ATCOSIM Air Traffic Control Simulation Speech Corpus

Konrad Hofbauer, Stefan Petrik and Horst Hering

README
(Version 1.0)

Abstract

This DVD (or ISO image) contains the complete ATCOSIM corpus. The ATCOSIM Air Traffic Control Simulation Speech corpus is a speech database of air traffic control (ATC) operator speech. It consists of ten hours of speech data, which were recorded during ATC real-time simulations using a close-talk headset microphone. The utterances are in English language and pronounced by ten non-native speakers. The database includes orthographic transcriptions and additional information on speakers and recording sessions.

1 Speaker Profiles

Spk. ID	Nationality	Native Tongue	Gender	ATC Sector	Utterances	Int. ID
sm1	German	German	Male	Söllingen	1167	a
sm2	German	German	Male	Söllingen	1848	b
sm3	German	German	Male	Söllingen	808	c
sm4	German	German	Male	Söllingen	1162	d
gf1	Swiss	Swiss French	Female	Geneva	238	e
gm1	Swiss	Swiss French	Male	Geneva	384	f
gm2	Swiss	Swiss French	Male	Geneva	378	g
zf1	Swiss	Swiss German	Female	Zürich	1716	i
zf2	Swiss	Swiss German	Female	Zürich	1739	j
zf3	Swiss	Swiss German	Female	Zürich	638	k

2 Content

Four directories are located in the top-level directory of the corpus.

DOC (Documentation Directory)

readme.pdf This introductory document.

filelist.txt Full listing of all files and directories in the ATCOSIM distribution.

license.txt License terms and disclaimer.

atcosim_report.pdf Complete documentation of the corpus.

eec_simulation Directory containing documents about the simulation during which ATCOSIM was recorded.

validation Directory containing the validation report.

papers Directory containing publications about the ATCOSIM corpus.

artwork Directory containing a DVD jewel case inlet.

WAVdata (Sound Files Directory)

The WAVdata directory contains the recorded speech signal data. Each file corresponds to one controller utterance. The file format is single-channel Microsoft WAVE with a sample rate of 32kHz and a resolution of 16bits per sample. The 10,078 files are located in a sub-directory structure with a separate directory for each of the ten speakers and sub-directories thereof for each session of the speaker. The speaker directories are named according to the speaker ID, where the first letter stands for the controller's control centre (geneva, söllingen or zürich), the second letter for the gender of the controller (female or male), and the digit on the third position being a consecutive numbering for controllers with identical gender and control centre.

The session directories are sub-directories of the speaker directories and are named by the speaker ID, followed by underscore, followed by a consecutive two-digit number that identifies the session within that speaker.

The utterance files within the session directories are named by the speaker ID, followed by an underscore, followed by the two-digit session number, followed by a three-digit utterance number within this session. We refer to this sequence as the 'full utterance ID'. The full file name is therefore the full utterance ID, followed by the file extension '.wav'. For example, the file 'zf2_04_010.wav' is the tenth utterance in the fourth session of the second female Zürich speaker.

TXTdata (Transcription and Meta Data Directory)

All files described herein are text files in plain-text 7-bit ASCII encoding. They are thus also compliant to e.g. ISO/IEC 8859-1 (ISO Latin-1) and Unicode (UTF-8) encoding, as no special characters outside the 7-bit ASCII range are used.

***.txt files** The TXTdata directory contains the same directory structure as the WAVdata directory. It contains a plain-text file for each utterance which consists of the orthographic transcription of the utterance. The file name is the full utterance ID as described above, followed by the file extension '.txt'.

fulldata.csv file In the root of the TXTdata directory, the file fulldata.csv contains the complete annotation and meta data for all utterances and should be the primary data source when using the corpus. The file is a comma-separated value (CSV) file according to RFC 4180, and should therefore be simple to import in a large number of database programs, spreadsheet programs and programming languages. The CSV file represents the annotation data in a table-like manner. Each line in the file corresponds to one utterance. Lines are terminated by a Unix-style LF (Line feed, 0x0A) newline character. Each line consists of several data fields, which are separated by commas. The data fields itself do not contain any commas, double quotes or newline characters. The first line is a header line which, also separated by commas, briefly describes the meaning of each field (column). A full description of each field is given in the corpus documentation.

wordlist.txt The wordlist.txt file contains an alphabetically sorted list of all occurring words, including location names, airline radio call-signs, truncated words and special mark-up characters, codes and symbols.

HTMLdata (Browsable Transcriptions Directory)

The files in HTMLdata directory are HTML files which present the data in a table form so that they can be displayed in a standard HTML web browser. These files are provided purely for convenience and should not be used for further processing, as certain special characters are escaped in the HTML code and also conversion errors might have occurred. The fulldata.csv file should be the primary source of information.

The functionality provided by these files may vary depending on the operating system and web browser used, and also depending on the configuration thereof. The files were tested using Mozilla Firefox 2.0 (with installed Quicktime Plug-in) on Microsoft Windows XP (SP2) and Apple Mac OS X 10.4.10.

As the tables are comparably large, they might take a long time to load. The column headers of the tables use abbreviated titles, with the full titles being shown as tool-tips. A click on the 'Play' field next to each utterance may start a JavaScript which replays the audio of the corresponding utterance in a separate browser window or tab. In those tables that are dynamic, a single-click on one of the column headers sorts the entire table according to this column. On a state-of-the-art-in-2007 desktop computer the JavaScript-based sorting of the dynamic tables takes approximately one minute.

fulldata_static.htm The file contains the same information as included in the fulldata.csv file, but presented as a static HTML table.

fulldata_dynamic.htm The file contains the same information as included in the fulldata.csv file, but presented as a dynamically sortable HTML table.

overview_sortedby_*.htm The files show only the most relevant data fields. This provides a better overview and more space to display the actual transcription. The data is presented in a static HTML table, which is pre-sorted according to the criterion indicated by the filename.

overview_dynamic.htm The file shows only the most relevant data fields, presented as a dynamically sortable HTML table.

wordlist.htm The file provides a list of all occurring words including location names, airline radio call-signs, truncated words and special mark-up characters, codes and symbols. It also includes the number how often each word occurred in the corpus and provides an exemplary list of utterances that contains the corresponding word. The same functionality as above to replay the utterance is included.

3 Distribution

The corpus is publicly available and provided free of charge, except for potential shipping and handling costs. The entire corpus including the recordings and all meta data has a size of approximately 2.5 gigabyte and is available in digital form on a single DVD-R, or as an electronic ISO disk image at `http://www.spsc.tugraz.at/ATCOSIM`.

To obtain a DVD copy of the corpus contact:

EUROCONTROL Experimental Centre (Horst Hering)
Centre du Bois des Bordes, B.P. 15
F-91222 Brétigny-sur-Orge CEDEX
France
Email: horst.hering@eurocontrol.int