The Oxford Aesop Corpus 2010

Anastassia Loukina and Greg Kochanski

October 29, 2010

1 Overview

This corpus contains the data collected at the Oxford University Phonetics Laboratory for the project "Comparing dialects and languages using statistical measures of rhythm" funded by ESRC RES-062-23-1323. Further information can be found at `www.phon.ox.ac.uk/speech_rhythm`. To refer to the corpus please cite one of our papers available at this page.

The main corpus of data consists of short paragraphs and children poetry read by native speakers of Southern British English, Russian (Moscow and St.-Petersburg), Greek (Athens), Taiwanese Mandarin and French (Paris). The paragraphs were selected so that they did not contain any dialogue. Most poems contained 8 syllables per line.

Speakers were 20-28 years old, born to monolingual parents, and had grown up in their respective countries. At the time of the recording, all speakers were living in Oxford, UK. Non-English participants had lived outside their home country for less than 4 years. Recordings were made through a condenser microphone and a lapel microphone in a soundproof room in the Oxford University Phonetics Laboratory and saved direct to disc at a 16 kHz sampling rate. Texts were presented on a screen in standard orthography for each language.

All speakers of Greek, French and Russian and read the same 45 texts and retold Cinderella. Mandarin speakers read 73 shorter texts and also retold Cinderella. English speakers were divided into 2 groups (12 speakers each). The first group read the same 45 texts as the speakers of other languages and retold Cinderella. The second group did not read or retell Cinderella. Instead these speakers repeated 4 texts three times. Each repetition was recorded on a separate day.

In addition to short texts, all speakers also read up to 700 short sentences which were intended to use for training an automatic speech recognition system. These sentences are offered in a separate archive called 'ASR sentences'.

2 Description of the data

The experimental data itself consists of speech recordings, and they are stored in subdirectories. It also contains the orthographic texts, automatically generated transcriptions and metadata files with information about each file.

2.1 Metadata files

The corpus has one metadata file for each tar.gz file. The metadata files are in FIAT format. Further information about this format can be found at `http://www.phon.ox.ac.uk/files/pdfs/fiat.pdf`. Fiat files may be read by the fiatio python module in the gmisclib package available from `http://sourceforge.org/projects/speechresearch`, with documentation at `http://kochanski.org/gpk/code/speechresearch/gmisclib/gmisclib.fiatio-module.html`. They contain a line-by-line description of each utterance.

Each metadata files contains the following information:

- **d** - a directory containing all files for a particular utterance. For filenames within each directory see 2.2
- **language** - the language of the text (English, Russian, French, Greek or Chinese)
- **subjectID** - the unique ID of the speaker
- **gender** - the gender of the speaker ('m' of 'f')
- **group** - the group to which the speaker was assigned (1 or 2 for English speakers, 1 for all other speakers. See 1 for further information on groups.
- **type** - A type of text ('poetry', 'fable' for Aesop's fables, 'Cinderella' for reading of the Cinderella, 're-telling' for retelling of Cinderella, 'paragraph' for other texts
- **text** - A unique numerical code for each text read by the speaker.
- **tempo** - in some cases speakers were instructed to read 'faster than usual' or 'slower than usual'. This is reflected by 'fast', 'slow' or 'normal' values in this column. Note that these values only reflect the instructions given to the speakers and may not reflect the actual speech rate.

- **repetition** - for texts that were read three times, the number of the repetition (1, 2 or 3)
- **RecordStartTime** - the timestamp for the beginning of the recording
- **RecordEndTime** - the timestamp for the end of the recording

In addition to the metadata files, each archive contains this Readme.pdf file and a textfile with all stimuli and their codes.

2.2 Data Files

There are several files inside each directory:

- **raw.wav**

  The original recording, in Microsoft WAV format. It is a two-channel file. The first (lower numbered) channel (0 or 1) contains the recording done using a (5 mm diameter) lapel microphone; the second (higher numbered) channel contains the same recording done using a (15 mm diameter) condenser microphone. The lapel microphones sometimes malfunctioned, therefore we recommend using the recordings obtained from the condenser microphone. The malfunctions should be detectable, as they either led to high amplitude noise or near silence; about 20% of the lapel recordings have malfunctions.

- **text.txt**

  The orthographic text which was presented to the speaker. FThe files use Unicode encoding (UTF-8).

- **text.phones**

  This is an ASCII file which contains automatically generated transcription in X-Sampa format. The transcriptions do not always reflect false starts, hesitations or repetitions.