MULTILINGUAL AUTOMATIC SPEECH RECOGNITION FOR KINYARWANDA, SWAHILI, AND LUGANDA: ADVANCING ASR IN SELECT EAST AFRICAN LANGUAGES

Yonas Chanie
Carnegie Mellon University
ychanie@andrew.cmu.edu

Moayad Elamin
Carnegie Mellon University
melamin@andrew.cmu.edu

Paul Ewuzie
Carnegie Mellon University
pewuzie@andrew.cmu.edu

Samuel Rutunda
Digital Umuganda
samuel@digitalumuganda.com

ABSTRACT

This paper presents a multilingual Automatic Speech Recognition (ASR) model for three East African languages—Kinyarwanda, Swahili, and Luganda. The Common Voice project's African languages datasets were used to produce a curated code-switched dataset of 3,900 hours on which the ASR model was trained. The work included validating the Kinyarwanda dataset and developing a model that achieves a 17.57 Word Error Rate (WER) on the language. Across all three languages, the Kinyarwanda model was finetuned and achieved a WER of 21.91 on the three curated datasets, with a WER of 25.48 for Kinyarwanda, 17.22 for Swahili, and 21.95 for Luganda. The paper emphasizes the necessity of considering the African environment when developing effective ASR systems and the significance of supporting many languages when developing ASR for languages with limited resources.

1 INTRODUCTION

Significant improvements have occurred in speech technologies due to the advancement in the architectural design of deep learning models and the availability of large speech corpora. Specifically, the current ASR models have shown impressive results in well-renowned languages for speech-to-text tasks. Despite this, most low-resource languages do not have accurate speech recognizers. Many of these languages have more than one million native speakers, some of whom have low literacy levels and can benefit from speech technologies to access information.

To build a robust Automatic Speech Recognition (ASR) in Africa, one must take into account the African context. Africa is linguistically rich with over 2000 languages, and given its history, there are many cases of people speaking multiple languages. A person might speak a local language, a regional language, and an administrative language which in most cases is a European language (English, Fresh, Spanish, or Portuguese). People living in border areas might speak the language of the neighboring region/country. This might cause a person to either switch languages within one conversation, speak one language using his mother tongue, or to code-switch, thus requiring an ASR capable of supporting multiple languages.

In this work, we created a multilingual ASR model for 3 of the most spoken east African language, Kinyarwanda, Swahili, and Luganda. Using the common voice dataset, we curated and created the dataset to accomplish the multi-language task. After training the model we achieved a Word Error Rate (WER) of 21.91 in all 3 languages, with the WER of 25.48, 17.22, and 21.95 for each language respectively.

2 Background

2.1 Automatic Speech Recognition

Automatic speech recognition (ASR) is a technology that enables computers to understand and transcribe human speech (Gaikwad et al., 2010). This technology has a range of applications, such as voice-controlled devices, transcription services, and accessibility tools for individuals with speech impairments. The use of ASR technology has increased in recent years due to advances in machine learning and the availability of large amounts of data for training.

ASR technology has its roots in signal processing, where engineers have long sought to develop algorithms for extracting information from speech signals. It is well-known that the speech signal not only conveys linguistic information (the message) but also a lot of information about the speaker himself: gender (Rakesh et al., 2011), age (Drager, 2011), social, and regional origin (Walker, 2007), health (Rosen et al., 2006), and emotional state (Polzin & Waibel, 1998) and, with relatively strong reliability, his identity (Benzeghiba et al., 2007). With the advent of deep neural networks, the performance of ASR models has improved significantly, using all attributes of speech to produce a matching text. The deep neural network replaced the traditional Gaussian mixture for the acoustic likelihood evaluation while keeping all other components of the ASR hybrid model (Hinton et al., 2012). Recently, there have been new breakthroughs in ASR modeling with speech communities transiting from hybrid modeling to end-to-end modeling, using a single network for the whole speech-to-text process (Bahdanau et al., 2016).

