UNESCO Radio Archives: AI for Audio Metadata Enrichment

Peter Sullivan – UBC iSchool
Eng Sengsavang – UNESCO Archives

Driving Question

How can AI enable better description of archival audio?

Project Goals

Understand relationship between a physical record and its digital surrogate

Evaluate whether diplomatic analyses apply to various genres of audio recordings

Analyze whether and how AI models improve performance during metadata creation, control, and enrichment.

Analyze risks, challenges, and potential biases

UNESCO Radio Archives At A Glance

Size: ~16,000 Recordings (~1000 described)

UNESCO Radio Archives At A Glance

Size: ~16,000 Recordings (~1000 described)

Time: 1950s – 1980s

UNESCO Radio Archives At A Glance

Size: ~16,000 Recordings (~1000 described)

Time: 1950s – 1980s

Languages: 70+ (Including multilingual)

UNESCO Radio Archives At A Glance

Size: ~16,000 Recordings (~1000 described)

Time: 1950s – 1980s

Languages: 70+ (Including multilingual)

Topic Coverage:

Education Culture Natural Sciences

Communication and Information

Social and Human Sciences UNESCO History

UNESCO Radio Archives At A Glance

Size: ~16,000 Recordings (~1000 described)

Time: 1950s – 1980s

Languages: 70+ (Including multilingual)

Topic Coverage:

