INTERSPEECH 2009 BRIGHTON

Structure and Annotation of Polish LVCSR Speech Database

Katarzyna Klessa[1], Grażyna Demenko[1,2]

[1]The Institute of Linguistics, Adam Mickiewicz University, Poznań, Poland
[2]Poznań Supercomputing and Networking Center, Polish Academy of Sciences, Poznań, Poland

klessa@amu.edu.pl, lin@amu.edu.pl

Abstract

This paper reports on the problems occurring in the process of building LVCSR (Large Vocabulary Continuous Speech Recognition) corpora based on the internal evaluation of the Polish database JURISDIC. The initial assumptions are discussed together with technical matters concerning the database realization and annotation results. Providing rich database statistics was considered crucial especially regarding linguistic description both for database evaluation and for the implementation of linguistic factors in acoustic models for speech recognition. The assumed principles for database construction are: low redundancy, acoustic-phonetic variability adequate to dictation task, representativeness, balanced, heterogeneous structure enabling separate or combined modeling of phonetic-acoustic structures.

Index Terms: language resources, large corpora annotation, standardization and evaluation

10.21437/Interspeech.2009-529

1. Introduction

Text and speech language corpora are indispensable to meet the challenges of various contemporary speech technology applications. The importance of creating the resources was already confirmed by research and in practice thanks to several existing large language corpora (Vermobil [1], SpeeCon [2]). One of the largest existing corpora is English 'Fisher' Corpus delivering 2300 hours of transcribed conversational telephone speech, and 75000 vocabulary items that covers training data (eg. [4]). An excellent overview of the available resources is offered by ELDA [3]. However, with a view to broaden the scope of reusability of the corpora it would be of interest to obtain detailed knowledge of the structure, evaluation and validation of the proposed corpus solutions.

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 and but also to the linguistic structures.

Obviously, it is not possible to comply with 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. An ideal factor analysis for the set of approximately 2000 speakers should bring about repeatable results at least for a group of several speakers. 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 were also commenced with ELDA [3].

The structure of the present paper is as follows: Section 2 and 3 report on general characteristics of the database structure accompanied by statistical information about the collected speech data and annotation files. Sections 4 and 5 provide the discussion and comments of the findings and the future prospects for the optimization of database realization and specification.

2. JURISDIC database structure

The JURISDIC speech database is a large continuous speech database for speech recognition, probably the largest one for the Polish language available at the moment (above 1500 annotated sessions of speakers from 16 regions of Poland, plus another 500 experimental recordings, for more details cf. [5, 6]).

The JURISDIC database is intended to provide material for both training and testing of speech dictation of common and legal texts, including isolated word systems, word-spotting systems and vocabulary independent systems which use either whole word or sub-word modeling approaches. The JURISDIC common specification is a mixture of semi-spontaneous (controlled dictation) and read/dictated 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. 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 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 meet international standards for language resources [2, 3, 7] with necessary adjustments resulting from the language specificity, the future applications and the substantial size of the database.

2.1. Data quality and management

The database delivers two-channel speech data recorded with two microphones (a headset and a table one) using software created specifically for the purpose (with the exception of the "Court" sub-corpus which was recorded through one channel with only one, headset microphone). The recording sessions took place in a quiet office environment. The speakers have at least secondary school certificate and were instructed to speak at a moderate speech rate, neutral style, as in a dictation task.

All annotations and recording scenarios as well as speaker and labeler information are stored in a SQL database and at the stage of annotation may be accessed via an especially created software enabling comfortable management of speech, text and user data - Annotation Database Manager. The Manager was designed using the Client-Server architecture based on MSDE 2000, and Windows 2003 Server, the client applications were programmed in C# programming language. The Manager is multi-functional, it is additionally equipped with a lexicon search engine available for the labelers as a spelling reference. The project's lexicon created according to standards [8, 9] contains: common words, application words and proper names (260000 entries). Additionally, a frequency lexicon (450000 entries) was included to complete the vocabulary coverage. The Manager supports creating working lexicons based on the annotation text data and then to provide comparative analyses with the reference lexicons (e.g. in order to supplement the reference lexicon with new items from the annotation data). The Annotation Database Manager is integrated with the following external tools: Transcriber [10], Wavesurfer [11], Salian [12] allowing controllable operation on the respective data formats which is important, taking into consideration the huge amount of data and multi-user character of the system.

2.2. The structure of the annotated corpora

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

- "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. The coverage of triphones in the scenarios of text type B read is as follows: within-word triphones: 10593, triphones containing an accented vowel: 8492, unaccented triphones 10650, phrase-final triphones: 4495. The number of triphones in the present text corpus is significant however so far, the general triphone statistics for Polish have been investigated only for fragmentary data not sufficient to provide an informative comparison or give the actual percentage of the total number of Polish triphones covered by the present data.
- "Court"- spontaneous speech of judges recorded during real court trials in courtrooms.
- "Lawyer" - read speech covering specialized language recorded in lawyers' offices.