In the early years of ASR modeling, building multilingual ASR systems was difficult since we needed acoustic models with shared hidden layers (Ghoshal et al., 2013) (Heigold et al., 2013) while ensuring each language has its lexicon and language model. The end-to-end models made the process easier as it takes the union of all languages and tokenizes them which is further used for the training of the ASR model (Cho et al., 2018) (Kim & Seltzer, 2018). However, multilingual ASR models are always prone to mixing languages during recognition which is why some models defer to conditioning on Language Identity (LID) as this guides the ASR model to generate the transcription for the target language by reducing confusion from other languages (Kannan et al., 2019). The gain of using LID is limited in end-to-end models, especially when streaming audio as it is not very reliable.

Recent advances have presented new approaches to achieving remarkable success in performing ASR tasks by using transformer models (Mohamed et al., 2019) (Radford et al., 2022). Chan et al. (2016)'s Attention is used by the transformer model to identify important parts of the input sequence or speech. This technique has allowed for the creation of more complex models that are more capable of performing language and speech tasks. Pairing this transformer model with semi-supervised learning (SSL) techniques has also spurred the design of the Wav2Vec model (Schneider et al., 2019) which converts audio signals into representations that then can be used in downstream tasks. Wav2Vec is trained on large amounts of unlabeled audio data which makes its training methodology important to languages that have less labeled data. Recently, conformer models (Gulati et al., 2020b); combine convolutional neural network (CNN) components and transformers to capture local and global dependencies in an audio sequence more efficiently than the stand-alone variants of the individual models.

2.2 ASR on African Languages

The space of resource-constrained languages has seen a good amount of effort to create models that can learn and perform well using small amounts of data, a few hours of annotated data in this case. The ALFFA project is focused on distributing ready-to-use or ready-to-train Kaldi ASR systems and associated corpora for sub-Saharan African languages. The ASR directory currently includes Kaldi recipes for Amharic (Tachbelie et al., 2014), Swahili(Gelas et al., 2012), Hausa, and Wolof (Gauthier et al., 2016). The issue of data availability has pushed many researchers to either collect their data or use alternative data sources. Yılmaz et al. (2018) explore the creation and use of a Soap Opera speech corpus to create ASR models for code-switching between five South African languages while Doumbouya et al. (2021) used radio archives to create the West African Radio Corpus and the West African Virtual Assistant Speech Recognition Corpus then used the first to train the West African wav2vec. Mohamud et al. (2021) conducted a mobile application-based data collection project in which they tested the performance of self-supervised learning techniques on limited hours of three languages, Wolof, Ga, and Somali. Ritchie et al. (2022) experienced with multilingual modeling as well as semi-supervised learning for 15 African languages and found that a combination of SSL and finetuning with the small available data will create robust models across the tested languages.

2.3 Languages

There are three major language families that represent the majority of languages spoken all over Africa, Afroasiatic (Arabic, Hausa, Somali, Amharic ...etc.), Nilo-Saharan (Kanuri, Fur ... etc.), and Niger-Congo (Igbo, Yoruba, and the Bantu family) as seen in figure 1.

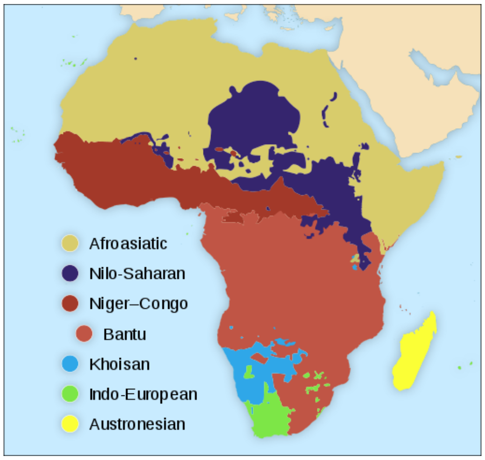

Figure 1: Map of African language families (Commons, 2017)

The Bantu family contains the largest number of languages spoken in Africa, covering the largest speech area of all other families. Swahili, Kinyarwanda, and Luganda are amongst the most spoken language in East Africa, all are in the Bantu family of languages and uses the Latin alphabet. Swahili is the largest of them, spoken by over 70 million people throughout East and Central Africa, Kinyarwanda follows spoken by over 13 million people in Rwanda, Uganda, and the Democratic Republic of Congo (DRC) and Luganda has over 6 million speakers in Uganda.

