ASR Papiamentu Model Report

Sonny George, Chris Tam

December 16, 2023

1 Introduction

We present a Whisper-based transcription model for Papiamentu, a Portuguese creole spoken mainly in the Caribbean islands of Aruba, Bonaire, and Curaçao. We build off of the results of our previous work, a corpus of Papiamentu audio, transcript, and metadata. We evaluated and finetuned a Whisper model on this data, and experimented with the effect of asymmetric speaker partitioning to test the generalizability of our approach. Our efforts shed some light on the limitations of training ASR models on religious texts, and we discuss their resulting consequences for low-resource language technologies.

2 Data

Our existing corpus consists of around 11000 audio files (around 5-10 seconds each) from three sources: webscraped Bible data, firsthand elicited speech recordings, and internet media. The vast majority of our data is the Bible data, where Papiamentu recordings of the New Testament on bible.is were aligned and separated at the sentence level. This data consists of many different speakers portraying various figures throughout the each book. We also collected first-hand speech recordings of Papiamentu from Brandeis professor and Curaçao native Pito Salas, who verified and read from a script of sample utterances generated by GPT4. Finally, we found and added to our corpus a poem from an online archive and a promotional recording on social media, both with accurate transcriptions of Papiamentu.

3 Method

3.1 Fine-tuning Whisper

We transferred our corpus from its original spreadsheet format to a private HuggingFace Dataset by using the *pandas* library. We initially partitioned the data randomly into an 80/20 test/train split. As Papiamentu is heavily influenced by Spanish, we ran the default Spanish whisper-tiny model on this dataset as a baseline, which scored a word error rate (WER) of 103, which corresponds

to over one error per word. This is clearly not a useable model, and the error clearly stems from the vocabulary differences between the two languages.

To improve upon this, we used the training partition to finetune this whisper-tiny-es model (using the HuggingFace finetuning Trainer framework) with the following default hyperparameters:

- training_steps: 750
- learning_rate: 1e-05
- seed: 42
- optimizer: Adam with betas=(0.9,0.999) and epsilon=1e-08
- lr_scheduler_type: linear
- lr_scheduler_warmup_steps: 100
- mixed_precision_training: Native AMP

The model continued to improve until it reached a WER of 23, a significant step considering the size of the dataset. Further training saw decreasing training and validation loss, but increasing WER. We interpreted this as an indicated that the model was overfitting, and that the training set was perhaps too highly representative of the evaluation set. To remedy this bias, we attempted a form of speaker diarization for intentional partitioning.

3.2 Speaker-informed partitioning

3.2.1 Previous method (principal speaker) and limitations

In our previous project, we assigned single speakers to books with the assumption that the 'principal' speaker of each book, representing its author, would make up the majority of its utterances. This was an optimistic, but not ideal way to perform diarization. In fact, many biblical figures speak in each book of the Bible, each with their own voice actor, and some of these are not even 'principal' speakers. To capture this, we turn to the individual audio files themselves, and attempt to characterize them automatically.

3.2.2 Speaker clustering and partition assignments

To perform speaker diarization with no prebuilt tools for Papiamentu, we use a spectral clustering algorithm to group *resemblyzer* speaker embeddings into speaker categories. We evaluate the clustering algorithm with a hand-labelled subset of audio files and their deduced speaker, with the highest scoring algorithm making the fewest errors (bleed) across these annotations.

As shown in Figure 1, the best clustering algorithm split the data into a set of fourteen clusters of varied size and composition. These clusters are visually split into subsets, indicating which books their audiofiles come from. The hue of

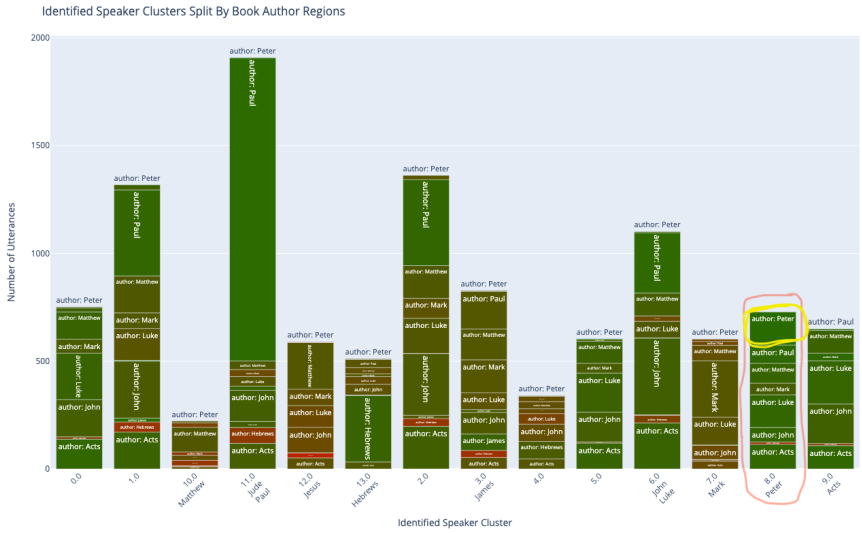

Figure 1: Bar chart of speaker clusters

these subsets reflects how close they are in vector space to the cluster's centroid - intuitively, more green means more aligned, and red means more noisy. These clusters were checked by manually listening to random segments of each, which for the most part corresponded to the same speaker.

The circled portion of the graph points out the 'Peter' cluster, which not only had the best representation of the book of Peter, but contained book subsets with high 'confidence'. As this cluster differentiates itself the greatest from the rest of the data, we choose this as our evaluation set to test the generalizability of our data. We also choose the set of elicited 'Pito' utterances as our test data, as we can classify them with 100% certainty, and to best represent non-Bible data.

The results of retraining our transcription model from scratch with this new paradigm are shown below.

model	split	wer	cer
0-shot (whisper-tiny spa)	test	98.77	57.65
0-shot (whisper-tiny spa)	dev	106.01	57.58
Fine-tuned whisper-tiny	test	60.99	26.62
Fine-tuned whisper-tiny	dev	30.29	14.90

We see that the fine-tuned model (trained with the same hyperparameters) achieves a decent WER of 30 on the dev set on the Peter cluster, whose voice it should theoretically not have heard. It improves, but performs worse on the elicited data, with a WER of 60. Character error rates (CER) follow a similar downward trend.

4 Discussion

Whisper, having been trained on hundreds of thousands of hours of data, is prone to overfitting on smaller datasets like ours involving low-resource languages. Though our first experiment showed that the model could classify unseen utterances well, provided they were from the Bible, an unbiased train/test split using randomized or stratified data does not address the problem of bias, and our initial evaluations were sure to overestimate the model's capabilities.

Indeed, in ensuing experiments, when evaluated purely on an unseen cluster, the model performed slightly worse, and performed considerably worse on unseen contemporary data. This larger errors could be attributed to several reasons, such as the difference in vocabulary used in a more modern setting and the difference the in recording environment. These results stress the need for more varied, non-religious set of audio and transcription data for low-resource languages, where even a handful of fully transcribed contemporary videos could make a world of difference for ASR technology.