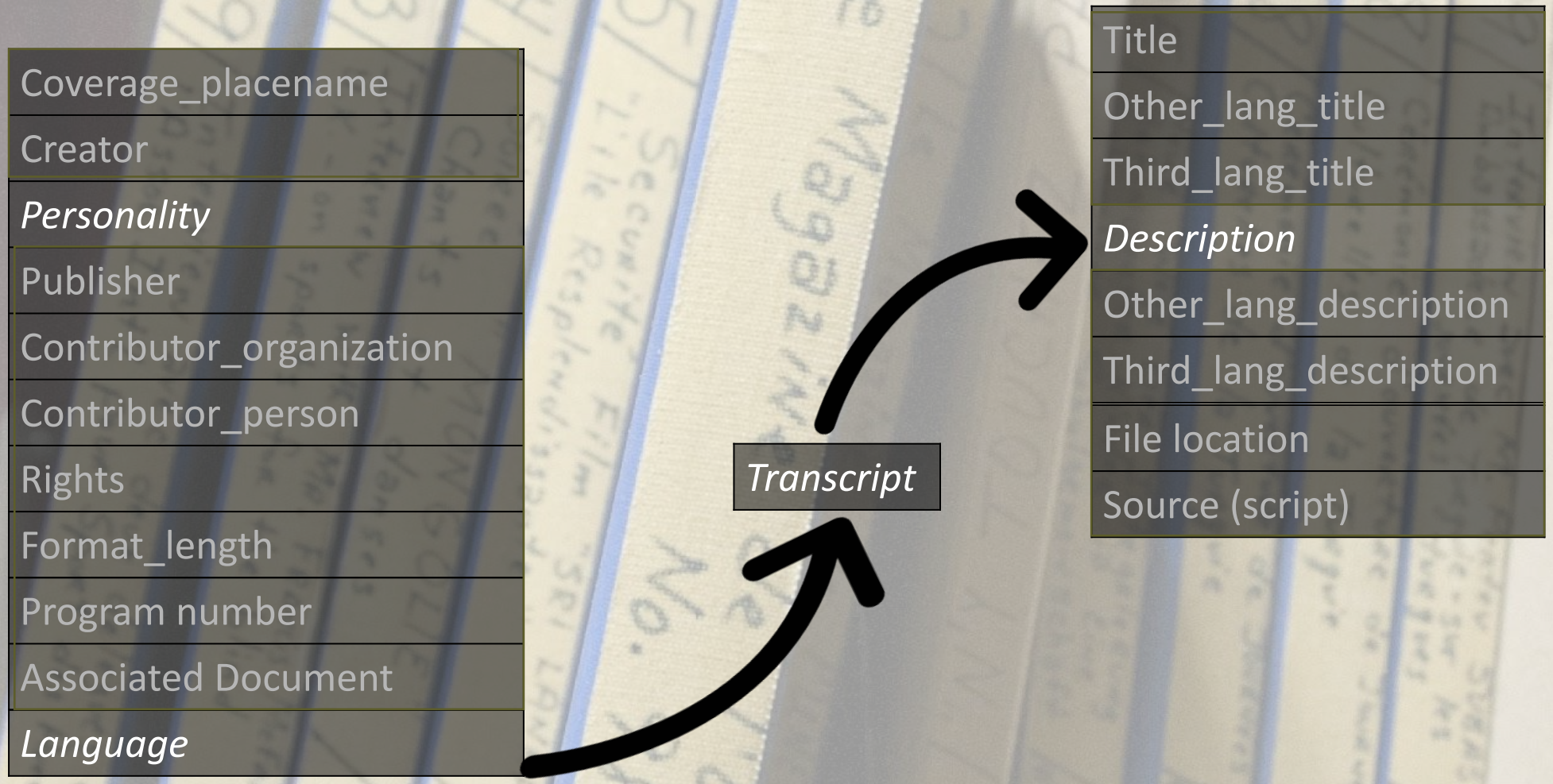
Education Culture Natural Sciences

Communication and Information

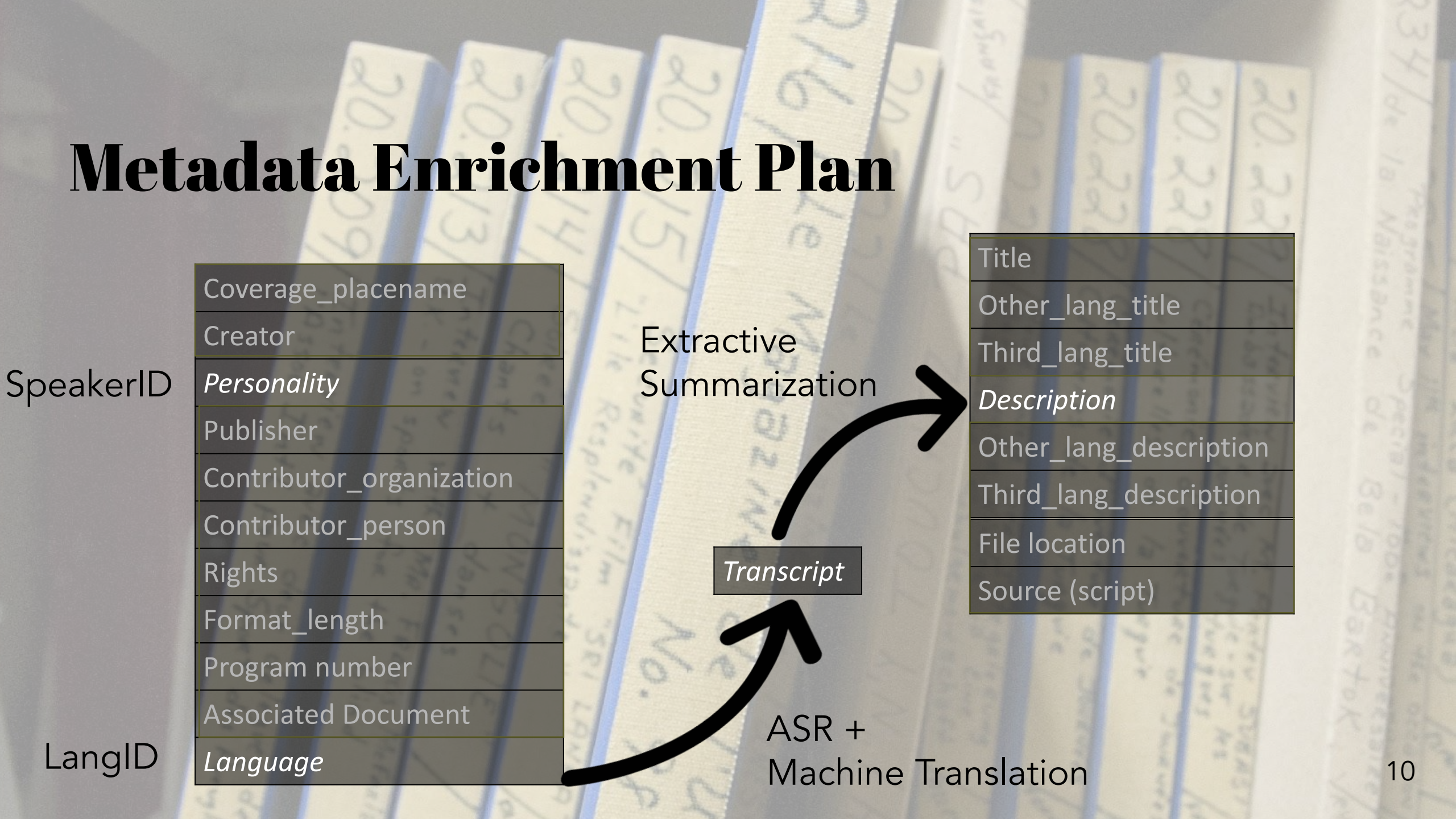
Social and Human Sciences UNESCO History

