Contents

Table of Contents

Introduction

This database collection is a Polish speech database collected by Voicelab company over the microphone.
The presented corpus contains recordings of 263 424 utterances of Polish speech data which were obtained from 200 speakers.
The overall recording time is about 279 hours.

Speech File Format

The utterance are stored as 16 Khz, 16 bit, Mono channel, and uncompressed.
All files are stored in *200f* directory.

Directory Structure

The three-levels directory structure is defined as /200f/SpeakerID/SessionID
where each part is defined in Table 1-1.

SpeakerID	number from 0 to 9999
SessionID	Defined as VLRecordingSpeakerID_date_id

Table 1-1: Desktop Speech Directory Structure

The files are under SpeakerID folder, their names are defined as VLRecordingSpeakerID_date_id.,
with each file the speaker-folder contains a transcription file named VLRecordingSpeakerID_date_id.txt
in UTF-8 format.

Database Design and Collection

The Voicelab project of collecting a large amount of data in the Polish language resulted in about 4 000 hours of speech recorded by more than 3 000 people. People recorded in their home using desktop computers with headsets (real data) and a website. We created a corpus to cover most of the speech sounds in Polish. For the whole project the phonetic balancing has been done on the triphonic level through a large computational effort which produced a corpus of around 50 000 sentences which were chosen from around 1bln (10^9) sentences. The resulting 50 000 sentences were manually reviewed.

In this document we present 280 hours of speech data which we collected during our project.

Recording Devices

For recording, people used desktop computers with headsets.

Speaker Recruitment

The entire collection was performed in Poland. An email database was used in the recruitment process. Speakers were to record in their home for at least 60 minutes.

Database Design

The described database contains files with a total length of 1008850 sec.

Transcription

Speakers had been asked to record their voice while reading text on the website. The main task was to validate the quality of that transcription. This was possible with a special quality tool developed by Voicelab. For the whole project we hired 6 people for quality checking.

Speaker Demographic Information

Gender Balance

The database consists of 103 male speakers and 97 female speakers.

Age Distribution

For this project, speech data were collected in the following age categories:

Age group	# Speakers	# Speakers (%)
15 – 30 years	171	85.5%
31 – 45 years	25	12.5%
46 – 60 years	4	2%

Speakers' Age Distribution

Appendix A. Speaker Information

Average length describes the sum of all files for each speaker divided by the number of files.

Total time sec is the parameter describing the total time recorded by each speaker.

Avg snr is the average speech noise reduction parameter counted by a specific algorithm.

avg active speech is a parameter describing the relation between the active speech time and the total file length in percentages**.**

gender "m" represents men, "k" represents women.