ATCOSIM - Air Traffic Control Simulation Speech Corpus Validation Report

Stefan Petrik

October 31, 2007

Abstract

The ATCOSIM speech corpus provided by Eurocontrol Experimental Centre has been validated against a list of specified checks. The validation covered completeness, formal checks and manual checks of randomly selected samples. The overall quality of the corpus is good and there should be no problem in using the corpus for speech applications. Minor improvements to the documentation have been proposed and may be incorporated without much effort. The ATCOSIM corpus is in a useable state.

1 Introduction

This document summarises the validation procedure of the ATCOSIM corpus, performed externally at Graz University of Technology. The ATCOSIM corpus is a speech corpus of simulated air traffic control (ATC) operator speech provided by Eurocontrol Experimental Centre, Brétigny. It was originally recorded in terms of ATC simulation sessions in 1997, and orthographically transcribed and annotated in 2007. The corpus consists of ten hours of speech data, recorded with a close-talk headset microphone. The utterances are in English language (following common ATC phraseology), pronounced by ten non-native speakers. The corpus provides orthographic transcriptions and meta data about speakers and the transcription process. The aim of this corpus is to provide a language resource for studying real-world air traffic controller speech and a basis for development of speech applications, without being tailored to a specific application.

2 Validation results

The validation was done on corpus version 0.8 and carried out following the guidelines of the Bavarian Archive for Speech Signals (BAS)[SD02]. The basic methodology employed was to inspect the corpus from a user's point-of-view. Due to the fact that the ATC simulations were meant for a different purpose, no prior specification of corpus and recordings is available. Therefore, the data was examined as it had been provided, and then tested against the documentation enclosed with the distribution.
The following section contains all validation steps as specified in the validation contract, together with the detailed methodology and the results. For each subsection, a summary of the tests performed is given, together with additional comments and an indicator whether it had been passed. The indicator can be either of:

- √ PASSED
- ⊙ ACCEPTABLE, modifications are suggested
- × FAILED

2.1 Completeness, terminology, readability and parsability of data

- Signal files
  The corpus includes 10,078 .wav-files of speech signals, stored in the directory `WAVdata`. It is divided according to speakers and recording sessions. Table 1 shows a summary of the signal

files found in the corpus. The numbers exactly match the figures given in the documentation files `readme.pdf` and `atcosim_report.pdf` (c.f. sec. 2.2).

speaker	sessions	signal files	annotation files
gf1	1	238	238
gm1	2	384	384
gm2	2	378	384
sm1	7	1167	1167
sm2	9	1848	1848
sm3	5	808	808
sm4	6	1162	1162
zf1	8	1716	1716
zf2	7	1739	1739
zf3	3	638	638
total	50	10078	10078

Table 1: Signal and annotation files found in the distributed corpus.

Files are named according to the following scheme, with regard to speaker and recording session:

$$
\begin{aligned}
\textit{gender} &:= \texttt{m (male)} \mid \texttt{f (female)} \\
\textit{digit} &:= [0..9] \\
\textit{sector} &:= \texttt{g (Geneva)} \mid \texttt{s (Soellingen)} \mid \texttt{z (Zuerich)} \\
\textit{sessionID} &:= \textit{digit digit} \\
\textit{utteranceID} &:= \textit{digit digit digit} \\
\textit{speakerID} &:= \textit{sector gender digit} \\
\textit{fileID} &:= \textit{speakerID} \ \texttt{\_}\ \textit{sessionID} \ \texttt{\_}\ \textit{utteranceID} \ \texttt{.wav}
\end{aligned}
$$

All signal file names meet this convention which is also documented in the documentation files `readme.pdf` and `atcosim_report.pdf` (c.f. sec. 5.3.1). No white spaces are used in file and directory names. Furthermore, all signal files were readable and of size of more than 0 Bytes.

Test summary:
File counts: $\surd$
Naming conventions: $\surd$
Readability: $\surd$
Filesize > 0 Bytes: $\surd$
Status: $\surd$.

