ASR Papiamentu Corpus Report

Sonny George, Chris Tam

November 15, 2023

1 Introduction

For this project, we present a corpus of Papiamentu speech and transcripts for ASR applications. It consists of a collection of audio files and a master spreadsheet containing clip information and audio segmentations:

column	description
utterance_id	numeric identifier
file_name	file name in master .zip
train_dev_test_split	train/test/dev assignment
duration_ms	utterance duration (ms)
unpunctuated_text	text content w/out punctuation
speaker	speaker name
embedding	resemblyzer embedding (len 256)
speaker_was_inferred	whether speaker was inferred from heuristic

As well as a speaker table containing relevant metadata:

column	description
speaker	speaker name
speaker_is_male	true if male, false otherwise
speaker_comment	comment about speaker (accent, age, etc.)
recording_environment_comment	environment details (studio, music)
license	audio license

Relevant scripts are hosted on the accompanying GitHub: scrape_bible.py should be run first followed by force_align.py and build_corpus.py to generate the metadata spreadsheet along with the sliced audio files.

2 Corpus creation

The audio files are mostly sourced from recordings of the New Testament from bible.is. We also include miscellaneous clips found on the internet, as well as firsthand data collected on campus from a native speaker.

2.1 Data collection

2.1.1 bible.is

New Testament recordings were downloaded as raw audio from the bible.is website. Transcripts were scraped from the online reader using BeautifulSoup and written to an intermediate CSV with book and verse data (punctuation is preserved at this stage).

book_id	chapter	verse	verse_text
MAT	1	1	Ata lista di e antepasadonan...
MAT	1	2	Abraham tabata tata di Isak;...
MAT	1	3	Huda di Pèrès i Zèrag (nan mama...
MAT	1	4	Ram di Aminadab; Aminadab...

This data is then processed for sentence-level segmentation using the spaCy Spanish tokenizer (small). After this, punctuation is discarded to create text files for alignment.

2.1.2 Online content

There is not much transcribed audio for Papiamentu on the internet - there does not exist a comprehensive research dataset for the language, and there seem to be no recorded audiobooks besides the Bible.

We were able to find a poem recording as well as several videos on social media that did contain on-screen transcriptions, including language-learning videos and narrated shorts, of which one was manually converted to text. These data only total several minutes in length.

2.1.3 Firsthand recordings

To collect firsthand recordings, a sample transcript was generated by querying GPT4 for hundreds of paradigmatic Papiamentu phrases. This transcript was verified, edited, and read by Brandeis professor and Curaçao native Pito Salas, resulting in a single recording of around 65 utterances.

```
Bon dia, kon ta bai?
Mi kier sa e orario di vuelo.
Por fabor, sube e volumen.
Kiko e temperatura ta awe?
Set e alarm pa las siete.
Kambia e kanal por fabor.
```

Figure 1: Example of pre-edited GPT-4 Output

2.2 Processing

All audio was resampled to one channel with a rate of 16K Hz. All matching transcript and audio files were aligned via automatic requests to the Munich Automatic Segmentation System (MAUS) force aligner, then parsed into further sub-segments using both the natural verse distinctions and further sentence segmentation with spaCy.

3 Metadata

We append speaker identity information to the master spreadsheet. For the Bible.is data, there is a main designated speaker per book (intended to represent the book's author), with accompanying voices for other characters.

For our current corpus, we only differentiate between these principal speakers, but we are investigating automatic speaker diarization at the sentence level.

After labeling speakers, we partition audio samples into train/test/dev using an informed stratified sampling method. To ensure we are reasonably avoiding speaker bias across partitions, we leverage the Resemblyzer. Resemblyzer is a Python library that allows you to derive a high-level representations of a voice through a pre-trained voice-encoding model, an implementation of the paper Generalized End-To-End Loss for Speaker Verification (in which it is called the speaker encoder).

After embedding every utterance, we use universal manifold approximation and projection (UMAP) to visualize the utterances for each section of the Bible with a unique author-narrator. Knowing that there are other speakers mixed into these sections, we use this visualization to make informed partitions that are characterized by the presence unique masses of utterances within the resemblyzer vector space.

After partitioning the Bible data, we add our other data to the partitions with heuristics for how they might best supplement the potentially underrepresented aspects of the overall split.