The ease of travel across the East African Community allows these three languages to be spoken across the entire federation, adding a level of intermingling when spoken by people. This loose use of the languages makes it hard to create and deploy a single language model in this region, as speakers tend to move naturally across these languages, making the creation of multilingual systems a must.

3 Methodology

3.1 Data

The data we used was obtained and created using the Mozilla CommonVoice dataset. We used the monolingual Common Voice datasets for Kinyarwanda, Luganda, and Swahili as baselines and we combined the datasets to create the final multilingual dataset. The information in hours can be found in table 1 below.

For Kinyarwanda, we found many issues in the recordings and the transcriptions by manually inspecting the data and analyzing the high error values found in a baseline ASR model. Common voice adds a validation step, allowing community members to validate whether a certain recording and its transcriptions match. We found that this validation step does not affect the data and that the data used for training, and testing the previous models had issues. To add a validation step, to remove some of the bad transcriptions and recordings, we checked the number of words in the transcription

Table 1: Monolingual datasets (hours)

Split	Kinyarwanda	Swahili	Luganda
Train	1,284.8	48.59	102.1
Validate	24.3	16.7	20.1
Test	21.96	16.86	19.78

per second of the recording (Word Rate), to reject some of the noisy samples that were found using manual inspection.

Using this method, words with over three words per second were rejected, as we found that the average word rate of the dataset was 1.75 words per second and the human average speech rate is 160.6 words per minute (Pindzola et al., 1989), which is 2.7 words per second. This has resulted in a smaller set of samples (see table 1) which when used to train a monolingual Kinyarwanda ASR model[1] achieved a WER of 17.57 which represents the SOTA model for the language. Adding a step to normalize the apostrophes, commas, and full stops across the model predictions and actual transcriptions, we were able to achieve a WER of 5.09.

Inspecting the data for this model, we found that some of the test samples are in Luganda, which the model was successful in predicting. This was a motivation to explore the similarities between Kinyarwanda and Luganda which led to the selection of the three languages we are studying in this paper.

To create the multilingual dataset, we combined the three datasets. Each sample in the final dataset had a random combination of random samples from the three datasets. This means that a sample can have either multiple samples of a single language or randomized samples from the languages. The data was created with a sampling rate of 16KHz, at a length of 10 to 20 seconds with 100ms pauses between the random samples. The resulting dataset information can be found in table 2.

Table 2: Multilingual datasets (hours)

Split	Full dataset	Size (samples)
Train	3,013.7	796,971
Validate	601.23	158,042
Test	300.64	79,053

3.2 Model

For this work, we fine-tuned a pre-trained Conformer (Gulati et al., 2020a) based Kinyarwanda Model [2] with CTC decoding (Graves et al., 2006) and BPE Tokenization (Wang et al., 2020). Conformer uses a combination of convolution layers and self-attention to achieve better performance, self-attention learning global features while convolution captures local relationships. Contextual temporal classification (CTC) decoding was used to better capture the temporal information or the alignment of the data without the need for segmenting input data. Byte pair encoding (BPE) tokenization is a type of tokenization technique that breaks down the text into smaller subword pieces learned from text (Shibata et al., 1999). BPE tokenization is useful in cases where out-of-vocabulary (OOV) exists and this technique has been used in a range of various NLP tasks, including speech processing, neural machine translation, and language modeling (Shibata et al., 1999) (Kunchukuttan & Bhattacharyya, 2016) (Choudhary et al., 2018) (Zhou et al., 2021).

