TCNSpeech: A Community-Curated Speech Corpus for Sermons

Wuraola Fisayo Oyewusi, Sharon Ibejih, Soromfe Uzomah, Elizabeth Joseph,
Cynthia Olawuyi, Folakunmi Ojemuyiwa, Benedicta Johnson-Onuigwe, Omolola Taiwo,
Akintunde Akinpelumi, Olabisi Adesina, Ayodele Noutouglo, Adeola Adeoba, Andrew Akoh,
Chukwuemeka Nwachukwu, Opeyemi Agbabiaje, Itunu Falade, Olukemi Erhunmwunsee, Oluwatobiloba Dada,
Oluwatobi Osibeluwo, Ehis Akene, Udim Akpan, Moira Amadi-Emina, Jaiyeola Marquis,
Michael Senapon Bojerenu, Gbolahan Olumade, Oluwagbemi Lesi, Timothy Ezeh, Oluwadamilola Oguntoyinbo,
Tosan Mogbeyiteren, Felicia Oresanya, Samuel Chika & Sodiq Akinjobi
Digital Tech Community Group, The Covenant Nation(TCN) Lagos,Nigeria
ccgdigitaltech@gmail.com

Abstract

In this work we present TCNSpeech, a community-curated multispeaker sermon corpus for speech recognition tasks. It contains a total of 24 hours of English audio data recording, chunked and transcribed. The context of the dataset is domain-specific for sermons in Nigerian English accent and a use case for community data curation. The dataset is made publicly available.

1 Background

Despite the progress being made in the adoption and design of Automatic Speech Recognition (ASR) systems, there is still a range of barriers that influence solution development and user satisfaction. When training resources are scarce, variations such as speaker gender, speaking rate, regional accent, speaking style are challenging to model Benzeghiba et al. (2007).

This work introduces TCNSpeech, a context-specific multispeaker corpus of Church sermons curated by volunteers in a digital technology community group. The goal is to leverage the community to transcribe a dataset that captures the nuances of sermon speaking style and content type in a Nigerian accent. The content types referred to are biblical terms such as names, locations, bible chapters, and church experiences such as admonishing, praying, speaking in tongues, playing music, etc. There are different tribes with marked influence on intonation, the inflection on how they speak English. So for clarity, the reference to Nigerian accent is limited to the speakers in the sermons transcribed who are mostly from Lagos, the southwestern part of Nigeria.

While this dataset is curated as part of a larger project for bespoke Speech to Text to power real-time automatic transcription for sermons and songs by the host Church, it is going to be openly available as a contribution to science.

1.1 Community Focused Data Curation

One of the effective ways for data curation is to leverage a community that understands the data and is passionate about the use case. There are several use cases of communities collaborating to gather and curate datasets that spotlight their domain. For example in Africa, the Maskahane community [1] translated as “We build together” has collaborated to build a range of datasets for different natural language processing tasks such as Named Entity Recognition Adelani et al. (2021), and Machine Translation Nekoto et al. (2020). SautiDB [2] is also leveraging the community to gather SautiDB-Naija corpus, a novel corpus of non-native (L2) Nigerian English speech Afonja et al. (2021).

[1] https://www.masakhane.io/

[2] https://sautidb.web.app/home

1.2 SERMONS AND RELATED CONTEXT

Sermon is interpreted as a cautionary speech of religious content, spiritual shepherd appeal to believers in Church. We consider a church sermon not only in the atmosphere of the temple but in the biblical context qualifying it as any spiritual instruction of the priest to the laity Morozov (2015). There are a number of openly available datasets for Automatic Speech Recognition (ASR) related to church content such as MaSS: A Large and Clean Multilingual Corpus of Sentence-aligned and Spoken Utterances Extracted from the Bible Boito et al. (2019), Speech to Text System: Pastor Wang Mandarin Bible Teachings Kao. As of the time of this publication, we are not aware of other openly available corpus related to sermons in the Nigerian context.