Altogether, the dataset includes 11615 total utterances averaging 6.9 seconds with a maximum of 24.3 seconds. In total, there are 22.4 hours of utterances. These utterances, along with their speaker embeddings, are visualized in the figure at the end of the document.

4 Strengths

We have made efforts to compile a varied dataset beyond just Bible recordings, with several additional minutes of transcribed audio from the internet and firsthand recordings. This slightly expands the range of our vocabulary and improves speaker variety for model training.

By the heuristic of assuming bookwide Bible narrators, we can roughly categorize the different speakers in our dataset. Stratifying the data off of these

distinctions, our train/test/dev splits are more unbiased than if we had simply partitioned randomly.

Finally, our corpus creation pipeline is automatic, directly converting downloaded recordings and online text into consistent, sentence-level chunks. Our pipeline achieves perfect coverage of all of the available Bible.is audio chapters. Our method can likely be adapted for other languages covered by bible.is with some tweaks and a working force aligner.

5 Weaknesses

Despite our extra data collection, our corpus still has limited speaker diversity and covered subject matter. Our dataset is heavily skewed towards male speakers and contains no children speakers.

For all Bible data, the correct labeling of any given utterance is not guaranteed, since multiple speakers speak throughout the dramatic reading of any given book. In other words, Bible speakers that do not align with the author-narrator of the book are mislabelled.

Lastly, our data also only covers structured, written language, not noisy, everyday spoken language. It is unlikely to contain the common filler words or slang used in typical conversation.

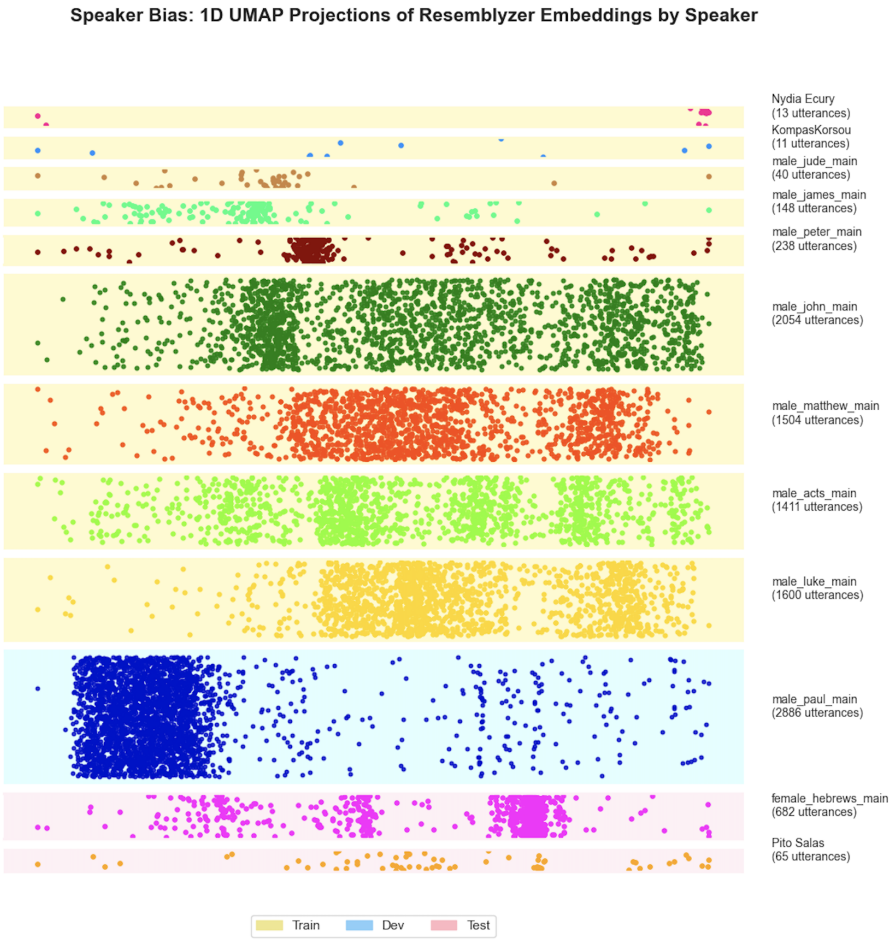

Figure 2: Speaker Bias: 1D UMAP Projections of Resemblyzer Embeddings