CS Flash Talk: Automatic Speaker Diarization

Author: Judy Fong

October 12, 2022

Outline

- What is speaker diarization?
- Example outputs
- Scoring quality of output
- Goal: Make the output more accurate

What is Automatic Speaker Diarization?

- Subfield within language technology
  - Relatively recent field within speech
- ~Segmenting in the subtitling field
- Who spoke when
- Infers from manually created datasets likely labels for conversations

Example Output – RTTM file

rttm

```
SPEAKER <recording-id> <channel> <start-time> <duration> <NA> <NA> <speaker-number> <NA> <NA>
```

SPEAKER 5004310T0 1 0.03 5.52 <NA> <NA> Roger <NA> <NA>
SPEAKER 5004310T0 1 6.60 8.46 <NA> <NA> Virgil <NA> <NA>
SPEAKER 5004310T0 1 15.49 3.75 <NA> <NA> Roger <NA> <NA>
SPEAKER 5004310T0 1 19.44 1.32 <NA> <NA> Virgil <NA> <NA>

Example Output – Subtitle File (.srt)

1
00:00:00,030 --> 00:00:05,550
-Roger:

2
00:00:06,600 --> 00:00:15,060
-Virgil:

3
00:00:15,490 --> 00:00:19,240
-Roger:

4
00:00:19,440 --> 00:00:20,760
-Virgil:

file: 5004310T0.mp3
Audio channel: 1

Scoring quality of output

- Diariation Error Rate (DER)
  - check how well a recipe/set of algorithms performs

- The formula:

$$DER = \frac{\text{False Alarm} + \text{Miss} + \text{Overlap} + \text{Confusion}}{\text{Reference Length}}$$

Goal: Make the output more accurate

- Currently domain dependent
  - Talk show
  - Movie
  - Meeting

- Current Icelandic DER: 26.27%
  - ¼ are wrong
  - No prior information
  - Icelandic National Broadcasting Service news and Kastljós (talk show)

Thank you! Takk fyrir!

More about diarization: https://wq2012.github.io/awesome-diarization/

Center for Analysis and Design of Intelligent Agents

Gervigreindarsetur Háskólans í Reykjavík