[1] Available on Huggingface under the MbazaNLP community space: `https://huggingface.co/mbazaNLP/`

[2] Kinyarwanda Conformer Model: `https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/stt_rw_conformer_ctc_large`

3.3 Evaluation

To evaluate the performance of the models trained in this paper, we used the Word Error Rate (WER) as a metric. The word error rate can be defined as the ratio of word insertions (I), substitutions (S), and deletions (D) needed for the transcription to match the prediction to the total number of spoken words (N) (Park et al., 2008).

$$WER = \frac{S_{word} + D_{word} + I_{word}}{N_{word}} \tag{1}$$

We also evaluate using the Character Error Rate (CER) which refers to the ratio of character insertions (I), substitutions (S), and deletions (D) needed for the transcription to match the prediction to the total number of characters in the transcription (N)

$$CER = \frac{S_{character} + D_{character} + I_{character}}{N_{character}} \tag{2}$$

For both of these values, we aim to reduce these ratios when evaluating our models.

4 Experiments & Results

Using the pre-trained Kinyarwanda conformer-based model, we finetuned the model on the proposed dataset to evaluate the performance on the test split of the dataset and baseline dataset splits. The Conformer model has 121 million parameters and is trained using NeMo Toolkit (Kuchaiev et al., 2019) setting the hyperparameters as shown in the table 3. The tokens were created from the training set of the dataset using the sentence-piece (Kudo & Richardson, 2018) tokenization technique resulting in 128 subtokens.

Table 3: Training Hyperparameters

Hyperparameter	Value
Epoch	120
Batch Size	24
Tokenizer	BPE
Optimizer	AdamW
Learning Rate	2.0
Scheduler	NoamAnnealing
Weight Decay	0
Mixed Precision	16fp
SpecAugment (Time Mask)	10sec
SpecAugment (Freq Width)	27Hz

The results evaluated on the test splits of the dataset are presented in table 4. Our result shows that while we were not able to improve the WER for Kinyarwanda compared to the monolingual baseline model, we can note that the CER is lower and our multilingual model is able to correctly predict at the character level. Generally, the higher WER as compared to lower CER could be attributed to the fact that the model has good performance at predicting. Missing a character greatly affects the WER performance since the word has to be replaced.

In addition, inspecting audio samples that have higher WER, we have noticed that most of the prediction errors happened due to noise in the background, unfinished audio samples, differences in pronunciation, and our model either missing or adding an additional repeating letter.

Table 4: Word and Character error rate on the test splits

Test Set	WER	CER
Code-Switched	21.91	6.38
Kinyarwanda	25.48	7.79
Swahili	17.22	5.96
Luganda	21.95	5.15

5 Conclusion

5.1 Summary & Discussion

In this work, we explored creating multilingual models for ASR across three East African languages, Kinyarwanda, Swahili, and Luganda. We first explored validating and cleaning the Kinyarwanda dataset available through the Common voice project, creating a model that achieves 17.57 WER on Kinyarwanda. We then used this language and the monolingual datasets for the three languages to create 3,900 hours of multilingual data that was then used to train a model that achieves a WER of 21.91 on the created dataset and 25.48 on Kinyarwanda, 17.22 on Swahili and 21.95 on Luganda. We also evaluated the test set using CER and we were able to get 6.38 on the new dataset, 7.79 on Kinyarwanda, 5.96 on Swahili, and 5.15 on Luganda. The results show that the model has better performance at predicting character level.

5.2 Future Work

As we have seen with Kinyarwanda, improving the datasets will improve the model accuracy. These improvements to the current datasets can include removing defunct recordings, perfecting the punctuation, and using a tokenizer for each separate language. Connecting with native speakers of Luganda and Swahili will allow us to reproduce the error analysis and investigate the dataset more, which requires collaboration with the East African research community. The models trained are tested on the curated dataset, which is a limitation as we need to test the model in more natural multilingual and codeswitching situations. This can be explored by collecting speech data from actual speakers that represent real human speech. We also plan to include other African languages, including exploring the similarities between non-Bantu languages, for example, the connections between Amharic and Arabic, and other similar situations in the region.

