Russian telephone speech corpus TeCoRus.

TeCoRus has been collected to provide the developer with Russian speech material to work on the speech technology, primarily, in the field of telecommunication. The speech corpus is designed to obtain a controlled amount of speech for training conventional statistic context-dependent phone models. The corpus contains enough data to support the development and testing of the algorithms for noise cancellation, channel equalization, assessment of speech technology system performance and so on.

TeCoRus material consists of two parts

1. The phonetic part. The primary target of the phonetic part is to provide phonetically rich speech material to build the boot set of context-dependent phone models. This part contains 3050 utterances read by 6 speakers. Each utterance is recorded into a separate audio file accompanied with an annotation text file. The annotation contains the corresponding orthographic text, phonetic transcription and phonetic segmentation of the utterance.

The speech material of the phonetic part is recorded into two channels:

- narrowband telephone channel (4kHz, Moscow fixed telephone network);
- broadband microphone channel (11kHz computer Koss SB35 mike).

The size of the phonetic part is 600Mb.

2. The speech part of the corpus is a collection of the short answers extracted from the interviews and also includes both read and spoken material (the controlled answers and spontaneous speech on the preselected topic). Specifically, the speech part of the corpus contains:

- one-word positive and negative answers
- digits (isolated and digit strings, consisting from 2 to 16 digits)
- numbers
- names of months and weekdays
- Russian first names
- Spelling of Russian alphabet
- time and date
- 7 phonetically rich sentences
- short stories

Each utterance is recorded into a separate audio file and supported with an annotation text file.

There are 100 speakers in the speech part of the corpus.

The annotation file contains the orthographic transcription of the actual utterance and its canonical form, with marks for the non-speech events.

The size of the speech part is 2 Gbyte.

Most of the speech part data is recorded simultaneously into two channels:

- narrow band telephone channel (4kHz, Moscow fixed and cell telephone networks)
- broad band microphone channel (11kHz Shure SM10 and computer Koss SB35 mikes).

On the whole, the speech corpus TeCoRus, including TeCoRus data itself (audio files and annotation text files) and a pronunciation dictionary, requires 2.6 Gbytes.