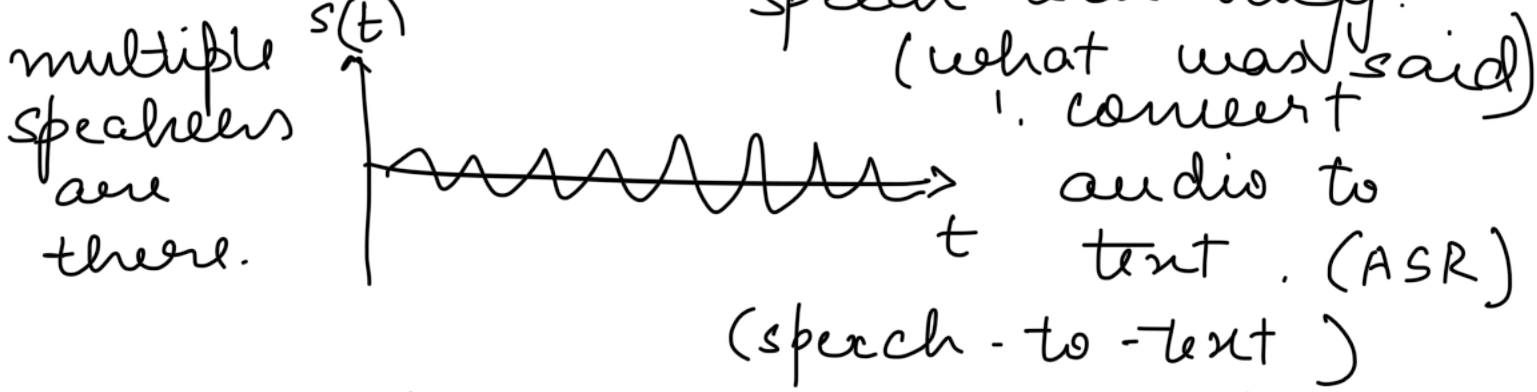
Week-6 - DLP

L1 Speaker diarizaⁿ - Intro

- Speaker diarisation → a applicaⁿ in speech technology.



also called as transcription.

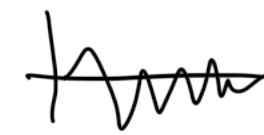
- instead of 1 big text sth.

S1. →
S2. →
S1. →
} like this would be better

↳ speaker diarisaⁿ (who spoke when)

2. ↳ we will use speaker identificaⁿ model to achieve this.

(Whisper ASR model) ← by Open AI
dominated by American English.

Eg.  → [ASR] 2.03 - 2.08 ___

2.09 - 2.17 ___

↓↓

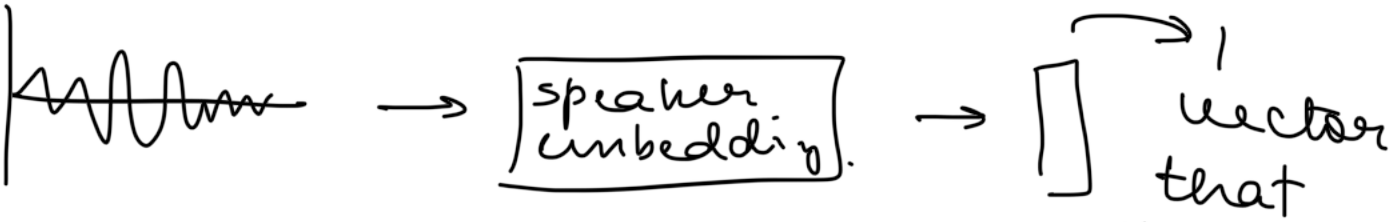
Speaker Identificaⁿ model

↳ Speaker embedding extraction

Speech Toolkit
Kaldi
Esp net
Speech Brain

ECAPA-TDNN Model will be used.

L2 Speaker Identification



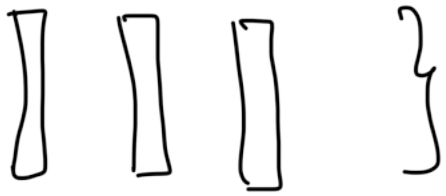
For any 2 vectors,

[ ] & [ ] we can calculate some kind of distance (euclidean dist.) to show their degree of closeness.

Or

→ we can use the angle btw. 2 vectors to explain closeness.

$$\cos\theta = \frac{x^T y}{\|x\| \cdot \|y\|}$$ → cosine distance



calculate the cosine dist. btw. these embeddings.

if, $\langle s_1, s_2 \rangle < \langle s_1, s_3 \rangle$

then, s_1 is closer to s_2 than s_3.

- for our case, we will assume that there are only n no. of speakers in my recording.

1.3 Clustering Techniques

- Agglomerative Clustering → keep merging clusters, until I hit n (desired no. of clusters).

Assumption → I know there 2 people only.

Initially, we have

7 segments ⟶ 7 embeddings $E_1 \dots E_7$

calculate,

$\langle E_1, E_2 \rangle$ — — — $\langle E_1, E_7 \rangle$

$\langle E_2, E_3 \rangle$ — — $\langle E_2, E_7 \rangle$

$\langle E_6, E_7 \rangle$

find the minimum cosine distance

Eg. $\langle E_4, E_7 \rangle$ was max in

$\hookrightarrow \langle E_4', E_4' \rangle$

and repeat this. → finally we reach a stage where there are only 2 embeddings. ∴, 2 speakers.

L4 demo

speechbrain == 0.5.16
faster-whisper
pyannote.audio
whisper

} use lib.

- Convert any format to .wav using ffmpeg.