2.3. Recordings statistics

Table 1 presents the number of recorded and annotated utterances within the three major sub-corpora of the database: "Police&Office" with sub-categories (read and semi-spontaneous speech), "Lawyer" (read speech) and "Court" (spontaneous speech) (cf. paragraph 2.2 above, and [5]).

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 (Table 1) 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 or from speakers with speech disorders, a subset of about 200 sessions are one-channel recordings of read speech. Altogether, taking the experimental recordings into account, the database comprises speech provided by 2137 speakers (total duration of the whole database: above 1155 hours).

3. Annotation

3.1. Annotation procedures

The starting point for annotation specification applied for the present corpus were SpeeCon annotation guidelines (deliverable D214 [13]) 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.

Table 1: *The number of annotated utterances and the total duration of JURISDIC sub-corpora.*

Sub-Corpus	Description	Duration (hours)	Annotated Utterance Count
Police & Office	Semi-spontaneous speech, dictation	111	92891
	Read B1 - phonetic and grammar structure coverage: longer, complex sentences	134	83160
	Read B2 - phonetic and grammar coverage: short sentences	113	124371
	Read B3 - common short phrases, bigrams, pause contexts	5	9540
	Read CR - police reports	224	88579
	Read CC - legal texts, comparably long, sophisticated sentences	143	35722
	Read C - diversified vocabulary and phrases for lexical coverage	54	44316
Lawyer	Legal texts - specialized formal texts	56	12012
Court	Courtroom recordings, spontaneous speech	15	4342
	Total:	**855**	**494933**

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 [14]. 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 [12] according to rules used also in Polphone [15]. 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).

3.2. Annotation information

The annotation files may serve as a rich source of information about the database contents, most importantly about the speech signal files, but also, somewhat indirectly, about the

speakers strategies (also the annotators habits) as related to text types or to specific tasks (read versus spontaneous/dictation speech tasks). By the speaker strategies we mean diversification of the number and specificity of fillers, speaker noises and other speech and non-speech events depending on the speech task, as well as the differences in phrasing for the spontaneous and semi-spontaneous speech. As for the annotators habits - the data might be useful for analyses and comparisons of perceptive analyses since although the annotators are given the same guidelines and specifications it is still expected that they will differ in their judgments (cf. [14]).

3.2.1. Word statistics

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

The list of the most frequent vocabulary items found in the annotation files were compared to a list of Polish stopwords [16]. 62 items out of 110 were exactly the same in the two lists, several items differed only by inflection ending. The more serious differences were connected partly with the fact that the reference list contained more inflected forms of particular items (e.g. the inflected forms of the word *który*, Eng. *which/that*). Instead of the elaborate inflection range for the most frequent words the present top-frequency list contained the domain-specific vocabulary (e.g. *artykułu* (the genitive form of the word *artykuł)*, Eng. *article* or *paragraf*, Eng. *paragraph*, and also numbers or address words).

Generating the annotation frequency list allowed also to identify the probable spelling errors (words having the lowest frequency are the suspect group). The list of possibly misspelled words obtained in this way enclosed 1,3 percent of the whole word list.

3.2.2. Special events and noise labels

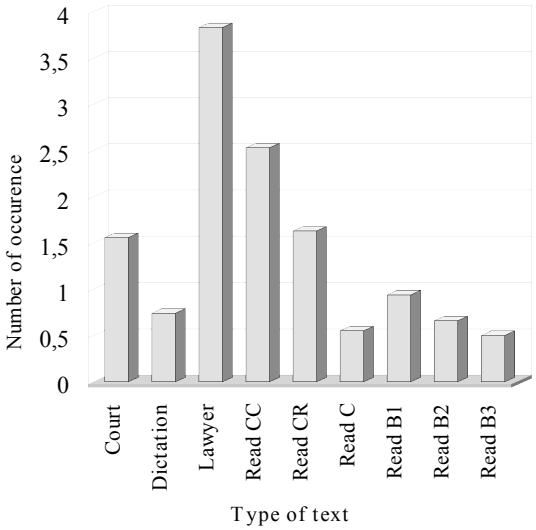

Figure 1: *The average number of speaker noises annotated per utterance in each of the three main database sub-corpora.*

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
- wave file truncation markers

Figures 1 and 2 depict the number of fillers and speaker noises for various text types within the three database sub-corpora. Since the "Police & Office" corpus is quite complex in terms of the component text types, they were shown as separate bars in the figures. The abbreviations are explained in the "Description" column in Table 1.

The first observation 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.

