Abstract

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, automatically segmented, and orthographically transcribed. The utterances are in English language and pronounced by ten non-native operational controllers.

1. Speech Corpus for ATC

- Application Domain
  - Civil air traffic control (ATC)
  - Radar screen and party-line analogue voice radio
  - One controller responsible for 10-20 aircraft

- Speech Characteristics
  - Specific language, phraseology and vocabulary
  - Specific environment and speaking style
  - Frequently non-native speakers

- Peculiarities and Difficulties
  - Not free, read or command&control speech
  - Spoken Language Techn. virtually non-existent
  - High reliability requirements
  - Speech recordings difficult to obtain or produce
  - Very few corpora available

2. Existing ATC Related Corpora

3. ATCOSIM Recording Context

- Operational recordings legally problematic
- EUROCONTROL Experimental Centre, France
- Large-scale ATC real-time simulation for operational concept evaluations with professional controllers

4. The ATCOSIM Corpus

- Speech Recordings
  - Close-talk microphone
  - 50 sessions,
  - 52 h of recordings,
  - 10 h of speech
  - 10 non-native speakers
    (4 female, 6 male)
  - 10078 utterances
  - 108883 words

- Transcriptions (Examples)
  - zero lloyd five one seven proceed direct to frankfurt
  - speedway three five four ten contact milan one three four five two
  - speedway three five four turn right ten degrees report new heading
  - transaero six one seven proceed to framburg for roman indi
  - good afternoon belgian airforce forty four non r "r" x "m identified
  - [INFORMATION] aeronautica six four seven four report your heading
  - [FRAGMENT] lufthansa six four seven identified
  - [UNKNOWN] alitalia six four seven four report your heading
  - [UNKNOWN] speedway three five four ten identified
  - [UNKNOWN] lufthansa six one seven Identified OFF \r it's over now \r OFF
  - [UNKNOWN] zero lloyd [INFORMATION] charlie papa alfa guten tag radar contact

- Meta Data
  - Speaker IDs, comments, duration, ...
  - Supplementary HTML tables

- Documentation
  - Corpus manual, transcription guidelines, ...
  - Validation report (following BAS guidelines)

5. Availability

The corpus is available as electronic download from http://www.SPSC.TU Graz.at/ATCOSIM.