2 METHODOLOGY

As previously mentioned, this project leveraged both community and open source tools. For reproducibility, we will share as much as possible about our process:

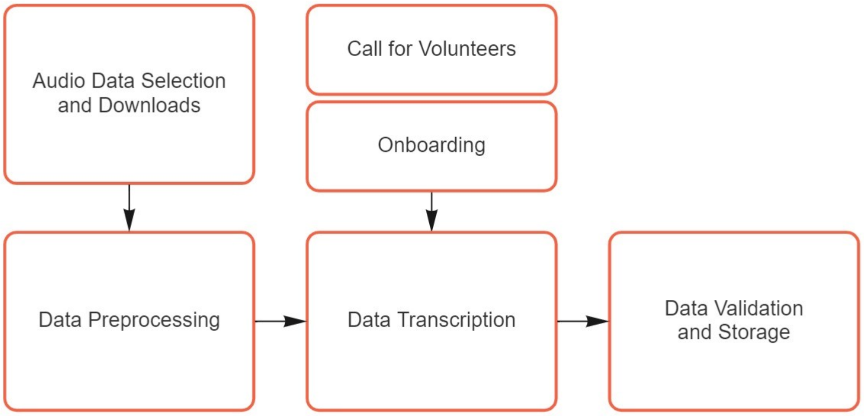


Figure 1: Methodology

Concept Note and Annotation Guide
Setting an appropriate big picture and creating a detailed annotation guide were pertinent since data transcription was done by volunteers without experience. A detailed concept note with information on the entire project and milestones was created, then an annotation guide for the audio data transcription task.

The annotation guide[3] contains instruction on how to open assigned folders, name completed transcription, effective listening that captures all that is being said including laughter, ehm, stammers, etc., how to write (1. Uniform spelling of pronunciations: Spelling names in full like verse instead of vs, Corinthians instead of Cor. or Pastor instead of Pst., 2. Writing all numbers, including bible chapters and verses as figures and not in words such as James 2 verses 1) and how to transcribe specific instances like music interludes, speaking in tongues with place holders such as [music] for songs or instrumentals, [unknown] for spoken words that are unclear, [clap] for claps, [speaking in tongues] for speaking in tongues, [prayer] for prayer times.

Tools Selection
The project explored the use of only readily available tools such as emails to assign tasks, Google Drive for data storage, Google Form for registration and Google Sheets for tracking.

Call for Volunteers
The call for volunteers was made within the closed digital technology community group and there

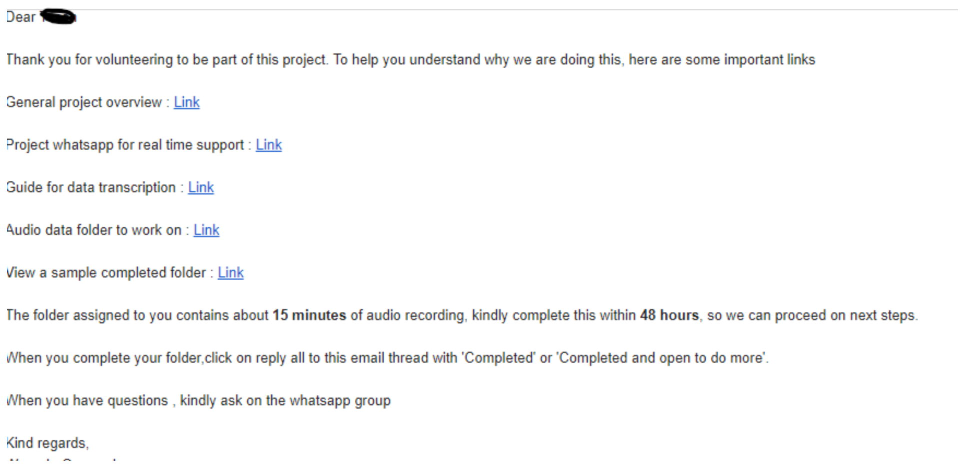
[3] https://bit.ly/audio_data_transcription_guide

were up to 71 volunteers who showed interest. The application form was simple and asked for information such as a person's name, email address, and WhatsApp phone number.