We also plan to extend the work, after allocating more resources, to include bigger models like fine-tuning the Whisper model (Radford et al.), neural scoring as well as exploring semi-supervised training on the large amounts of unlabelled data that is available through the radio, YouTube and other open platforms.

Acknowledgments

We would like to thank the FAIR Forward - Artificial Intelligence for All[3] for their support in providing computing power and being the reason behind this research through their capstone project proposal. We also thank German Research Centre for Artificial Intelligence (DFKI) [4] for their technical support working on the GPUs. We thank Vitaly Lavrukhin, Aleksandra Antonova, and Boris Ginsburg from NVIDIA for their support and advice in training the NEMO conformer model. We also thank the Mbaza-NLP community[5] for their contribution in testing the model in the various languages.

[3] FAIR Forward: `https://www.giz.de/expertise/html/61982.html`

[4] DFKI: `https://www.dfki.de/en/web`

[5] Mbaza NLP: `https://mbaza.org/`

REFERENCES

Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel, and Yoshua Bengio. End-to-end attention-based large vocabulary speech recognition. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 4945–4949. IEEE, 2016.

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.

William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 4960–4964. IEEE, 2016.

Jaejin Cho, Murali Karthick Baskar, Ruizhi Li, Matthew Wiesner, Sri Harish Mallidi, Nelson Yalta, Martin Karafiat, Shinji Watanabe, and Takaaki Hori. Multilingual sequence-to-sequence speech recognition: architecture, transfer learning, and language modeling. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pp. 521–527. IEEE, 2018.

Himanshu Choudhary, Aditya Kumar Pathak, Rajiv Ratan Saha, and Ponnurangam Kumaraguru. Neural machine translation for english-tamil. In *Proceedings of the third conference on machine translation: shared task papers*, pp. 770–775, 2018.

Wikimedia Commons. Map of african language families, 2017. URL `https://commons.wikimedia.org/wiki/File:Map_of_African_language_families.svg`.

Moussa Doumbouya, Lisa Einstein, and Chris Piech. Using radio archives for low-resource speech recognition: towards an intelligent virtual assistant for illiterate users. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 14757–14765, 2021.

Katie Drager. Speaker age and vowel perception. *Language and Speech*, 54(1):99–121, 2011.

Santosh K Gaikwad, Bharti W Gawali, and Pravin Yannawar. A review on speech recognition technique. *International Journal of Computer Applications*, 10(3):16–24, 2010.

Elodie Gauthier, Laurent Besacier, Sylvie Voisin, Michael Melese, and Uriel Pascal Elingui. Collecting resources in sub-saharan african languages for automatic speech recognition: a case study of wolof. *LREC*, 2016.

Hadrien Gelas, Laurent Besacier, and Francois Pellegrino. Developments of Swahili resources for an automatic speech recognition system. In *SLTU - Workshop on Spoken Language Technologies for Under-Resourced Languages*, Cape-Town, Afrique Du Sud, 2012. URL `http://hal.inria.fr/hal-00954048`.

Arnab Ghoshal, Pawel Swietojanski, and Steve Renals. Multilingual training of deep neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pp. 7319–7323. IEEE, 2013.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pp. 369–376, 2006.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. Conformer: Convolution-augmented Transformer for Speech Recognition. May 2020a. doi: 10.48550/arXiv.2005.08100. URL `http://arxiv.org/abs/2005.08100`. arXiv:2005.08100 [cs, eess].

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*, 2020b.

Georg Heigold, Vincent Vanhoucke, Alan Senior, Patrick Nguyen, Marc'Aurelio Ranzato, Matthieu Devin, and Jeffrey Dean. Multilingual acoustic models using distributed deep neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pp. 8619–8623. IEEE, 2013.

Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97, 2012.

Anjuli Kannan, Arindrima Datta, Tara N Sainath, Eugene Weinstein, Bhuvana Ramabhadran, Yonghui Wu, Ankur Bapna, Zhifeng Chen, and Seungji Lee. Large-scale multilingual speech recognition with a streaming end-to-end model. *arXiv preprint arXiv:1909.05330*, 2019.

