Introduction

The working group discussed voice formats of SELF-MADE-ATC. The working group consists of 10 people. Eight of them were on-site present in Braunschweig.

Motivation

- **Foundation of ASRU in ATC:** Voice recordings are the essential foundation for Automatic Speech Recognition and Understanding (ASRU) in Air Traffic Management.
- **Recorded Data vs. Machine Learning Pipelines:** Operational ATC audio is typically recorded and archived for legal compliance and incident investigation and not pre-engineered for machine learning pipelines.

Comparison between the ideal ASRU standpoint and the current operational reality experienced with ANSP recordings:

Dimension	The Ideal Case (Pure ASRU Standpoint)	The Operational Reality
Audio Format & Sampling Rate	16 kHz linear PCM mono recordings, which serve as the standard baseline for modern Automatic Speech Recognition models.	Usually 8 kHz encoded as A-Law or U-Law. While ASR models can handle this after transforming it into 16 kHz linear PCM, having a higher-quality native signal available would simplify processing.
File Length & Segmentation	One individual file per transmission avoiding multi-minute or multi-hour recordings that contain multiple transmissions separated by silence.	Often delivered in large chunks of audio containing multiple transmissions. Requires automated splitting tools afterwards, which do not work perfectly and occasionally split files at the wrong places.
Speaker Role Identification	Known speaker role per transmission (identifying whether the speaker is acting as a Pilot or an Air Traffic Controller)	Often not available in recorded data archives. Requires automated speaker role detection, which leaves room for errors.

Questions given as guidance to the participants

- How are voice communications currently captured, stored, and retrieved across different organizations (e.g., for safety oversight and incident analysis)?
- Does current recording infrastructure preserve transmission markers (such as PTT or explicit start/end timestamps) or do archives primarily consist of continuous multi-hour audio streams with long periods of silence?
- Is the speaker role readily identifiable (e.g., for a recorded transmission, is it known whether an ATCO or pilot is speaking)?

- Even if data is not stored in a Machine Learning-ready format today, what existing recording system interfaces or streams could provide the necessary segmentation and metadata?

Results

Agreements

- **Live ASRU Systems Covered by Existing Standards (ED-137):**
  - For live ASRU deployments, existing standards already contain all required information. Specifically, under ED-137, an ASRU system can establish a session with the Voice Communication System (VCS) and specify the use of the uncompressed G.711 codec.
  - In live sessions, speaker role identification is available via Push-to-Talk (PTT) and Squelch (SQU) signalling, although Cross-Talk (simultaneous/overlapping transmissions) can still pose a problem.
- **Offline Recorded Data Configuration:**
  - For offline recorded data, it is theoretically already possible to record all relevant metadata such as Squelch and PTT, but currently it is up to the individual customer (ANSP) to configure their recording system in that manner.

Open Challenges

- **Airspace, Sector & Airport Identification (WG-126):** It is important to know which specific sector or airport the audio data originates from so that that it can be stored as metadata as well and Automatic Speech Recognition (ASR) models can be tailored to the respective airspace. The standard currently being developed by EUROCAE WG-126 could provide a standardized solution for this.
- **Utilizing Existing Unannotated Legacy Audio Archives:** A key open challenge remains how to make better use of the large amounts of offline audio data that are already available today, but were recorded without the necessary meta-information (such as PTT, squelch). The question is especially important for training the understanding part. Even training data for the speech-to-text models without ATCO-pilot classification can be used, because in most cases a common ATCO-pilot model is trained. If some audio files contain more than one utterance, this also usable for training speech-to-text models.