Intro to Speech Processing

AKA WHY ARCHIVISTS SHOULD CARE ABOUT THOSE BOXES OF TAPES

PETER SULLIVAN

ITRUST AI – Lanzarote 2022

Overview

- Why care about AV records?
- Issues with AV record accessibility
- Making records accessible through AI
  - Which language is being spoken? – Language ID
  - Who is speaking? – Speaker ID & Speaker Diarization
  - What is being said? – Automatic speech recognition (ASR)
- Spanish Language ASR Demo (OpenAI's Whisper)

Motivation

- AV processing tools lag text-based methods (Van Noord et al. 2021)
- Example AV records:
  - Taped interviews
  - Legal depositions
  - News and radio broadcasts

Driving question:

How can AI support reference and access needs for AV archives?

What are the challenges from a ML perspective to these tasks?

What do these records look like?

Production date:	1993-02-06
Description:	The evening session at the candidate training convention at the Chateau Cartier, Montreal, Quebec.
Language:	English / Français

Production date:	1975-11-27
Description:	Remarks to the news media after a meeting of the Cabinet, about: the RCMP; Jean Marchand; Marc Lalonde; party fundraising; a meeting with Maurice Nadon; and the Minister of Communications.
Language:	English

Description:	Interviews with Sue Nattrass from the Sydney Opera House [brief inturruption midway through interview] and John Mattheson from Lyric Opera of Queensland.

Title:	[Weizmann Institute of Science, Board Meeting]
Production date:	1969-10
Description:	Board of Directors meeting, held in New York.

Description:	Progressive Conservative General Meeting and scrum (28 February 1981) ; Closing remarks to Progressive Conservative General Meeting, Ottawa (1 March 1981) ; Scrum following caucus (4 March 1981).

Speech Processing Tasks

Which language is being spoken?

- Spoken Language ID

Who is speaking?

- Speaker ID (who?) and speaker diarization (when?)

What is being said?

- Automatic Speech Recognition, Spoken Language Translation

Pre-processing pipeline

Start → Down sample → Segmentation → Feature extract

Archival Master

16khz audio

1D array (amplitude)

5-15s segment

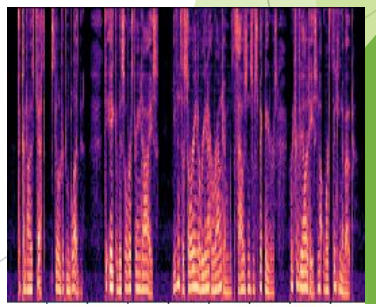
1D array (amplitude)

