On Barriers to Archival Audio Processing

Peter Sullivan
University of British Columbia, Canada
prsull@student.ubc.ca

Muhammad Abdul-Mageed
University of British Columbia, Canada
muhammad.mageed@ubc.ca

Abstract

In this study, we leverage a unique UNESCO collection of mid-20th century radio recordings to probe the robustness of modern off-the-shelf language identification (LID) and speaker recognition (SR) methods, especially with respect to the impact of multilingual speakers and cross-age recordings. Our findings suggest that LID systems, such as Whisper, are increasingly adept at handling second-language and accented speech. However, speaker embeddings remain a fragile component of speech processing pipelines that is prone to biases related to the channel, age, and language. Issues which will need to be overcome should archives aim to employ SR methods for speaker indexing.

1 Introduction

Multinational organizations such as the United Nations (UN); the International Federation of Red Cross and Red Crescent Societies (IFRC); and the United Nations Educational, Scientific, and Cultural Organization (UNESCO) maintain audio archives that are of profound cultural and historical value. However, incomplete descriptive metadata often hinder their access by the public (Zervanou et al., 2011).

Audio archives present a complex terrain for contemporary speech processing technologies, owing to the varied domains these recordings encapsulate. The long running MALACH project's effort to tackle emotional, disfluent, and accented speech (Picheny et al., 2019; Psutka et al., 2002) gives of sense of this complexity. In this study, our attention is directed towards extensive multilingual repositories, which pose challenges for speaker recognition (SR) and language identification (LID) technologies due their long chronological span and inclusion of second-language (L2) speech.

A primary objective in enhancing accessibility to these recordings involves identifying the speakers within a specific recording. While documentation of speakers is *sometimes* available, it is only at the document level, making this task closer to speaker indexing (Sturim et al., 2001) than standard SR. Moreover, concerns regarding the robustness of speaker embeddings to voice modifications associated with aging, as well as the accuracy of language-specific predictions (Hutiri and Ding, 2022), significantly challenge the straightforward utilization of off-the-shelf SR technologies within the ambit of these long-running, multilingual archives.

In this exploratory investigation using a selection of radio audio archives from UNESCO, we explore the impact of these factors on robustness of zero-shot application of off-the-shelf tools, to identify paths towards speaker indexing in age- and language-variable environments. Our dataset involves 171 hours of archival data covering the period of 1952-1980, involving 20 languages (See Table 1). Our work offers the following contributions:

1. We characterize a relatively neglected sphere within speech processing scholarship: multilingual audio archives.

2. We carry out a cross-age analysis to investigate robustness of the speaker embeddings.

3. We analyze the robustness of speaker embeddings in multilingual speech scenarios, uncovering unique insights that are otherwise hidden without access to datasets such as ours.

4. We investigate the utility of a number of off-the-shelf language identification tools for accented LID.

The rest of the paper is organized as follows: Section 2 offers a brief overview of speech processing as it relates to challenges present in archival audio. Section 3 is a description of our datasets.

Section 4 details our data processing methods. Section 5 gives an overview of the LID and SR models used. Section 6 is a breakdown of our different experiments. In Section 7, we provide our experimental results and discuss these in Section 8. We discuss limitations of our work in Section 9 and provide ethical implications in Section 10. We conclude in Section 11.

2 Background

2.1 Language ID

Recent work in LID has moved from the discrete segment representations popular in earlier i-vector (Dehak et al., 2011) and x-vector (Snyder et al., 2018) works, to the convenience of end-to-end deep neural models either based on ResNets (Cai et al., 2018) or Transformers (Babu et al., 2021; Radford et al., 2023; Pratap et al., 2023). Part of this has been enabled by the creation of LID datasets such as VoxLingua107 (Valk and Alumäe, 2021) and FLEURS (Conneau et al., 2023), allowing for direct training of large end-to-end models. Meanwhile, another aspect of this evolution has been the inclusion of LID into models designed as 'jack-of-all-trades' tools such as Whisper (Radford et al., 2023) and Massively Multilingual Speech (MMS) (Pratap et al., 2023), which are built to support automatic speech recognition (ASR) and speech translation in addition to LID.

Recent work in LID has raised attention to less well developed areas of exploration including L2 LID (Kukk and Alumäe, 2022), LID for multilingual users (Titus et al., 2020), and domain generalization of LID (Sullivan et al., 2023). For archival audio, these three aspects become even more relevant, as the long chronological nature of the archives and diverse population of multilingual speakers demand the use of highly robust LID systems. At the same time, the operational context (including lack of funding and know-how) means that off-the-shelf tools are the only viable option, neglecting transfer learning or domain adaptation approaches, regardless of their efficacy.

