CORPORA - SPEECH DATABASE FOR POLISH DIPHONES

S. Grochowelski
Institute of Computing Science
Poznań University of Technology
ul. Piotrowo 3a, 60-965 Poznań, Poland
email: grochow@poznan.put.poznan.pl

ABSTRACT

In the paper the attempts for creating the first databases for Polish are presented. Among two databases, supported by The Polish National Research Committee, and COPERNICUS project (1304 "BABEL: A Multi-Language Database" for Polish, Bulgarian, Estonian, Hungarian, Romanian) the first of them is presented in detail. The speech material contains 365 utterances (alphabet letters, digits, 200 first names, 114 sentences) uttered by 45 speakers. In the paper the design ideas, recording conditions, annotation rules, the method of automatic segmentation and labelling used in CORPORA are presented.

1. INTRODUCTION

There is a great need for large, segmented and labelled databases for speech synthesis, speech recognition and phonetic research purposes. Unlike for many other languages, there is no any large database for Polish. Currently, there are two projects leading to such a database: the project supported by The Polish National Research Committee, and COPERNICUS project (Project 1304 "BABEL: A Multi-Language Database" for Polish, Bulgarian, Estonian, Hungarian, Romanian).

The first one, CORPORA", is conducted in Institute of Computing Sciences in Poznań University of Technology, and will be described in details in this paper. The second one - in Speech Acoustics Laboratory in Institute of Fundamental Technology Research in Warsaw, and is intended for preparing the speech data for EUROM_1 database [1]. The second one contains 10 digit, 100 numbers (selected such that the main phonotactic possibilities of the language number system were covered), 40 short passages each containing some thematically connected sentences, 35 sentences composed to compensate for the phoneme-frequency imbalance in the passages. Speech signals are digitized at a sampling frequency of 20 kHz and 16 bit resolution.

The disk space needed for one speaker is about 40 MB.

It is worth mentioning that a new consortium, called SpeechDat(E), is being formed. The aim of this consortium is to collect speech databases of East European languages [2]. SpeechDat databases are focused on the telephone environment including fixed and mobile telephone networks.

2. THE DESIGN OF CORPORA

2.1. Speech material

The speech material was selected to satisfy two requirements: to include Polish letters, digits and 200 first names uttered in isolation, and to include minimum 900 Polish diphones according to their frequency of occurrence in Polish. A study of the statistical properties of diphone distribution in Polish was undertaken on a corpus of spontaneous speech from radio recordings [3]. To satisfy the second requirement we introduced 114 sentences which with the above mentioned letters, digits and first names give 966 diphones. The list of diphones can be found in [4]. In practice the number of diphones will be greater because during the recordings the unexpected pronunciations were noticed.

2.2. Recording conditions

The recordings were made in different regions in Poland, in computer room environment, using the same type of equipment in each centre (the real time PC based digital spectrographs VOLYZER elaborated by author at al. [5], currently accessible in almost all Polish phonetic laboratories). Speech signals were digitized at a sampling frequency of 16 kHz and 13 bit resolution. The signal was normalized and only 12 most significant bits were used. To minimize the errors the recording process was controlled by the specialized software package TIMREC. The sentences to read were being displayed on the screen in sequence, the recording was being disabled when the signal level was too
low or too high prompting the speaker to repeat the utterance, the endpoints were automatically identified, the file names were automatically generated according to the prompt text, etc.

Each recording session was supervised by the qualified phonetician. Examining the real time [5] spectrogram, waveform and selected spectra of the utterance (see Fig.3) he could order to repeat the sentence as well as to correct the automatically identified endpoint of the sequence.

The disk space needed for one speaker is about 13 MB. We recorded 45 speakers on one CD-ROM (28 male, 11 female and 6 children voices). Totally 16,425 utterances have been collected. They contain almost 180,000 Polish diphones.

3. SEGMENTATION AND LABELLING

We decided that all the recordings should be segmented and labelled according to the phonemic transcriptions of the utterances.

Manual segmentation and labelling is an extremely time consuming and tedious process. We considered several approaches to the automatic segmentation and labelling of the recordings [6], [7], [8], [9]. In [7] it has been shown that an automatic labelling system, originally trained on a multi-speaker continuous speech database of utterances in one language, can be automatically updated to label utterances in other languages without the need for any hand labelled data representative for this new language. Although the authors obtained 24-28% segmentation errors they hope to decrease the percentage of segmentation to something like 10-15% in the near future.

In our case we considered two possibilities. The first one was to build the HMM models for all Polish phonemes based on the reference annotations produced by a trained phonetician, as in the case of the COPERNICUS project directed by Gubrynowicz [10]. The second one consisted in using the nonlinear warping between the labelled utterance and the reference utterance annotated manually. Because of the simplicity of the second approach we performed two reference annotations (for male and female voices) prepared manually by trained phoneticians and then we tried to obtain the optimal annotations using the dynamic programming technique.

In the Fig. 4 there is the screen of our program LABELER. At the bottom there is a manually segmented reference utterance "tim mojna atakovac". The markers at the top spectrogram have been automatically optimally aligned using dynamic programming with the reference markers. The acoustic analysis was performed every 5 ms using 15 ms Hamming window. Each frame represented 100 point log spectrum based on the 240 point WFTA (Winograd Fourier Transform Algorithm) [5].

This approach worked quite well and therefore we elaborated in this way the annotations for all the 16,425 recordings (45 speakers x 365 utterances per speaker). Since our database should be treated as the reference database, all the annotations have to be visually inspected and, if necessary, manually corrected.

4. ANNOTATION RULES

To automatically annotate all the recordings we had to prepare two reference annotations. We chose the recordings from one male and one female speaker for manual annotation to be performed by professional phonetician. The rules depend on what is the intended use of the annotated speech material. Since the CORPORAS is the first Polish database we started with phonemic transcriptions.

The most frequent problems were connected with the cases when acoustical features of the short parts of the utterances didn't correspond to what was heard. In the Fig. 1 the isolated segment "ej" is heard as "i" whereas in the whole word "wołejś iepce" it is heard as "ej".

![Fig. 1. Segment "ej" is heard in isolation as "i".](image-url)
In such a case we labelled this part as "ej" whether acoustic cues were seen in the waveform or not. The same problem was with two immediately succeeding plosives with no visible release in the first one. When two plosives are heard the boundary is placed randomly in the silence area, as in the Fig. 2 in the case of "kot psotnî".

Fig 2. Although 't' is not visible on the spectrogram it is audible and therefore it is marked randomly in the silence area in the label string.

The detailed annotation rules can be found in [4]. Each waveform file in *.wav format is accompanied with the appropriate text file with the phonemic labels. One character in the text file (phonemic label) corresponds to 5 msec time segment.

5. DATABASE FACILITIES

Along with the data files, there are several database tools which provide means for searching the required speech units in time and frequency domain. There are also appropriate programs for conversions from *.wav files to *.hik, as well from *.fon (transcriptions files) to *.lab files [11].

6. CONCLUSION

In the paper the annotated speech database for Polish is described. It contains almost 6 hours of speech uttered by 45 speakers. By selecting almost 1000 Polish diphones we intended to capture the majority of coarticulation effects in contemporary Polish.

7. ACKNOWLEDGMENTS

This research was supported by The Polish National Research Committee under grant nr 8 T11C 023 08.

8. REFERENCES


Fig. 3. Real time spectrogram, waveform and selected spectra were displayed during recordings.

Fig. 4. The LABELER's screen. At the bottom there is a manually segmented reference utterance "tim można atakować". The markers at the top spectrogram have been automatically optimally aligned using dynamic programming with the reference markers.