- Meta data files

  Meta data is stored in the directories `HTMLdata` and `TXTdata`. It consists of speaker and utterance information, an indicator for file corruptness, comments by the transcriptionist, the length of the recorded utterance, and the position of the utterance on the initial data collection tapes. The meta data is presented in HTML and CSV format.

  The HTML files are encoded in ISO-8859-1 (Western) character encoding. In addition to the HTML files, Javascripts are enclosed which enable playback of wave files and sorting of table columns within .htm files. The files entitled "dynamic" are very slow for loading and sorting of table columns, resulting even in a warning message in the Firefox web browser. This problem, however, is mentioned in the documentation. The sorting feature is interesting, but the "static" file versions are better suited for browsing.

  Furthermore, an additional CSV file (`fulldata.csv`) comprising all available meta-data and transcriptions is available in the directory `TXTdata`. For easy database import, this file follows the comma separated values (CSV) format. Lines are terminated with UNIX style line endings (LF only). File format and line terminators were successfully verified.

Test summary:

File existence:	√	
Readability:	√	(slow for "dynamic" HTML files)
Line terminators	√	(single LF for `fulldata.csv`)
Filesize > 0 Bytes:	√	

Status: √.

- Annotation files (transcriptions)
  For each signal file, an annotation file is included in the directory `TXTdata`. The directory structure is the same as of the signal files, with annotation files following the same naming conventions, apart from the different file extension (`.txt` instead of `.wav`). As listed in table 1, the number of signal and annotation files is identical and the one-to-one correspondence was verified. No white spaces are used in file and directory names.

  The annotation files themselves are in plain text format. The mime-type and encoding are `text/plain; charset=us-ascii`. There are no line terminators used, as there is only one line per file. In one case (`zf1/zf1_02/zf1_02_197.txt`), this results in a very long line which may turn out to be problematic for some text editors.

  Test summary:

File counts:	√	
Naming conventions:	√	
Readability:	√	
Filesize > 0 Bytes:	√	
Character encoding	√	(us-ascii)
Line terminators	√	(no line terminators)
Leading/trailing spaces	√	(one trailing space, consistently)

  Status: √.

2.2 Superfluous files

No superfluous files were found in the corpus.

Test summary:

File list inspection	√

Status: √.

2.3 Technical specifications of signal files

The technical specification of the signal files was tested with the UNIX program sox[1]. The observed formats are in accordance with the specification in the documentation.
In addition, a number of tests was performed to estimate the quality of the recordings. The *file length* distribution was calculated to trace spurious recordings of extraordinary length. The *clipping ratio*, defined as the proportion of samples in a file that is equal to the maximum/minumum value divide by all samples in the file helps detecting distorted recordings. Finally, the *mean sample value* can be used to trace files with large DC-offsets. The resulting statistics for these tests are shown in table 2. No files were found which showed significant deviations.

[1] `http://sox.sourceforge.net`

LEN [s]	count
0-1	541
1-2	416
2-3	1593
3-4	3414
4-5	2347
5-6	979
6-7	456
7-8	193
8-9	77
9-10	34
10-11	13
11-12	5
12-13	4
13-14	0
14-15	2
15-16	0
16-17	0
17-18	0
18-19	0
19-20	4

CR [%]	count
.00-.01	9583
.01-.02	251
.02-.03	112
.03-.04	56
.04-.05	30
.05-.06	13
.06-.07	11
.07-.08	5
.08-.09	3
.09-.10	3
.10-.11	2
.11-.12	1
.12-.13	4
.13-.14	0
.14-.15	1
.15-.16	0
.16-.17	1
.17-.18	0
.18-.19	0
.19-.20	2

MSV	count
-.005-.004	0
-.004-.003	0
-.003-.002	0
-.002-.001	151
-.001-.000	9555
.000-.001	37
.001-.002	41
.002-.003	71
.003-.004	89
.004-.005	44
.005-.006	24
.006-.007	19
.007-.008	11
.009-.010	10
.010-.011	6
.011-.012	4
.012-.013	6
.013-.014	5
.014-.015	1
.015-.016	4

Table 2: Statistics for signal length (LEN), clipping ratio (CR), and mean sample value (MSV).

