Authors: Aleš Pražák, Luboš Šmídl

Title: Czech Parliament Meetings

Identifiers: ZCU_CZ_Parliament

Subject keywords: speech corpus; acoustic model; speaker identification and verification

Sponsors:

Description:

The corpus consists of recordings from the Chamber of Deputies of the Parliament of the Czech Republic. It currently consists of 88 hours of speech data, which corresponds roughly to 0.5 million tokens. The annotation process is semi-automatic, as we are able to perform the speech recognition on the data with high accuracy (over 90%) and consequently align the resulting automatic transcripts with the speech. The annotator's task is then to check the transcripts, correct errors, add proper punctuation and label speech sections with information about the speaker. The resulting corpus is therefore suitable for both acoustic model training for ASR purposes and training of speaker identification and/or verification systems.

File description:

Speech data and corresponding annotations

The archive contains 18 sound files (WAV PCM, 16-bit, 44.1 kHz, mono) and corresponding transcriptions in XML-based standard Transcriber format (http://trans.sourceforge.net)

Annotation conventions description:

Besides word-by-word transcription of the speech, the following non-speech events are also captured in the annotation:

- `[speaker]` – noises produced by a speaker (loud breath, coughing, lip smacking, sneezing, etc.)
- `[hesitation]` – fillers stemming from speaker's hesitation ("Er" and "Uhm", etc.)
- `[noise]` – other environmental noises that significantly stand out from the background noise level (knocking, rustling, car horn sounds, etc.)
- `[non-speech]` – whole segment containing only silence, noise, music and/or other non-speech events
- `[background-speech]` – overlapping speech from additional speaker(s)
- `[unintelligible]` – unintelligible segment of speech
- `[pron=pronunciation]`word`[/pron=pronunciation]` – for recording non-standard (unexpected) pronunciations of words enclosed within the tag. Also used for writing down the full word forms of abbreviations, acronyms and various special symbols (like section sign §)

Example of the annotation in the Transcriber tool