Genre: Interviews Speeches Educational Program

Metadata Enrichment Plan



Metadata Enrichment Plan



The story so far...

Initial focus group to examine ASR transcripts and Language ID

- Transcripts reasonable IF language ID predictions correct

Started diplomatic analysis of transcripts (ongoing)

- Identify structure that can enable easy extraction of description

Zero Shot Language ID & Speaker Embedding Experiments

Initial Language ID Lessons

Errors w/ certain language pairs

English predicted as being Welsh, Spanish predicted as Galician.

Ambiguity with Balkan languages.

A11406.gl.txt – IDed as Galician
0.0-4.0: MÉXICO, PRÓXIMA
12.0-15.0: Aquí, la UNESCO en Paris.
15.0-20.0: Tenemos el gusto e el inmenso placer de recibir en nuestros estudios

A11903.gl.txt – IDed as Galician
0.0-3.0: Aquí, na Unesco, en París.
3.0-6.0: Temos o gozo de receber nos estudios
6.0-8.0: o professor Alain Jox

The story so far...

Initial focus group to examine ASR transcripts and Language ID

- Transcripts reasonable IF language ID predictions correct

Started diplomatic analysis of transcripts (ongoing)

- Identify structure that can enable easy extraction of description

Zero Shot Language ID & Speaker Embedding Experiments

Diplomatics

Consistent form across similar types of recordings.

Example: First 30 seconds of interviews almost always include who is being interviewed, and about what.

0.0-8.0: I have with me in the studio today Dr. George Stoddard, Dean of the School of Education of New York University.
8.0-21.0: He's one of the experts who've been convened to a UNESCO meeting to study the effect of mass media, that's films, press and radio, on juvenile delinquency. Dr. Stoddard.
22.0-36.0: Well, really the purpose of the meeting is somewhat broader than that. It's to study all the influences of the mass media, particularly the cinema and television, on the behavior of children.

0.0-7.0: Now here in the studio is Mr. Frederick Bellinger from the United States of America and for
8.18-15.18: the past year he's been in Egypt working for UNESCO Technical Assistance. Now I believe
15.22-22.06: Mr. Bellinger, you are a chemical engineer. What exactly were you doing in Egypt?
22.06-28.64: Under the UNESCO Technical Assistance Program I was asked to go to Egypt to assist the National
28.64-35.64: Research Council in definitely planning and starting a basic and applied research effort
36.24-39.76: in the industrial chemistry field.

Diplomatics

Consistent form across similar types of recordings.

Example: Last ~10 seconds “sign off” for reports and programs or thanking guests for interviews.

901.0-905.0: Sound of radio
910.0-916.0: Those noises from beyond the Earth mark the end of this, the fourth program in the series,
916.0-919.0: Signposts for the Atomic Age.
919.0-922.0: They are edited and introduced by Richie Calder.
922.0-927.0: The program was produced by Rex Keating in the studios of UNESCO, Paris.