2.2 Speaker embedding

Modern SR makes use of speaker embeddings (i.e. x-vector, r-vector etc.) often extracted from ResNets-based extractor networks (Desplanques et al., 2020; Wang et al., 2023) and trained on VoxCeleb (Nagrani et al., 2017). Despite the progress made in the field, a recent study of bias in

Table 1: Dataset Overview. n is the number of recordings. For LID, we use only the first 30 seconds of spoken audio from each recording.

Split	n	Hours	Languages
LID	484	4	20
Cross-age	692	104	17
Cross-lingual	463	67.5	20

SR (Hutiri and Ding, 2022) indicate a number of issues, including demographic biases impacting age, gender, language, and nationality. To our knowledge little work has been performed to investigate the impact of language (including multilinguality) on SR effectiveness. However, there is a body of work that has looked at the impact of cross-age scenarios on SR (Singh et al., 2023; Qin et al., 2022). With Qin et al. (2022) employing adversarial learning to learn age-invariant embeddings, and Singh et al. (2023) estimating the impact of age on cross-age verification. To our knowledge no work has considered investigation of cross-age scenarios in multilingual corpora.

Work in speaker indexing, has been largely neglected since anchor model based approaches (Sturim et al., 2001), but we hope that this study will pave the way to future work in speaker indexing by identifying existing obstacles with off-the-shelf tools.

3 Datasets

VoxPopuli - L2 English An existing source of L2 audio, in a similar domain to the one under consideration is the VoxPopuli collection of EuroParl recordings (Wang et al., 2021). In particular we are interested in the 29 hour subset of L2 English speech coming from 15 different accent backgrounds.

Radio Archives From our partnership with UNESCO, we utilize a mid-20th century radio recording archives collection that is currently undergoing metadata enrichment. While the dataset is currently not published, the metadata used and links to the recordings will be made available. The recordings in this collection consist of a variety of material including speeches, interviews, news briefs, and educational programs. The audio was recorded between the 1950s and early 1980s, and digitized in the late 2010s. See Table 1 for more detailed statistics.

4 Data Processing

While the VoxPopuli dataset is used with existing segmentation for our L2 English experiments, we build two datasets from the radio archives by filtering the known metadata to meet certain criteria. As much of the radio archives contain incomplete metadata, we restrict our selection to recordings identified as having a single known speaker on the recording,as well as a single languages spoken.

From this selection, we create two datasets: Our *cross-age* dataset is filtered by selecting speakers with multiple recordings in the same language occurring in different calendar years. The second *cross-lingual* dataset is filtered by selecting speakers with multiple recordings in different languages.

For both datasets, we further filter by diarizing the recordings using Pyannote's (Bredin et al., 2020) speaker diarization pipeline (version 3.1), and selecting recordings where there is a single speaker accounting for more than 75% of the duration. We utilize 16khz single channel copies of the recordings for the study.

5 Models

Whisper (Radford et al., 2023) is a multilingual speech processing model that allows for ASR, speech translation, and LID. We use all three versions of the large (1.5B parameter) model.
MMS (Pratap et al., 2023) is similarly a multilingual speech processing model allowing for ASR, speech translation, LID, and additionally text-to-speech. The model has been primarily trained on the Bible and other religious audio, with a focus on scaling the number of covered languages. We use the (1B parameter) 126 language version.
WeSpeaker ResNet34-LM (Wang et al., 2023) is a SR model trained on VoxCeleb (Nagrani et al., 2017).

6 Experiments

6.1 L2 LID

For our L2 LID experiments, we compare two well known off-the-shelf LID models: Whisper Large V(1-3) (Radford et al., 2023) and MMS L126 (Pratap et al., 2023). For our VoxPopuli dataset we simply report predictions on the given segments. However, for our cross-lingual radio dataset, we follow the suggested LID procedure of the Whisper model and select the first 30 seconds of audio to perform the prediction.

6.2 Speaker Embedding Robustness

To understand how robust speaker embeddings are to cross-age and cross-lingual effects, we use a pretrained ResNet model, Wespeaker ResNet34-LM (Wang et al., 2023), to extract segment representations. We take the average embedding for the majority speaker in each recording. For the *cross-age experiments*, we calculate the cosine similarity between representations and aggregate by difference in calendar years between recordings. We rely on calendar year as an approximation as we do not have recording dates for some recordings. For the *cross-lingual experiments*, we simply compare the cosine similarity between same language segments to their different language similarity scores.

7 Results

For L2 English performance (see Table 2 in Appendix A), The updated Whisper V3 model substantially improves the recognition of accented English audio attaining an accuracy of 94% compared to both Whisper V1,V2 and MMS L126.

For our mixed L1 and L2 multilingual archival audio (see Table 3 in Appendix A), we see similar results as to the VoxPopuli results. Notably, MMS performs better on the mixture of accented and unaccented speech (accuracy at 71.90%). Still, the Whisper V3 model performs best on this set of audio: accuracies of 88.01% for V1, 87.60 for V2, and 91.32% for V3.