Spectrogram

2D array

528671-1973-0026-072.WAV

```
[-0.00048828 -0.00018311 -0.00137329 ...
  0.00085449]
```



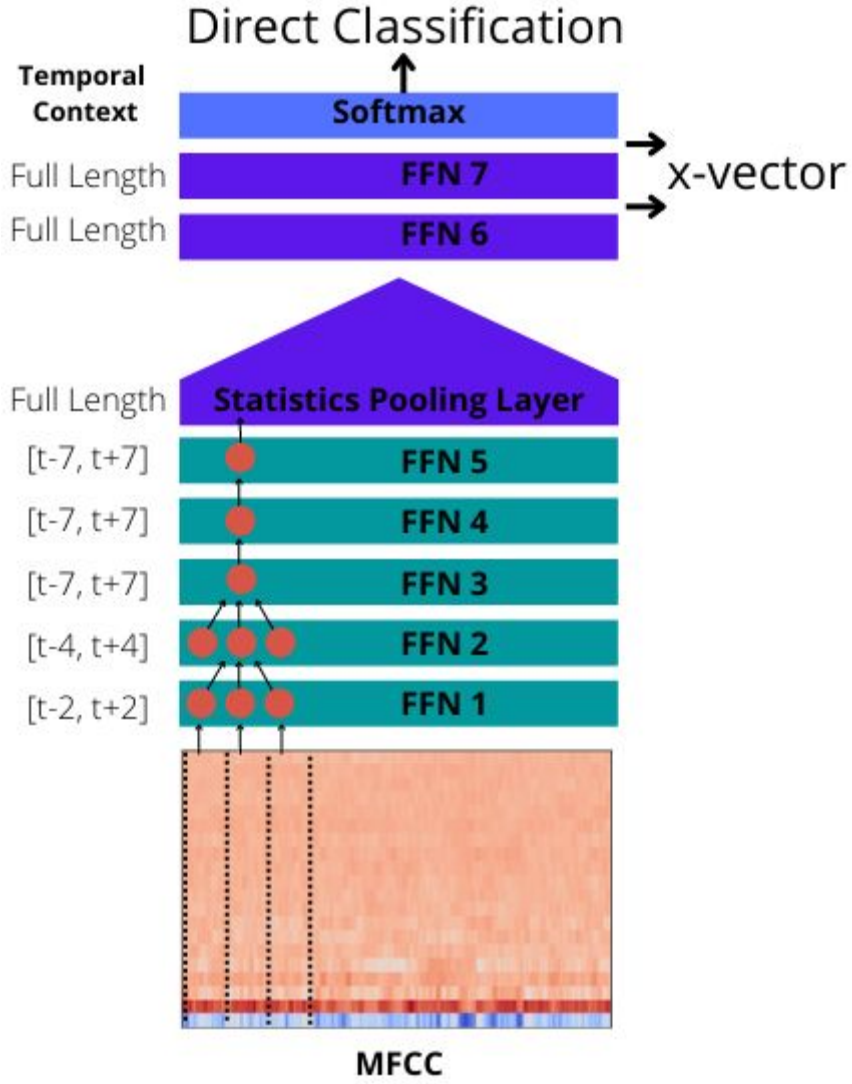
```
[[-63.278946 -55.87039  -68.521965 ... -65.844215 -55.37684  -60.337776]
 [-58.574966 -55.75315  -63.788086 ... -66.275925 -56.51124  -59.232697]
 [-56.25997  -55.574905 -61.46692  ... -66.709    -58.043304 -58.268936]
 ...
```

Language ID / Speaker ID

Input: Spectrogram or .wav

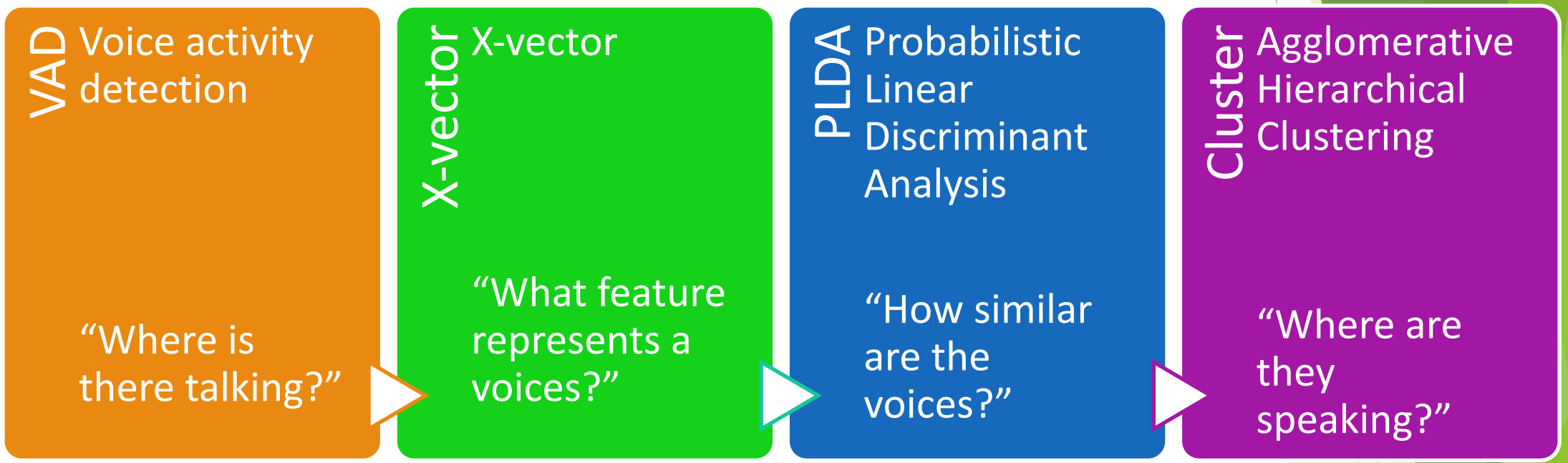
Direct Classification:
Know all possible options – softmax

