences taken from original police reports. The "Lawyer" sub-corpus (above 12 thousand utterances read by 158 subjects) is to some extent supplementary for the "legal text" subset of the "Police-Office" sub-corpus, produced with a view to enhance the quality of the dictation system for lawyer's offices. The total duration of the "Lawyer" sub-corpus is 56 hours. The "Court" sub-corpus is the smallest one (the total of 4342 utterances, 33 speakers, total duration: 15 hours), however from certain point of view it might be very interesting since it contains spontaneous speech recorded in the real-life situations. The speech is spontaneous in the meaning that it is not elicited or controlled by the experiment, however it is produced by judges speaking formally, in public thus it is expected to be comparably well-formed and not excessively expressive. A certain drawback for the courtroom recordings might be the fact that the noise level is significantly higher than in the remaining sub-corpora, however the noises are an inevitable circumstance for data recorded in such type of environment.

Apart from the annotated sub-corpora the JURISDIC database contains approximately 577 additional recording sessions (total duration: 300 hours). Part of these sessions are the product of the preliminary recordings, part of them was obtained in the environment of higher noise levels.

2.4.1. Recorded word statistics

The total number of word occurrences in the annotation files is more than 5 200 000 (excluding special events markers).

The length, and so the average numbers of words in the recorded utterances vary across the JURISDIC sub-corpora. The average number of words per phrases categorized by the database sub-corpora are depicted in Figure 3. The longest utterances were delivered by the specialized, semantically oriented read speech sub-corpora. The recordings of spontaneous speech from lawyers and attorneys also appeared to be comparably long on average. The semi-spontaneous speech section

contained mostly elicited dictation of separate words or short phrases.

2.4.2. Special events and noise statistics

Apart from ordinary text the annotations contain several special events labels: four types of noises (speaker noise [spk], filler [fil], intermittent noise [int], stationary noise [sta]); in cases when an event was present only in one of the recording channels an additional index for channel information may be attributed to the noise markers ([int:1] - intermittent noise in channel one, [int:2] - intermittent noise in channel two). (In the statistics presented in this study the channel information was ignored with a view to simplify the presentation of the results.). Mispronunciation or unintelligible speech markers, and wave file truncation markers are also applied.

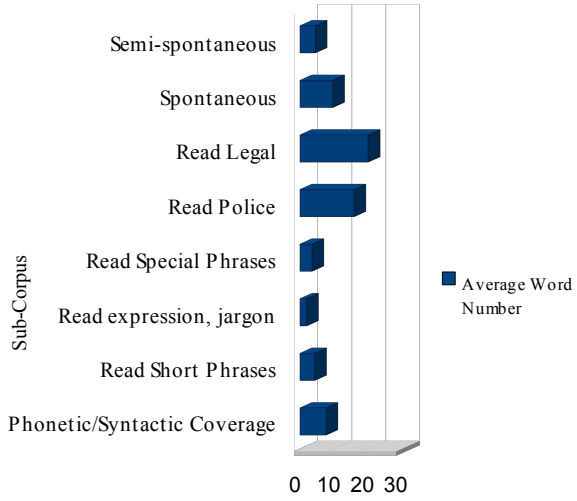

Figure 3: Average number of words per phrase for component sub-corpora in JURISDIC

The first observation for the special event and noise statistics is that the average number of the speaker noises is significantly higher than the respective average number of fillers for the read texts. For the semi-spontaneous, dictated utterances and (particularly) for the spontaneous courtroom utterances the tendency was the opposite: significantly more fillers were annotated.

The highest number of the speaker noises was noted for the domain-specific read texts, especially for the "Lawyer" sub-corpus. The sentences might have been difficult for the readers because of two reasons: the specialized vocabulary, complex syntactic structures, and comparably long sentences.