Looking at the robustness of the speaker embeddings, we notice a substantial drop in the similarity scores in both the cross-age setting (see Fig. 1) and cross-lingual setting (see Fig. 2). For the cross-age comparison, median cosine similarity scores continue to drop until stabilizing after a gap of 10 years. While cross-lingual performance of representations drops substantially as well, of note is the very large increase in standard deviation between the two settings. This latter trend is potentially problematic for treating all cross-lingual scenarios the same, and may be indicative that fluency as well as language similarity between compared languages may be factors.

8 Discussion

The performance of the Whisper V3 model appears to indicate it as a strong candidate for archives in processing multilingual audio files. Notably, while all of the Whisper models under examination have roughly the same number of parameters (V3 is
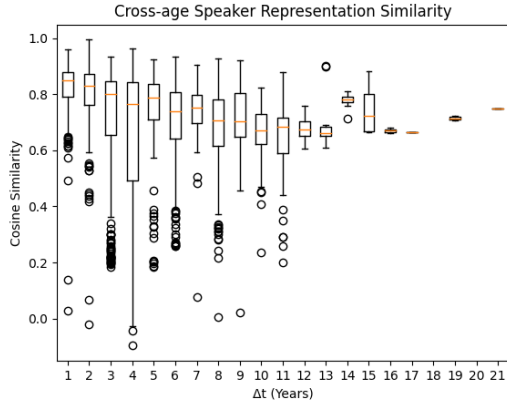
Figure 1: Longitudinal comparison of speaker embedding. Cosine similarity scores are plotted with median and quartiles marked. Outliers are noted as circles. Δt is the absolute difference in calendar years between each pair of recordings. Data past 15 years becomes quite sparse, as few speakers fit our filtering criteria.
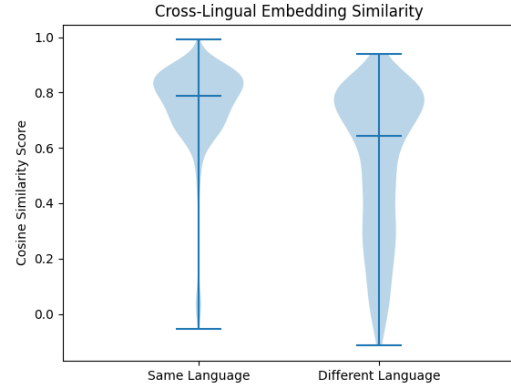

Figure 2: Cross-lingual comparison of speaker embedding cosine similarity scores. To better show the distribution, we present the results as violin plots, noting that the lower end of the Same Language plot may be representative of different speakers who were not filtered out of our automated approach. Distribution statistics: *Same language* Mean: 0.71, Median: 0.76, Std: 0.19; *Different language* Mean: 0.53, Median: 0.60, Std: 0.26

slightly larger due to increasing the input dimensions), the larger amount of training data used for V3 appear to substantially help it in identifying accented English. The brittle nature of the MMS model with regard to L2 speech is quite surprising, and demonstrates the importance of having diverse and challenging audio benchmarks for LID. While the longitudinal embedding comparison demonstrates the challenges of applying SR models across channel and age. The appearance of a flattening at around 10 years of difference indicates that it may be possible to account for this cross-age drift.

9 Limitations

Working with real world datasets presents substantial challenges and limitations. While the archival partner aims to put the radio recordings online sometime in the future, these are not yet available publicly. Similarly, working with aggregated data has limitations. For instance the small cluster of low similarity in the same language embedding comparison is likely indicative of misidentified speakers who slipped through the filtering process.

10 Ethics

Speaker identification tools have the potential for significant privacy violation, especially if applied as part of a surveillance system at scale. As seen from the study, the large cross-channel and cross language variability in speaker embeddings present significant risks for use in larger scales, where risk

associated with misidentification may also be quite high.

We believe our investigation of these tools for the purpose of improving archival accessibility and discovery are consistent with ethical practice, and our application to a limited closed set of public figures on public audio recordings further limits this potential for abuse.

11 Conclusion

We present an analysis into the robustness of SR and LID tools on cross-age, and cross-lingual audio on a unique archival radio dataset. Our work indicates the viability of existing LID tools, such as Whisper V3 (Radford et al., 2023), at handling the accented and multilingual speech common in recordings from multinational organizations. We also find that cross-age and cross-lingual application of SR introduce large drops in the cosine similarity scores, with the cross-age similarity dropping steadily over a period of a decade. The cross-lingual scores further introducing large increase in the standard deviation, potentially indicative of other factors such as language fluency or language similarity impacting the result. Additionally, this work demonstrates the value of archival audio in examining current speech processing approaches. The variety of such recordings offer a strong platform to study bias and domain adaptation.