Test summary:

File format	$\surd$	(Microsoft Wave file)
Encoding	$\surd$	(PCM linear)
Sampling rate	$\surd$	(32000 samples/s)
Resolution	$\surd$	(16 bits/sample)
Channels	$\surd$	(1)
Signal length	$\surd$	
Clipping ratio	$\surd$	
Mean sample value	$\surd$	

Status: $\surd$.

2.4 Documentation

The documentation for the corpus is stored in the `DOC` folder. Apart from the text file `license.txt` which contains the license description, the rest of the documentation is in PDF format. There is no software included to open these files (Acrobat Reader, or similar).

2.4.1 Completeness of documentation and consistency with speech corpus

Completeness of documentation was determined with regard to *administrative* information, *content* information, *speaker* information, *recording* information, and *annotation* information.
In general, the documentation appears very complete, as there is a lot of background information included concerning the recording setup (c.f. folder `eec_simulation`). However, some important details are missing. There is no full listing of all the files in the distribution. The document `readme.pdf` only gives a rough idea what can be found where. Furthermore, information about speaker profiles (particularly speaker age, experience) and a description of the characteristic ATC phraseology are either spread over multiple files or not explicitly referenced in the main documentation files `readme.pdf` and `atcosim_report.pdf`. Finally, information about the validation process (validation report, institution) is missing. Minor modifications would improve the informativeness significantly.
Since the corpus was created without prior specification of the speech tasks, the data cannot be tested against such a specification. For this reason, only the compliance of the found data with the description in the documentation is validated.

Test summary:

Contact for requests regarding the corpus	√	(c.f. `readme.pdf`)
Number and type of media	√	(1 DVD, c.f. `readme.pdf`)
Layout of media	√	(c.f. `readme.pdf`)
Content of each medium	×	(missing file listing, c.f. `readme.pdf`)
Copyright statement & intellectual property rights	√	(c.f. `license.txt`)
Validation date(s)	×	
Validation person(s)/institution(s)	×	
Clearly stated purpose of the recordings	√	(c.f. `atcosim_report.pdf, sec.1.2`)
Speech type(s)	√	(c.f. `eec_note_2001_01.pdf`)
Instruction to speakers (full copy)	√	(implicitly given by ATC phraseology)
Number of speakers	√	(10)
Distribution of speakers over sex, age, native language	⊙	(c.f. `atcosim_report.pdf, sec.2.2`)
Description/definition of native languages	√	(German, Swiss German/French)
Recording platform	√	(Sony DTC60ES)
Position and type of microphone(s)	√	(Sennheiser HME 45-KA)
Acoustical environment	√	(c.f. `atcosim_report.pdf, sec.2.1`)
Annotation manual, guidelines, instructions	√	(c.f. `atcosim_report.pdf, sec.4`)
Description of quality assurance procedures	√	(c.f. `atcosim_report.pdf, sec.4.5`)
Background of annotators	√	(c.f. `atcosim_report.pdf, sec.4.4`)
Training of annotators	√	(c.f. `atcosim_report.pdf, sec.4.5`)
Annotation tools used	√	(c.f. `atcosim_report.pdf, sec.4.1`)

Status: ⊙.

2.4.2 Readability of documentation files

The documentation files were tested on Windows and Linux platforms and were found to be readable with Acrobat Reader software. However, there is no software included for reading these files, nor any description of how to obtain it.

Test summary:

Readability on Windows, Linux, Macintosh	√	
Software reader	⊙	(not included, provide at least a link)

Status: ⊙.

2.5 Transcriptions

The transcriptions were tested for completeness, accordance with the guidelines, and accuracy by manual validation. The transcription guidelines are listed in `atcosim_report.pdf`. In sec. 2.4.1, the general information given about the transcription process was already found to be complete.