Not surprisingly, the number of fillers was highest for the spontaneous and semi-spontaneous speech (i.e. the courtroom recordings and dictation tasks from the "Police and Office" sub-corpus). For the read speech data the number of fillers was comparably high for the domain-specific texts and also for the longer phonetic coverage sentences. This may be explained by the difficulty of terminology and also by the presence of rare words (used in the sentences for high within-word triphone coverage).

While observing the proportion of mispronunciation and unintelligible speech markers in the data, the highest values were noted for the court recordings and for the read speech for the texts containing specialized vocabulary and structures. The numbers for the dictation tasks were lower which may indicate that although speakers hesitated they were quite intelligible in the more controlled, semi-spontaneous speech. The high occurrence for the domain-specific texts may again confirm their difficulty and problematic character. It may also be an information about the linguistic labelers competence - in case of difficult (or unknown) words they might have been inclined to insert mispronunciation or unintelligible speech markers.

As it was expected, the proportion of environment noises ([int] and [sta]) in the courtroom recordings is substantially higher than in the remaining corpora recorded in more controlled conditions.

3. Emergency Calls Database: Expressive Speech Corpus

An important supplement for the controlled contents of the JURISDIC database is the database of original recordings from the police emergency telephone number. It complements the JURISDIC database with information from spontaneous speech which is also to be considered in the ASR domain. As a total the corpus contains recordings from up to 8 thousand phone numbers, and above 30 thousand recordings of particular conversations (some phone numbers repeat in the records). The recordings included in this corpus are characterized by lower quality in terms of acoustics (phone conversation characteristics), however they are a valuable source of information on original emotional, expressive speech, including extreme emotions, hard to observe in any other recording conditions. The process of analyzing the data is in progress, the software used is the same as for JURISDIC, the same type of SQL database is used to store annotations (the data are stored separately from other corpora for security reasons). The annotation specification was formulated based both on the experience with JURISDIC regular data and also on the previous research in the domain of emotional speech and dialogue analysis (e.g. Cornelius, 2000, Devillers, Vidrascu, 2006, Karpiński, 2006).

3.1.Expressive Speech Description

Descriptors used in the annotation of the emergency calls expressive speech are defined in the following categories: dialogue acts, suprasegmentals, time/situation contexts, lexical events, non-verbal features, and background acoustics.

- Dialogue acts are divided into eight categories, which still is a kind of simplification of the complexity and multi-dimensionality of dialogue analysis. The first two categories are related to action commands or offers by interlocutors. Another two categories are connected with information transmission (providing or asking for information). Expressions whose primary use is to carry emotions are included to the next

category. Another group are *informal announcements*. Utterances for dialogue turn manipulation, conventional phrases are categorized as *dialogue management*.

- Suprasegmental descriptors are considered in terms of accent, intonation, rhythm, and loudness .
- *Context* is described in the dimension of *situation* (e.g. threat, complaint, demand...) or *time* (immediate, passed, potential).
- *Lexical descriptors* are marked in case of interjections, offensive expressions, modulants or neologisms.
- Non-verbal features are described using these categories: *emotions* (graded on a five step scale in the dimensions of valency, potency and arousal), e*xtreme emotions, fillers* and *speaker noises.* The last non-speech events category is *voice quality.*

The character of the above categorization is preliminary and will be the subject of more detailed analyses in the near future.

5. The future work and conclusion

Data collection, construction and annotation should be closely tied to the needs of dedicated application/applications to decide what kind of information would be needed. Rich annotation of the corpus on various levels of analysis as well as higher level phonetic evaluation procedures would be useful (i.e. prosodic features).

The planned future work is thus further examination of the obtained data, a sophisticated, language-oriented parsing of the annotation data giving information about the linguistic factors on various levels of analysis, as well as further identification of other possible error types. Two kinds of tools for the post-hoc semi-automatic evaluation of annotation are needed: a tool for random selection of annotation subsets enabling hand-validation supported by statistical analysis of spelling and more detailed analysis inter-labeler agreement (in preparation); a tool for an overall comparison between the annotation word lists and the reference lexicons (ready to use). The first future step will be the analysis of the results of that comparison.

