Class 11: Audio Recognition and Alignment

Theme: Audio

Computational Analysis of Text, Audio, and Images, Fall 2023

Aarhus University

Mathias Rask (mathiasrask@ps.au.dk)

Aarhus University

Table of Contents

Automatic Speech Recognition (ASR)

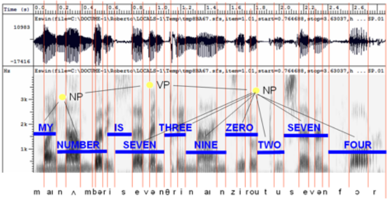
ASR is the process by which a machine can understand and transcribe spoken language into text.

Humans:

- Articulation produces sound waves
- Ear hears the waves
- Waves are sent to the brain for processing
- Largely speaker-*independent*

Computers:

- Digitization
- Acoustic analysis of the speech signal
- Linguistic interpretation



Complications:

- Linguistic components: accent, phonemes, and phonetics
- Signal components: Recording quality, background noise, and multi-speakers

ASR in Political Science

ASR is useful for political scientists for one reason in particular:
$\rightsquigarrow$ *automated* transcription of data sources

- ▷ Even better: political speech is often a grateful task for ASR systems

Data sources:

- TV/radio interviews
- Campaign debates
- Parliamentary debates

Facilitates:

- Direct text analysis (e.g. sentiment or topics)
- Indirect text analysis (e.g. hostile rhetoric removed in official transcripts)

Open-Source Tools

	Whisper	**Wav2Vec 2.0**
Learning	weak-supervision	self-supervision
Input	log-mel spectrograms	raw waveforms
Languages	> 100	$> 1,400$
Architecture	encoder/decoder	encoder
Output	processed	raw
Timestamps	segment-level	character-level
Error	$5 - 30\%$	$20 - 50\%$
Time (hours)	5.8	222.0

Other (paid) systems as well, but Whisper and Wav2Vec 2.0 are the best options:

- Google Cloud Platform
- Amazon
- Assembly AI

Word Error Rate (WER)

The most common way to assess ASR systems is the Word Error Rate (WER):

$$\text{WER} = \frac{S + D + I}{N}$$

where

- N is the total number of words in the target-text
- S is the number of substituted words (e.g. good is substituted with food)
- D is the number of deleted words (words that are missing in the ASR-text)
- I is the number of words in the ASR-text but not in the target-text

$\rightsquigarrow$ WER is a generalization of Levenshtein's distance to the word-level rather than the character-level

Exercise

Consider the following two sentences:

- Target: "The cat is sleeping on the mat"
- ASR: "The a cat is sweeping on mat"

1. Compute the WER
2. Discuss weaknesses of the WER metric and potential alternative metrics

$$\mathrm{WER} = \frac{S + D + I}{N}$$

- N is the total number of words in the target-text
- S is the number of substituted words
- D is the number of words missing in the ASR-text
- I is the number of words in the ASR-text but not in the target-text

Table of Contents

Speaker Diarization

Speaker diarization is the process of *segmenting* a speech signal $y(n)$ into K separate segments s_i for $i \in \{1, \ldots, K\}$

Each segment s_i is assigned to a separate speaker $j \in \{1, \ldots, J\}$, but the identity of each speaker is not known $\rightsquigarrow$ speaker labels are generic (e.g. 'A', 'B' or e.g. '1', '4')

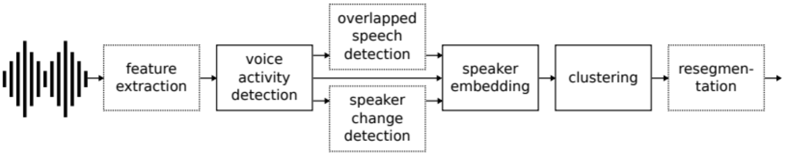
Applications:

- Audio transcription – a preprocessing step for ASR
- Speaker recognition – a preprocessing step for speaker recognition

$\rightsquigarrow$ decomposes audio to the speaker-level

Diarization Systems

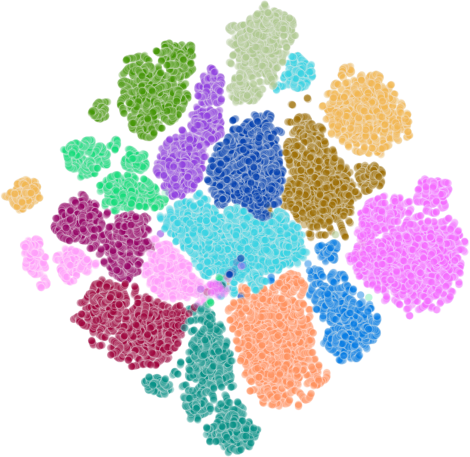
State-of-the-art diarization systems exploit neural networks and use end-to-end building blocks:



The output of the diarization pipeline hinges heavily upon the quality of the speaker embeddings

- Comparable to word embeddings: a fixed-length vector representation of a speaker's unique vocal traits, speaking style, and speech-related information
- Encoded with neural networks – pretrained models work surprisingly well