References

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al. 2021. XLS-R: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*.

Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. 2020. Pyannote. audio: neural building blocks for speaker diarization. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7124–7128. IEEE.

Weicheng Cai, Zexin Cai, Xiang Zhang, Xiaoqi Wang, and Ming Li. 2018. A novel learnable dictionary encoding layer for end-to-end language identification. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5189–5193. IEEE.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805. IEEE.

Najim Dehak, Pedro A. Torres-Carrasquillo, Douglas Reynolds, and Reda Dehak. 2011. Language recognition via i-vectors and dimensionality reduction. In *Proc. Interspeech 2011*, pages 857–860.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In *Proc. Interspeech 2020*, pages 3830–3834.

Wiebke Toussaint Hutiri and Aaron Yi Ding. 2022. Bias in automated speaker recognition. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 230–247.

Kunnar Kukk and Tanel Alumäe. 2022. Improving Language Identification of Accented Speech. In *Proc. Interspeech 2022*, pages 1288–1292.

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. VoxCeleb: A large-scale speaker identification dataset. In *Proc. Interspeech 2017*, pages 2616–2620.

Michael Picheny, Zoltán Tüske, Brian Kingsbury, Kartik Audhkhasi, Xiaodong Cui, and George Saon. 2019. Challenging the Boundaries of Speech Recognition: The MALACH Corpus. In *Proc. Interspeech 2019*, pages 326–330.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al. 2023. Scaling speech technology to 1,000+ languages. *arXiv preprint arXiv:2305.13516*.

Josef Psutka, Pavel Ircing, Josef V Psutka, Vlasta Radová, William J Byrne, Jan Hajič, Samuel Gustman, and Bhuvana Ramabhadran. 2002. Automatic transcription of Czech language oral history in the MALACH project: Resources and initial experiments. In *Text, Speech and Dialogue: 5th International Conference, TSD 2002 Brno, Czech Republic, September 9–12, 2002 Proceedings 5*, pages 253–260. Springer.

Xiaoyi Qin, Na Li, Weng Chao, Dan Su, and Ming Li. 2022. Cross-Age Speaker Verification: Learning Age-Invariant Speaker Embeddings. In *Proc. Interspeech 2022*, pages 1436–1440.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Vishwanath Pratap Singh, Md Sahidullah, and Tomi Kinnunen. 2023. Speaker Verification Across Ages: Investigating Deep Speaker Embedding Sensitivity to Age Mismatch in Enrollment and Test Speech. In *Proc. INTERSPEECH 2023*, pages 1948–1952.

David Snyder, Daniel Garcia-Romero, Alan McCree, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. 2018. Spoken Language Recognition using X-vectors. In *Proc. The Speaker and Language Recognition Workshop (Odyssey 2018)*, pages 105–111.

Douglas E Sturim, Douglas A Reynolds, Elliot Singer, and Joseph P Campbell. 2001. Speaker indexing in large audio databases using anchor models. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (cat. no. 01ch37221)*, volume 1, pages 429–432. IEEE.

Peter Sullivan, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2023. On the Robustness of Arabic Speech Dialect Identification. In *Proc. INTERSPEECH 2023*, pages 5326–5330.

Andrew Titus, Jan Silovsky, Nanxin Chen, Roger Hsiao, Mary Young, and Arnab Ghoshal. 2020. Improving language identification for multilingual speakers. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8284–8288. IEEE.

Jörgen Valk and Tanel Alumäe. 2021. VoxLingua107: a dataset for spoken language recognition. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 652–658. IEEE.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. *arXiv preprint arXiv:2101.00390*.

Hongji Wang, Chengdong Liang, Shuai Wang, Zhengyang Chen, Binbin Zhang, Xu Xiang, Yanlei Deng, and Yanmin Qian. 2023. Wespeaker: A research and production oriented speaker embedding learning toolkit. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Kalliopi Zervanou, Ioannis Korkontzelos, Antal Van Den Bosch, and Sophia Ananiadou. 2011. Enrichment and structuring of archival description metadata. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 44–53.

Appendix A LID Experiments

Table 2: L2 English LID Performance of Whisper (Radford et al., 2023) and MMS (Pratap et al., 2023)

Model	Accuracy
Whisper-Large v1	72.65%
Whisper-Large v2	72.65%
Whisper-Large v3	94.52%
MMS L126	11.10%

Table 3: Multilingual Archival Audio LID Performance of Whisper (Radford et al., 2023) and MMS (Pratap et al., 2023)

Model	Accuracy
Whisper-Large v1	88.01%
Whisper-Large v2	87.60%
Whisper-Large v3	91.32%
MMS L126	71.90%