2.5.1 Completeness and accuracy

Spelling and label symbols were found to be consistent with the specification. Only in two aspects, refinement is suggested. First, a list of all used airline radio call signs and location names would aid in understanding the sometimes unusual spellings. And second, the list of foreign words should be translated to English.

Test summary:

Unambiguous spelling standard used	√	(c.f. `atcosim_report.pdf, sec.4.2`)
Labeling symbols	√	(c.f. `atcosim_report.pdf, sec.4.2`)
List of non-standard spellings	×	(airline radio call sign and location list missing)
Character set used in annotations	√	(us-ascii)
Language dependent information	×	(translation of foreign words missing)

Status: ⊙.

2.5.2 Manual validation of transcriptions

To estimate the quality of the transcriptions, a small number of randomly selected utterances were transcribed anew and compared to the provided transcriptions. The re-transcription was conducted as follows:

1. A set of 130 utterances was randomly selected from the corpus by an automatic script.
2. The first 30 utterances were used to train the re-transcriptionist to the domain of ATC controller speech and to make him familiar with the signal quality and the transcription guidelines (c.f. `atcosim_report.pdf, sec.4.2`). These re-transcriptions were not counted, since they were produced by consulting the original transcriptions.
3. Then, the remaining 100 utterances ($\sim$ 1% of the whole corpus) were re-transcribed without checking the provided transcriptions. The only help provided to the re-transcriptionist was the lexicon as a reference for the (non-trivial) callsigns and place names used in the utterances.
4. Afterwards, the re-transcriptions were compared to the original transcriptions to analyse the differences. Table 3 lists the statistics and table 4 lists the mismatching utterances.
5. The actual accuracy of the original transcripts was determined by comparing original transcription and re-transcription to the audio files. This way, the errors causing the mismatches could be assigned to either the original or the re-transcription.

Despite the initial training period, many parts remained unintelligible to the re-transcriptionist resulting in a lot of `[UNKNOWN]` tags in the re-transcriptions. To still provide a meaningful analysis, these tags were treated separately in the statistics, shown in table 3.

	identical		`[UNKNOWN]`		different		total
	[1]	%	[1]	%	[1]	%	[1]
utterances	75	75	9	9	16	16	100
single tokens	1008	96.1	18	1.7	41	3.1	1049
→ substitutions			18	1.7	15	1.4	33
→ insertions			0	0	5	0.5	5
→ deletions			0	0	3	0.3	3
actual accuracy	1044	99.4			6	0.6	1049

Table 3: Results of the manual validation of transcriptions in absolute counts [1] and [%].

On utterance level, 75% of the transcriptions were identical, and another 9% only differed in terms of an `[UNKNOWN]` tag in the re-transcription. The remaining 16 utterances were analysed in more detail and the mismatches listed in table 4. Many substitution errors can be attributed to homophony or high phonetic similarity. On token level, 96.1% of tokens were identically transcribed, and another 1.7% of tokens were just tagged as unintelligible ([UNKNOWN]) in the re-transcription. It should be noted that the number of insertions (tokens found in the original transcription, but not in the re-transcription) and deletions (tokens found in the re-transcription, but not in the original transcription) is very low.
After a re-consultation of the audio files, 6 definite errors could be determined in the original transcriptions: 3 insertions, 2 substitutions, and 1 omission of a transcription marker). Therefore, the actual accuracy of the transcriptions is 99.4% on word level.

Test summary:

Accuracy	$\surd$	(99.4% actual word accuracy of the original transcriptions)
Spelling	$\surd$	(no spelling mistakes)
Digits/numerals	$\surd$	(all digits/numerals written out)
Case errors	$\surd$	(all items lower-case, tags all upper-case)
Punctuation	$\surd$	(no punctuation used)

Status: $\surd$.