Two-Dimensional Visualization of Speaker Embeddings



⇝ Embeddings are fixed-length *x*-vectors (Snyder *et al.*, 2018) computed on diarized speech segments and then reduced using t-SNE

Table of Contents

Speaker Recognition

Speaker recognition is the task of recognizing a speaker based on the speaker's voice characteristics:

- Verification: Is this speaker A's voice? (1:1 match)
- Identification: Is this one of N speakers' voices? 1:N match)

⇝ The speaker with the highest similarity compared to target audio is inferred as the speaker

- Two types:
  - Closed set: The audio only contains speech from a known set of speakers
  - Open-set: The audio contains speech from both known and unknown speakers

  ⇝ Thresholds are necessary to discriminate between known and unknown speakers: τ_{score} and τ_{diff}

    ▷ Thresholds are considered hyperparameters that can be tuned
- Generally requires a supervised setup, but a weakly supervised setup is possible if we have auxiliary targets

Combining Tasks

Speaker recognition is often done in combination with speaker diarization and sometimes also ASR:

1. Speaker diarization $\rightsquigarrow$ timestamps for when a speech segment starts and stops with each speech segment belonging to a single but unknown speaker
2. Speaker/speech recognition $\rightsquigarrow$ Who's the speaker and what does the speaker say?

Speaker diarization functions as a preprocessing step that simplifies the subsequent tasks – when combined, you have a powerful annotation pipeline (Rask, 2023)

Weakly-Supervised Speaker Recognition (Rask, 2023)

When speech-level transcripts are available for the recording we want to annotate, we can combine speaker diarization and ASR to perform weakly-supervised speaker identification using fuzzy string matching:

1. Diarize recording $\mathcal{R}$ with signal $y(n)$ into K segments s_i with $i \in \{1, \ldots, K\}$
2. Apply ASR on each segment s_i to obtain K candidate texts
3. Obtain M auxiliary targets from transcript
4. Preprocess and vectorize texts candidates and targets into vectors $\mathbf{C}$ and $\mathbf{T}$
5. Compute pairwise similarity between each element in $\mathbf{C}$ and $\mathbf{T}$ using similarity metric $\mathcal{M}$ (e.g. cosine) and construct a $K \times M$ matrix
6. Apply matching scheme to map candidates to targets to obtain speaker names for segments $i \in \{1, \ldots, K\}$
7. Generate speaker embeddings for each segment i and assign as reference audio for each identified speaker

$\rightsquigarrow$ Joint speaker diarization, ASR, and speaker recognition using unsupervised and weakly-supervised learning

Table of Contents

Alignment

Often we want to analyze modalities simultaneously (e.g. text-audio analysis) and not only in isolation.
⇝ To do this, modalities must be *aligned*

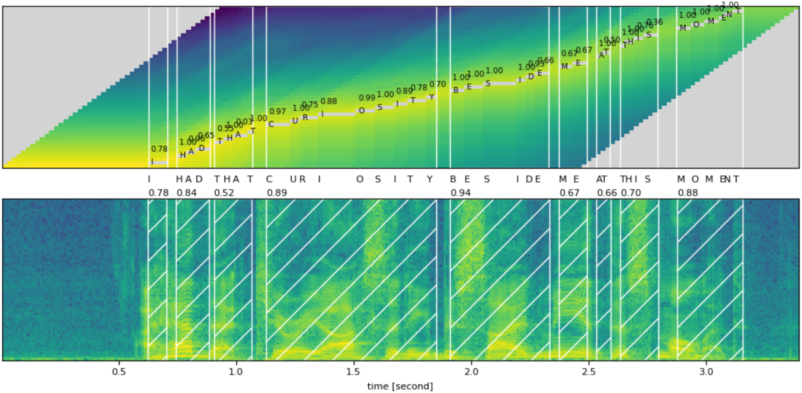
- Humans often process information conveyed in both text, audio, and images at the same time and in an interactive manner
- Combining modalities might improve predictive tasks (Rheault and Borwein, 2019)
- Alignment-level: semantic or temporal unit?
- Combining diarization and ASR is equivalent to aligning at the speech-level
- The level of alignment depends on where we think the *variation* is in each modality

Exercise

1. Discuss the difference between audio measurement when using the speech-level compared to the word-level.
2. Compare audio measures with text measures (e.g. using a dictionary) when we use speeches as the unit of analysis.

Word-Level Alignment

We can also align each audio and text at the level of each word:



- ⇝ Combines wav2vec2.0 with a phoneme model as overhead to perform character-level ASR
- ⇝ Alternative approach: Faster Whisper or WhisperX
- ⇝ Alignment is non-destructive – we can go back and forth between levels

Table of Contents

See you next week!

Theme: Audio

Computational Analysis of Text, Audio, and Images, Fall 2023

Aarhus University

References i

[1] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2018, pp. 5329–5333.

[2] M. Rask, "Automated annotation of political speech recordings," *Working Paper*, pp. 1–20, 2023.

[3] L. Rheault and S. Borwein, "Multimodal techniques for the study of a ect in political videos," Working Paper, Tech. Rep., 2019.