Audio Data Selection and Download
Sermons from both male and female speakers with few congregation interferences were selected across. Typical Church-related experiences like praying, singing, clapping, and speaking in tongues were included in the dataset. This selected data contains a mix of both male and female speakers downloaded in *'.wav'* format.

Data Preprocessing
The downloaded data with sizes ranging from 50 minutes to 5 hours were chucked into 10-second wav files at a sampling rate of 16000 Hz using the Audacity software [4]. The chunked files were then grouped into 90 files in folders of 10 to 15 Minutes each. On random sampling, it folders of 15 minutes were found optimum by the transcribers.

Data Transcription
Data transcribers received an onboarding email with specific instructions on what to do at each step. The image in Figure 2 shows an example of the onboarding email with a link to a WhatsApp group for real-time support by both coordinators and other transcribers



Dear [redacted]

Thank you for volunteering to be part of this project. To help you understand why we are doing this, here are some important links

General project overview : Link

Project whatsapp for real time support : Link

Guide for data transcription : Link

Audio data folder to work on : Link

View a sample completed folder : Link

The folder assigned to you contains about **15 minutes** of audio recording, kindly complete this within **48 hours**, so we can proceed on next steps.

When you complete your folder,click on reply all to this email thread with 'Completed' or 'Completed and open to do more'.

When you have questions , kindly ask on the whatsapp group

Kind regards,

Figure 2: Sample of onboaring email

Data Validation and Storage
Transcriptions were double-checked to ensure alignment with provided guidelines. Each transcription lines were also checked to ensure that their audio file names were correctly appended. While we made possible efforts to avoid transcription errors, this data may still contain some.

3 TCNSpeech

The total duration of the openly available TCNSpeech is 24 hours. The data multi-speaker and grouped by gender, so there are different folders for male and female voices, this split is to make provision for gender based use cases.Table 1 shows a break down of the gender distribution in the data. Each folder contains audio data chunks and their transcripts.For ease of matching, each line in the transcript is named like the audio file but excludes the ".wav" suffix.

3.1 ASR Experimentation and Result

As proof of concept, two speech to text models were trained using NVIDIA NeMo's QuartzNet 15x5 ASR architecture [5] at 4.43 and 13.23 hours of the TCNSpeech data. The training data was

[4] https://www.audacityteam.org/

[5] https://docs.nvidia.com/deeplearning/nemo/user-guide/docs/en/main/asr/configs.htmljasper-and-quartznet

Table 1: TCNSpeech Corpus

Gender	**Number of Samples**	**Speech Duration**
Female	3600	11 hours
Male	4950	13 hours
Total	8550	24 hours

augmented with about 5.77 hours of the Nigerian English [en-ng] multi-speaker speech dataset Research (2018/2019) which captures Nigerian English accents. The vocab used for the training contained all characters and symbols represented in the transcription. These were:

[’A’,’B’, ’C’, ’D’, ’E’, ’F’, ’G’, ’H’, ’I’, ’J’, ’K’, ’L’, ’M’, ’N’, ’O’,’P’, ’Q’, ’R’, ’S’, ’T’, ’U’, ’V’, ’W’, ’X’, ’Y’, ’Z’, ’a’, ’b’, ’c’, ’d’, ’e’, ’f’, ’g’, ’h’, ’i’, ’j’, ’k’, ’l’, ’m’, ’n’, ’o’, ’p’, ’q’, ’r’, ’s’, ’t’, ’u’, ’v’, ’w’, ’x’, ’y’, ’z’, ’0’, ’1’, ’2’, ’3’, ’4’, ’5’, ’6’, ’7’, ’8’, ’9’, "’",’-’, ’,’,’!’, ’?’, ’[’, ’]’, ’:’, ’ ’]

The models were trained at 100 and 75 epochs respectively, with a batch size of 8, and learning rate of 0.001. Table 2 shows the data size, data split and the Word Error Rate of each model

Table 2: Results of experimented models

Experiment	**Sermon data + en-ng data durations**	**Total data duration**	**Train**	**Validation**	**Validation WER (%)**
First	4.43 + 5.77	10.20	9.95	0.25	0.35
Second	13.48 + 5.77	19.25	19.00	0.25	0.31

Table 3 shows some examples of transcription by the trained models versus the ground truth transcription. Improvement in performance with the model trained on longer data is expected.