	utterance	original transcription	re-transcription
	zf2_01_122	... two thousand *or more*	... two thousand *on one*
	sm3_03_153	alitalia one four nine zurich ...	alitalia one four nine *at* zurich ...
	zf2_03_228	turkish *ah* nine two five ...	turkish *air* nine two five ...
	zf3_02_063	... further *climb zurich sector* one three ...	... further *climbs are exact* one three ...
×	gm1_01_070	speedbird five six nine *you're* identified ...	speedbird five six nine *you* identified ...
	gf1_01_217	... st prex passe= *ah* passeiry ...	... st prex passe= passeiry ...
	sm3_05_093	stand by *@aerovic* one zero six one ...	stand by *air wag* one zero six one ...
	zf2_01_133	... continue present heading ...	... continue *your* present heading ...
×	zf1_07_196	... one nine zero descend *to* flight level ...	... one nine zero descend flight level ...
	zf2_04_091	... contact *proceed via* trasadingen zurich ...	... contact *received by* trasadingen zurich ...
×	zf2_02_101	... resume *on* navigation inbound fribourg	... resume *own* navigation [UNKNOWN] fribourg
×	zf3_02_051	∼l ∼t ∼*u* one six zero two zurich	∼l ∼t one six zero two zurich
×	zf2_05_139	alright cross air five one eight ...	<*OT*> alright </*OT*> cross air five one eight ...
	gf1_01_071	aero lloyd *ah* five one seven	aero lloyd five one seven ...
	zf2_07_142	... direct fusse *rate* of descent ...	... direct fusse *wait* of descent ...
×	zf2_07_039	... radar contact i'll call you *back with* higher	... radar contact i'll call you *about the* higher

Table 4: Mismatching utterances between original and re-transcription. Mismatches are highlighted in *italics* and definite errors in the original transcriptions are marked with (×).

2.6 Lexicon

The corpus provides a file, containing a list of all words used in the transcriptions (`wordlist.txt`). It is a plain text file in us-ascii character encoding with UNIX style line endings (LF only). It contains 858 tokens, sorted alphabetically in ascending order. A phonetic mapping is not included.

2.6.1 Completeness and accuracy

An automatic test of the coverage and sorting order proved the lexicon to be error-free. All tokens not being part of the English language (i.e. tags, foreign words, etc.) were explained in the documentation.

Test summary:

Coverage	√
Sorting order	√
Encoding (language, alphabet)	√

Status: √.

2.6.2 Manual validation of lexical entries

The lexicon was manually inspected with regard to orthography and spelling conventions. The spelling conventions are specified in detail in the documentation (c.f. `atcosim_report.pdf`, sec.4.2, 4.3). With common English words, no spelling mistakes were found.

Test summary:

Spelling	√	(no mistakes for common English words)
Foreign words translated	×	(missing)
Digits/numerals	√	(all digits/numerals written out)
Case errors	√	(all items lower-case, tags all upper-case)
Punctuation	√	(no punctuation used)

Status: ⊙.

2.7 Readability of distribution media

The corpus is provided as single hardcopy DVD or ISO9660 image file for download. Both, the image file and the DVD containing corpus data and documentation were successfully mounted on Windows, Macintosh and Linux platforms.

Test summary:
Readability on Windows, Linux, Macintosh $\surd$
Status: $\surd$.

2.8 Tools and additional software

The corpus distribution does not include any tools or additional software. The software that was used for the production of the corpus, however, is referenced in the documentation. Since no specific tools are necessary to access the database contents, nothing needs to be included.

Test summary:
References to software tools in documentation $\surd$
Status: $\surd$.

3 Conclusion

The ATCOSIM corpus of air traffic control simulation speech was validated against its documentation. A number of automatic tests including completeness, readability, and parsability were successfully performed without revealing errors. Furthermore, manual inspections of documentation, meta-data, transcriptions, and the lexicon were done, which showed minor shortcomings that can be improved without much effort. Finally, the re-transcription of 1% of the corpus data showed transcriber agreement of 96.1% on word level and an actual word accuracy of 99.4% of the original transcriptions, proving them to be accurate.

The ATCOSIM corpus is in a usable state.

References

[SD02] Florian Schiel and Christoph Draxler. *Production and Validation of Speech Corpora*. Bastard, Munich, Germany, 2002.