Etape evaluation - metrics and tools

Olivier Galibert

Laboratoire national de métrologie et d’essais

Le progrès, une passion à partager
Evaluation themes

4 tasks

- Overlapping voices detection.
- Speaker diarization.
- Speech transcription.
- Named entities.
Overlapping voices detection

The task
- Detect the segments where voices overlap.

The metric
- Time precision and recall.

The format
- Standard ETF.
Speaker diarization

The task
- Detecting speech segments and assigning them to (unnamed) speakers.

The metric
- Standard diarization error rate.
  - Map hypothesis and reference speakers.
  - Divide the time in error by the reference time.
- Two evaluation setups:
  - One mapping per show.
  - One mapping for all the shows (cross-show diarization).

The format
- Standard MDTM.
Speech transcription

The task

- Transcribe all that is said, *including in overlapping parts*.
- Attribute every word to its speaker (as per the diarization).

The metrics

- Standard-ish word error rate, with words optimally distributed in multi-speaker zones.
- Speaker-attributed word error rate, where the mapping directs the word comparisons.

The format

- Based on CTM with a column added for the speaker label.
- Word times can overlap.
Named entities

The task

- Detect named entities per the Quaero guide.
- Done in manual transcription and automatic transcriptions.
- At a minimum, a rover of all submissions will be given.
- Real outputs can be used as long as the owners accept that they’re given out in the future evaluation package.

The metric

- Slot error rate, adapted and extended. See [IJCNLP 2011] for details.
Scoring tools

The LNE tools

- All GPLv3/LGPLv3.
- Not really ready yet, but very soon.

What currently exists

- Diarization with cross-show support (tested in Quaero and Repere).
- Speech recognition without overlapping (tested in Quaero and Repere).
- Named entities (tested in Quaero).