Table 3: Transcription Results of experimented models

Ground Truth	**First model prediction**	**Second model prediction**
[music] Praise the Lord Can we rise on our feet and just have a song It’s a bit hot isn’t it	[music] praise the lord *canw* rise on our fet and just have a song its a *bitd hout isnt* it	[music]Praise the Lord can we rise on our feet and just have a song *i* a bit hot isn’t it
[music] Matthew 14 lets start reading from verse 24	*l]ayusein ets not raning* from verse *2weour [msir]*	*[usic]* Matthew 14 lets start reading from verse 24
He used spit for this one he didn’t go near them they were five, and he just shouted you know ah Jesus Jesus go and show yourself to the priest	he *ue speed* for this one he *didnt bo* near them they were *falve* and he just shouted you know jesus jesus *who and shore yoursef* to the *prist*	He used *speaet* for this one he didn’t go near them they were *fave*, and he just *shoutted* you know ah Jesus Jesus go and *showe* yourself to the *prierst*
We thank you we are women who worship we understand that our worship is our warfare	we thank you *[lp]* we are women who worship we understand that our *wordship* is our warfare	We thank you we are women who worship we understand that our worship is our warfare
Water so he made his request known unto Jesus and Jesus said come and when	water so e ma his *requiet nowt* unto jesus and jesus said come	Water so he made his request known unto Jesus and Jesus said come *andd* when

4 Conclusion

We presented TCNSpeech, an open community curated speech corpus of sermons. It's a contribution to address domain and accent specific data availability for Automatic Speech Recognition tasks by leveraging a passionate volunteer community. On experimentation for a speech-to-text task, the best-performing model achieved a word error rate of 0.31.

References

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, et al. Masakhaner: Named entity recognition for african languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131, 2021.

Tejumade Afonja, Oladimeji Mudele, Iroro Orife, Kenechi Dukor, Lawrence Francis, Duru Goodness, Oluwafemi Azeez, Ademola Malomo, and Clinton Mbataku. Learning nigerian accent embeddings from speech: preliminary results based on sautidb-naija corpus. *arXiv preprint arXiv:2112.06199*, 2021.

Mohamed Benzeghiba, Renato De Mori, Olivier Deroo, Stephane Dupont, Teodora Erbes, Denis Jouvet, Luciano Fissore, Pietro Laface, Alfred Mertins, Christophe Ris, et al. Automatic speech recognition and speech variability: A review. *Speech communication*, 49(10-11):763–786, 2007.

Marcely Zanon Boito, William N Havard, Mahault Garnerin, Éric Le Ferrand, and Laurent Besacier. Mass: A large and clean multilingual corpus of sentence-aligned spoken utterances extracted from the bible. *arXiv preprint arXiv:1907.12895*, 2019.

Yen Peng Karl Kao. Speech to text system: Pastor wang mandarin bible teachings (speech recognition).

Evgeniy M Morozov. Current communication trends in church sermons. *Procedia-Social and Behavioral Sciences*, 200:496–501, 2015.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Tajudeen Kolawole, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Hassan Muhammad, Salomon Kabongo, Salomey Osei, et al. Participatory research for low-resourced machine translation: A case study in african languages. *arXiv preprint arXiv:2010.02353*, 2020.

Google Research. Crowdsourced high-quality nigerian english [en-ng] multi-speaker speech dataset, 2018/2019. URL `https://research.google/tools/datasets/nigerian-english-tts/`.