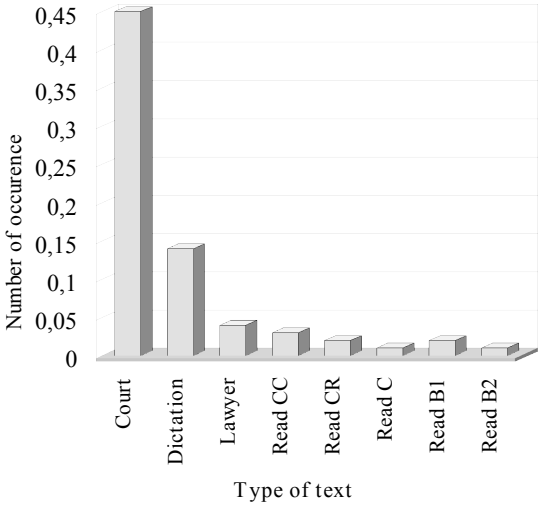

Figure 2: *The average number of fillers annotated per utterance for particular text types.*

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. Figure 2 shows, not surprisingly, that the number of fillers was highest for the spontaneous and semi-spontaneous speech (i.e. the courtroom recordings and semi-spontaneous dictation tasks from the "Police and Office" sub-corpus). The mean number of fillers was practically zero for B3 texts, that is why the respective bar is absent in the Figure 2.

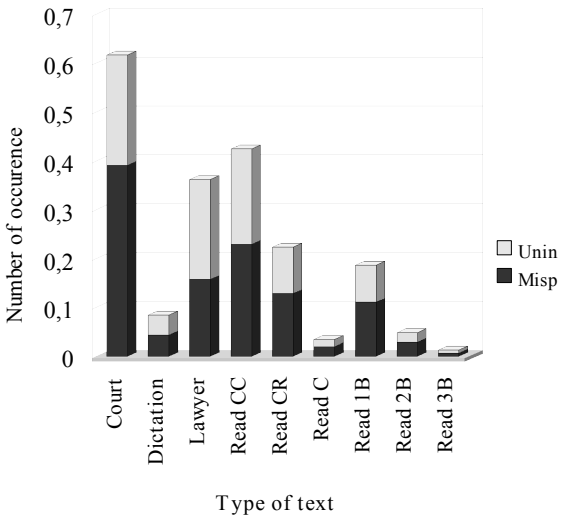

Figure 3: *The average number of mispronunciation and unintelligible stretch of speech markers (dark gray bars - mispronunciation, light gray - unintelligible).*

For the read speech data the number of fillers was comparably high for the domain-specific texts and also for the longer and more complex 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).

The proportion of mispronunciation (*Misp* - mispronunciation - dark gray bars) and unintelligible speech markers (*Unin* - light gray) is presented in Figure 3. 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 are lower which may indicate that although speakers hesitated (cf. the fillers occurrence) 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 indication as to the labelers' linguistic competence - in case of difficult (or unknown) words they might have been inclined to insert more 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 "Police & Office" or "Lawyer" corpora recorded in more controlled conditions.

Table 2 presents the number of annotated noises for the three major sub-corpora. The higher number of [int] label for the more controlled conditions may be the result of inserting it to mark a louder mouse click at the very end of recording ([int:2] marker was used for this purpose), this usage of label is comparably easy to identify since it usually occurs within the final silent time section (separated with time boundaries).

Table 2. *The average number of intermittent and stationary noises annotated per utterance for particular text types.*

Sub-corpus	[int] noise number	[sta] noise number
Court	0,67	1,43
Lawyer	0,24	0,08
Police & Office	0,12	0,05

4. Discussion and future work

The presented observations for the number of special speech events are a sample of statistical analyses conducted for the JURISDIC database. This kind of data mining together with recording statistics may be useful not only for the reasons of the evaluative description of the database but may also serve as a starting point for more general investigation of the speaker strategies depending of the experiment task.

Data collection, management 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 corpus: not only linguistic elements (words, triphones) but also phonetic evaluation would be useful (prosody, speech style, pronunciation types etc).

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.

5. Conclusions

It is proposed to modify the evaluation and validation criteria for LVCSR corpora. Apart from the assessment of recordings statistics it is regarded as equally important to evaluate linguistic information, and also the information about the non-speech events and other phenomena occurring in the collected data. Moreover, the rich linguistically annotated speech corpus would improve significantly acoustic models.

The above postulate was one of the premises for the present Polish speech corpus construction since the early stages of the corpus design. Considering the costs and effort put into the collection and annotation of the speech data it was assumed that the corpus should be reusable. Its main use is the optimization of automatic speech recognition for dictation system using acoustic models depending on speech rate (slow, medium, fast) and accent (accented, non-accented, phrase final triphones). Additionally it was intended to provide material for speaker identification and speaker characterization in terms of age, sex and dialect. 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. Studies concerning the specified fields have already been commenced or are planned in the near future, however for lack of space their description cannot be elaborated in the present paper.

6. Acknowledgements

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"

7. References

[1] 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.

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

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

[4] 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

[5] 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.

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

[7] 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.

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

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

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

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

[12] 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.

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

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

[15] 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.

[16] 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