210.26-217.3: This is Professor C. N. Vakil speaking from UNESCO headquarters in Paris and returning
217.3-220.18: you to UN Radio in New York.

285.84-289.04: That sounds very encouraging. Well thank you very much indeed Mr. Ballinger.

830.0-832.0: Thank you very much, Dr. *Stallone*

The story so far...

Initial focus group to examine ASR transcripts and Language ID

- Transcripts reasonable IF language ID predictions correct

Started diplomatic analysis of transcripts (ongoing)

- Identify structure that can enable easy extraction of description

Zero Shot Language ID & Speaker Embedding Experiments

Motivation

Right: Vittorino Veronese

(UNESCO General Director 1958-1961)

Recordings in:

French, Italian, Spanish, English, German

Photo Credit:
https://unesdoc.unesco.org/ark:/48223/pf0000067034

LangID

How robust are Language ID models to L2 speakers?

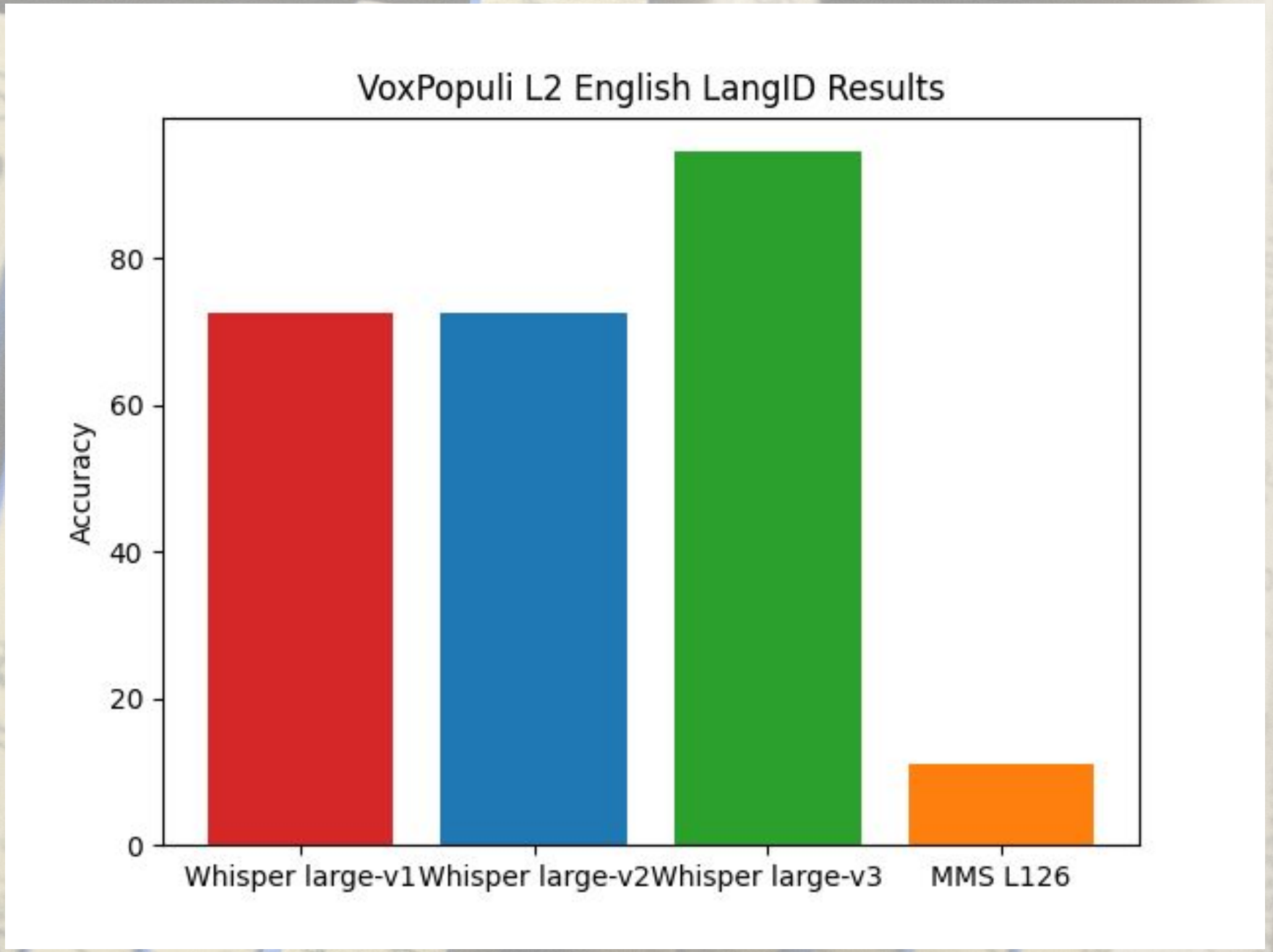
Data: L2 English VoxPopuli [4]
UNESCO Multilingual Speakers (L1 & L2)