One of the premises for the present Polish speech corpus construction since the early stages of the corpus design was creating a rich linguistically annotated speech corpus with a view to significantly improve the results of acoustic modeling as well as (to a lesser extent) language modeling.

Considering the costs and effort put into the collection and annotation of the speech data it was assumed that the corpus should be re-usable. Its main use is the optimization of automatic speech recognition for dictation system using acoustic models depending on speech rate and accent. Additionally it was intended to provide material for speaker identification and speaker characterization in terms of age, sex and dialect.

References

Cieri, Ch., Miller, D., Walker, K., 2004. The Fisher Corpus: a Resource for the Next Generations of Speech-to-Text. Proceedings of Language Resources and Evaluation Conference (LREC), published on the internet, accessed on 12, April 2009: http://papers.ldc.upenn.edu/LREC2004/LREC2004_Fisher_Paper.pdf

Cornelius, R.R. (2000). Theoretical approaches to emotion. Proceedings of the ISCA Workshop on Speech and Emotion (pp. 3–10). Northern Ireland.

Demenko, G., Grocholewski, S., Klessa, K., Ogórkiewicz, J., Lange, M., Śledziński, D. & Cylwik. N. 2008. Jurisdic–Polish Speech Database for taking dictation of legal texts. In: Proceedings of the Sixth International Language Resources and Evaluation, Ed. European Language Resources Association (ELRA), Marrakech, Morocco.

Demenko, G., Wypych, M., and Baranowska, E. 2003. Implementation of Grapheme-to-Phoneme Rules and Extended SAMPA Alphabet in Polish Text-to-Speech Synthesis. In: Speech and Language Technology, vol. 7. [Ed.] PTFon, 79-97, Poznań.

Devillers, L., Vidrascu, L. 2006 Real-life emotions detection with lexical and paralinguistic cues on Human-Human call center dialogs. The Proceedings of Interspeech Conference.

ELDA - homepage accessed on 10, April, 2009: http://www.elda.org/

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

Gibbon D., Moore R., Winski R., Handbook of Standards and Resources for Spoken Language Systems, deGruyter, 1997.

JURISDIC site: http://www.speechlabs.pl/en/project_ppbw/info

Karpiński, M. (2006) Struktura i intonacja polskiego dialogu zadaniowego. Wydawnictwo Naukowe UAM, Poznań

Klessa, K., Bachan, J. 2008. An Investigation into the intra- and inter-labeler agreement in the JURISDIC database. 2008. Speech and Language Technology, 2008, vol. 11.

Klubiński, M. Wyszukiwanie w repozytoriach tekstowych w języku polskim (Eng.: Searching text repositories for the Polish language), available on the internet, accessed on 10, April 2009: http://home.elka.pw.edu.pl/~mkozlow3/artykuly/M.Klubinski.pdf

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, Interspech 2007, 2145-1249.

Maier, E.,Mc Glashan, S., 1994. Semantic and Dialog Processing in the VERBMOBIL Spoken Dialog Translation System. In: Heinrich Niemann, Renato de Mori and Gerhard Hanrieder, Publisher, Progress and Prospects of Speech Research and Technology, CRIM/ FORWISS Workshop, p. 270-273, Infix Verlag, Munich 1994.

SPEECON homepage: http://www.speechdat.org/speecon/index.html, accessed on 10, April 2009.

Szymański, M. & Grocholewski, S. 2005. Semi-Automatic Segmentation of Speech: Manual Segmentation Strategy. Problem Space Analysis. In: Advances in Soft Computing, Computer Recognition Systems, 747-755, Springer Berlin.

Transcriber website, accessed on 10, April 2009: http://trans.sourceforge.net/

Wavesurfer website accessed on 10, April 2009: http://www.speech.kth.se/wavesurfer/

Ziegenhain, U. et al. 2002. Specification of corpora and word lists in 12 languages. LC-STAR Deliverable D1.1.