Open set classification:
Similarity of two x-vectors



ginal x-vector

Speaker diarization overview



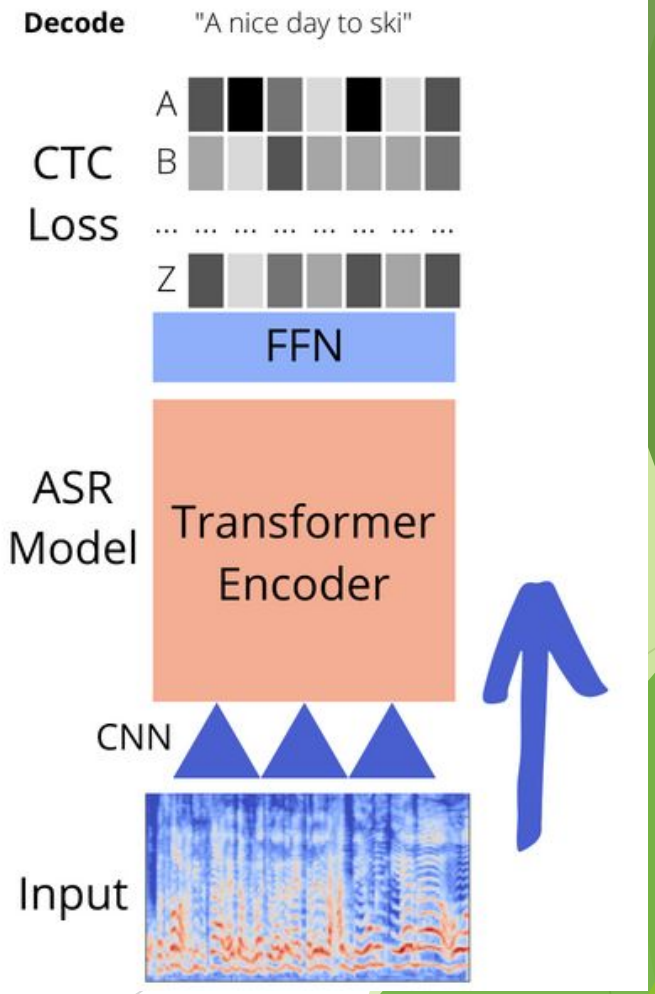
Automatic Speech Recognition (CTC)

Input: Spectrogram

Alignment: Connectionist temporal classification
(Graves et al. 2006)

Decoding: Dictionary re-weight of word probabilities
[See Hannun et al. 2014]

Pre-trained models:
HuBERT (Hsu et al. 2021),
wav2vec 2.0 (Baevski et al. 2020)
(these use 1d amplitude arrays as input)



Challenges

- Non-native speakers
- Cross talking
- Recording artefacts
- Noise
- Language differences from text
- Codemixing
- Amount of audio

Resources

- Feature extraction
  - Librosa
  - torchaudio
- Voice activity detection
  - WebRTC
- Speaker Diarization
  - pyAudioAnalysis
- Language Model Decoding
  - PyCTCDecode
- Pre-trained models
  - Huggingface model hub
  - Open AI Whisper
- Frameworks
  - SpeechBrain
  - Kaldi
- Paid Platforms
  - Microsoft Azure Video indexer

Whisper Demo

- QR Link to Colab Demo:

Bibliography

- Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, *33*, 12449-12460.
- Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006, June). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning* (pp. 369-376).
- Hannun, A. Y., Maas, A. L., Jurafsky, D., & Ng, A. Y. (2014). First-pass large vocabulary continuous speech recognition using bi-directional recurrent DNNs. *arXiv preprint arXiv:1408.2873*.
- Hsu, W. N., Bolte, B., Tsai, Y. H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 3451-3460.
- Park, T. J., Kanda, N., Dimitriadis, D., Han, K. J., Watanabe, S., & Narayanan, S. (2022). A review of speaker diarization: Recent advances with deep learning. *Computer Speech & Language*, *72*, 101317.
- Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018, April). X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 5329-5333). IEEE.
- Van Noord, N., Olesen, C. G., Ordelman, R., & Noordegraaf, J. (2021, February). Automatic Annotations and Enrichments for Audiovisual Archives. In *ICAART (1)* (pp. 633-640).