Models: Whisper (large-v1,2,3) [3], MMS (l126) [2]

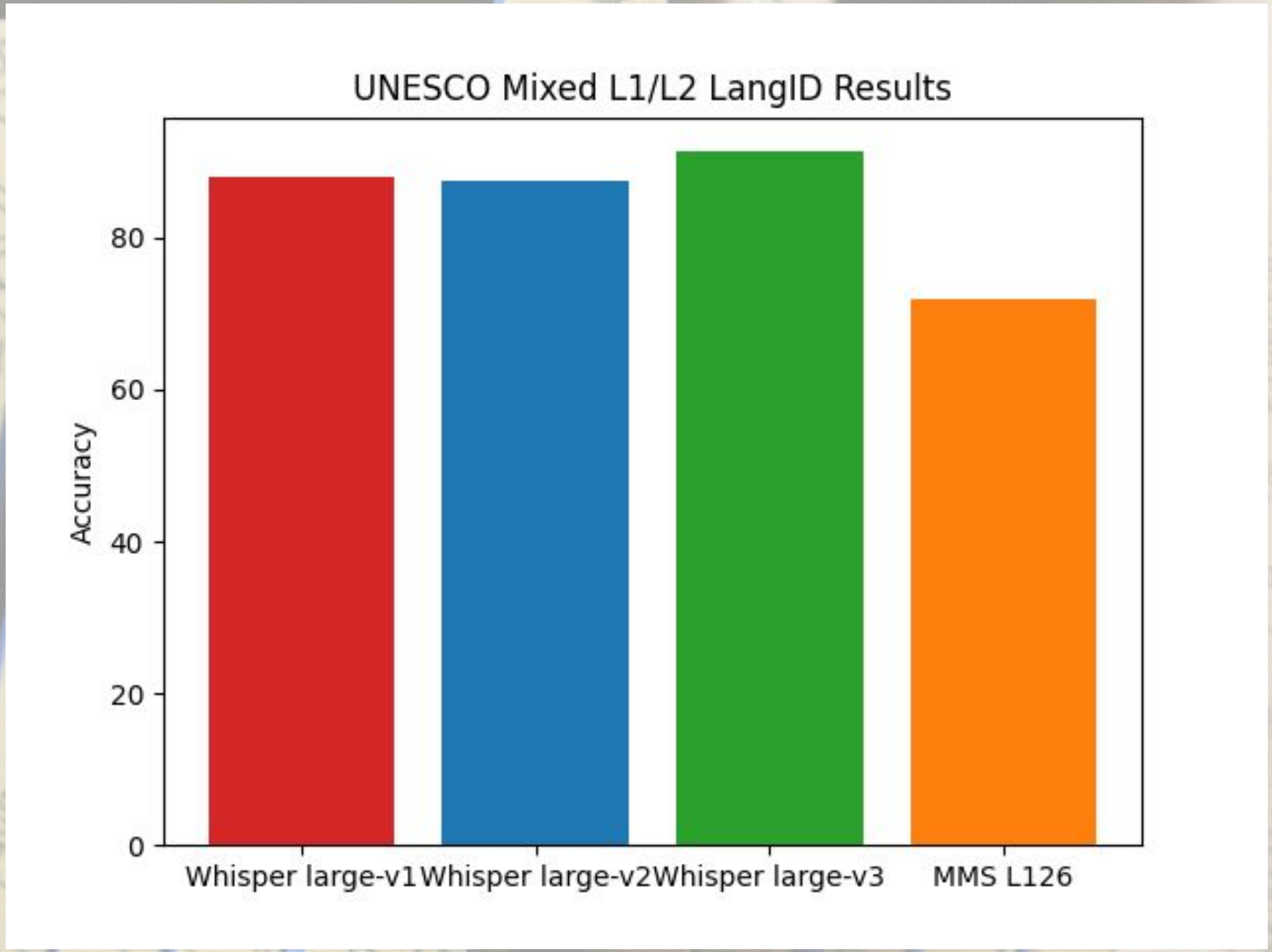
Benchmark off the shelf LID tools on L2 subsets