Suyoun Kim and Michael L Seltzer. Towards language-universal end-to-end speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4914–4918. IEEE, 2018.

Oleksii Kuchaiev, Jason Li, Huyen Nguyen, Oleksii Hrinchuk, Ryan Leary, Boris Ginsburg, Samuel Kriman, Stanislav Beliaev, Vitaly Lavrukhin, Jack Cook, et al. Nemo: a toolkit for building ai applications using neural modules. *arXiv preprint arXiv:1909.09577*, 2019.

Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*, 2018.

Anoop Kunchukuttan and Pushpak Bhattacharyya. Learning variable length units for smt between related languages via byte pair encoding. *arXiv preprint arXiv:1610.06510*, 2016.

Abdelrahman Mohamed, Dmytro Okhonko, and Luke Zettlemoyer. Transformers with convolutional context for asr. *arXiv preprint arXiv:1904.11660*, 2019.

Jama Hussein Mohamud, Lloyd Acquaye Thompson, Aissatou Ndoye, and Laurent Besacier. Fast development of asr in african languages using self supervised speech representation learning. *arXiv preprint arXiv:2103.08993*, 2021.

Youngja Park, Siddharth Patwardhan, Karthik Visweswariah, and Stephen C Gates. An empirical analysis of word error rate and keyword error rate. In *INTERSPEECH*, volume 2008, pp. 2070–2073, 2008.

Rebekah H Pindzola, Melissa M Jenkins, and Kari J Lokken. Speaking rates of young children. *Language, Speech, and Hearing Services in Schools*, 20(2):133–138, 1989.

Thomas S Polzin and Alex Waibel. Detecting emotions in speech. In *Proceedings of the CMC*, volume 16, 1998.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust Speech Recognition via Large-Scale Weak Supervision. pp. 28.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*, 2022.

Kumar Rakesh, Subhangi Dutta, and Kumara Shama. Gender recognition using speech processing techniques in labview. *International Journal of Advances in Engineering & Technology*, 1(2):51, 2011.

Sandy Ritchie, You-Chi Cheng, Mingqing Chen, Rajiv Mathews, Daan van Esch, Bo Li, and Khe Chai Sim. Large vocabulary speech recognition for languages of africa: multilingual modeling and self-supervised learning. *arXiv preprint arXiv:2208.03067*, 2022.

Kristin M Rosen, Raymond D Kent, Amy L Delaney, and Joseph R Duffy. Parametric quantitative acoustic analysis of conversation produced by speakers with dysarthria and healthy speakers. 2006.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*, 2019.

Yusuxke Shibata, Takuya Kida, Shuichi Fukamachi, Masayuki Takeda, Ayumi Shinohara, Takeshi Shinohara, and Setsuo Arikawa. Byte pair encoding: A text compression scheme that accelerates pattern matching. 1999.

Martha Tachbelie, Solomon Teferra Abate, and Laurent Besacier. Using different acoustic, lexical and language modeling units for asr of an under-resourced language - amharic. *Speech Communication*, 56, 2014.

Abby Walker. The effect of phonetic detail on perceived speaker age and social class. In *Proceedings of the 16th International Congress of Phonetic Sciences, Saarbrücken*, pp. 1453–1456. Citeseer, 2007.

Changhan Wang, Kyunghyun Cho, and Jiatao Gu. Neural machine translation with byte-level subwords. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 9154–9160, 2020.

Emre Yılmaz, Astik Biswas, Ewald van der Westhuizen, Febe de Wet, and Thomas Niesler. Building a unified code-switching asr system for south african languages. *arXiv preprint arXiv:1807.10949*, 2018.

Wei Zhou, Mohammad Zeineldeen, Zuoyun Zheng, Ralf Schlüter, and Hermann Ney. Acoustic data-driven subword modeling for end-to-end speech recognition. *arXiv preprint arXiv:2104.09106*, 2021.