MMS L126: brittle?

Whisper v3: a major improvement on L2 English



Whisper v3 still best option



Speaker Processing in Brief

Embeddings represents the characteristics of a speaker in a fixed-size array

Speaker Embedding Extractor (i-,x-,r-vector etc.)

Embedding

The same speaker *should* have similar embeddings

Speaker Processing for Archives

Target	Use	How	Challenges
Speaker Verification	Open Set	Security	PLDA Classifier	Demographic [1], GenAI
Speaker ID	Closed Set	Security	Direct Classification	Demographic [1], GenAI
Speaker Diarization	Open Set	Pre-Processing Transcripts	Clustering	ASR integrations [2] Cross talk [2], Domain [2]
Speaker Indexing For Archives	Closed Set w/ "others"	Information Retrieval	Cluster / Direct / Domain Adaptation ?	Multilinguality, Aging, Channel ?

Speaker Processing for Archives

Target	Use	How	Challenges
Speaker Verification	Open Set	Security	PLDA Classifier	Demographic [1], GenAI
Speaker ID	Closed Set	Security	Direct Classification	Demographic [1], GenAI
Speaker Diarization	Open Set	Pre-Processing Transcripts	Clustering	ASR integrations [2] Cross talk [2], Domain [2]
Speaker Indexing For Archives	Closed Set w/ "others"	Information Retrieval	Cluster / Direct / Domain Adaptation ?	Multilinguality, Aging, Channel ?

Cross-lingual Embeddings

How robust are embeddings for multilingual speakers?

Data: Single speaker recordings w/ multiple recordings in multiple languages

Diarize and filter out recordings without a majority speaker

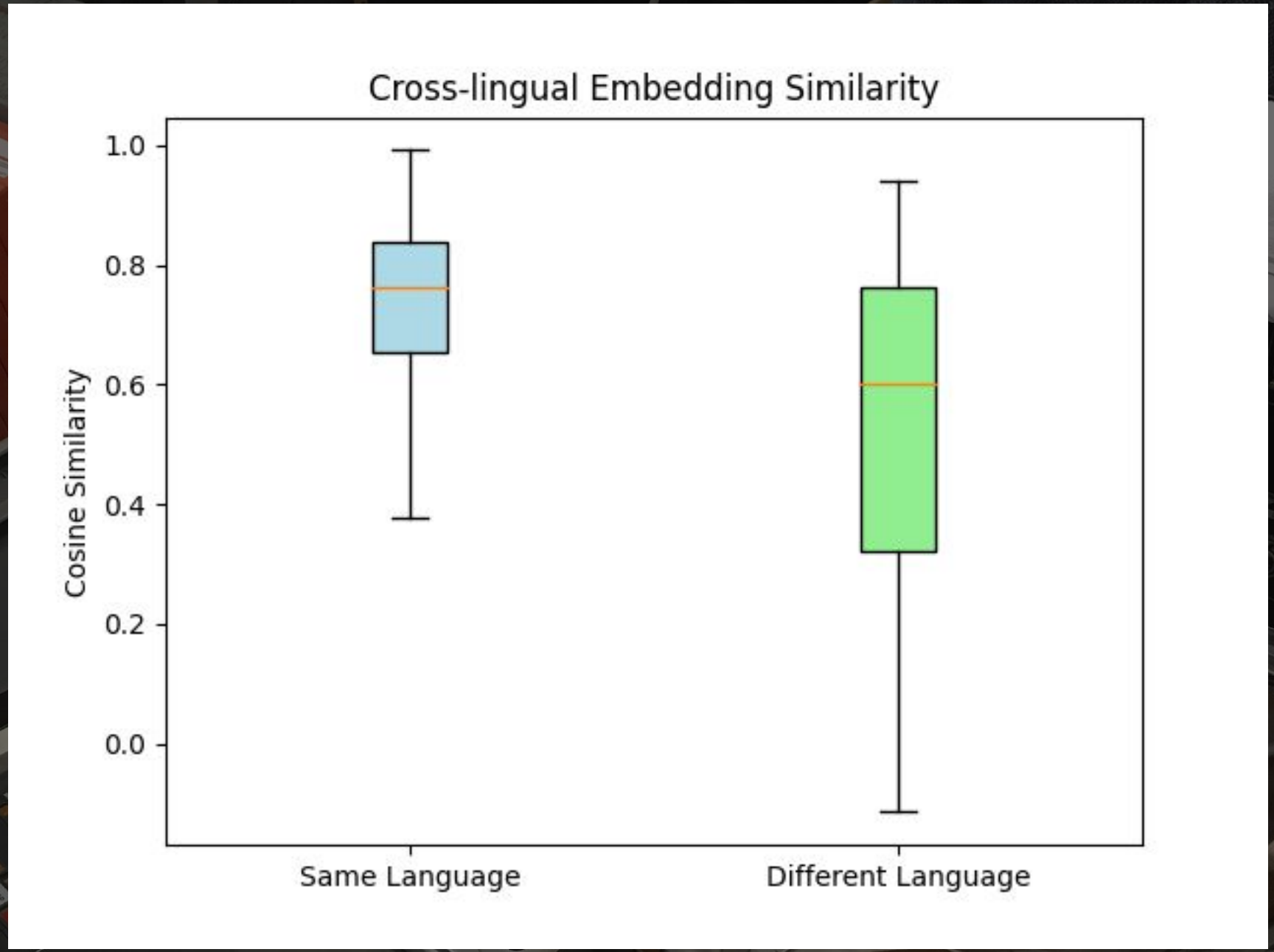
(n_speaker =75, n_recordings= 363, n_langs= 16)

Model: WeSpeaker Resnet34 [5]

Extract mean embeddings and compare monolingual vs. cross-lingual similarity

Monolingual:
Mean: 0.71
Median: 0.76
Std: 0.19

Cross-lingual:
Mean: 0.53
Median: 0.60
Std: 0.26



25

Future Steps

Diplomatics informed extractive summarization

Transcription and translation quality analysis

Domain adapted indexing of speakers


Mahalo!

Bibliography

[1] Hutiri, W. T., & Ding, A. Y. (2022, June). Bias in automated speaker recognition. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 230-247).

[2] 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.

[3] Pratap, V., Tjandra, A., Shi, B., Tomasello, P., Babu, A., Kundu, S., ... & Auli, M. (2023). Scaling speech technology to 1,000+ languages. *arXiv preprint arXiv:2305.13516*.

[4] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023, July). Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning* (pp. 28492-28518). PMLR.

[5] Wang, C., Rivière, M., Lee, A., Wu, A., Talnikar, C., Haziza, D., ... & Dupoux, E. (2021, August). VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation. In *ACL 2021-59th Annual Meeting of the Association for Computational Linguistics*.

[6] Wang, H., Liang, C., Wang, S., Chen, Z., Zhang, B., Xiang, X., ... & Qian, Y. (2023, June). Wespeaker: A research and production oriented speaker embedding learning toolkit. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1-5). IEEE.