Large vocabulary speech recognition for languages of Africa: multilingual modeling and self-supervised learning

Sandy Ritchie, You-Chi Cheng, Mingqing Chen, Rajiv Mathews, Daan van Esch, Bo Li, Khe Chai Sim

Google Research
{sandyritchie,youchicheng,mingqing,mathews,dvanesch,boboli,khechai}@google.com

Abstract

Almost none of the 2,000+ languages spoken in Africa have widely available automatic speech recognition systems, and the required data is also only available for a few languages. We have experimented with two techniques which may provide pathways to large vocabulary speech recognition for African languages: multilingual modeling and self-supervised learning. We gathered available open source data and collected data for 15 languages, and trained experimental models using these techniques. Our results show that pooling the small amounts of data available in multilingual end-to-end models, and pre-training on unsupervised data can help improve speech recognition quality for many African languages.

Index Terms: Africa, multilingual speech recognition, self-supervised learning, low-resource languages

1. Introduction

Africa is home to over 2,000 languages, but almost none of them have widely available automatic speech recognition (ASR) systems. The data required to train such systems is also only available for a few of the major languages, and not in the quantities typically available for high-resource languages. As more people in Africa come online in the 2020s and beyond, the need for high quality ASR systems for African languages becomes more pressing, and along with it the need to develop data sets, model types and training architectures which are optimized for these languages.

Two strands in ASR research which may provide solutions are multilingual end-to-end (E2E) modeling [1] [2] and self-supervised learning (SSL) [3] [4]. These techniques have been applied in the low-resource scenario with some encouraging results [5] [6] [7] [8], and pre-training has also been shown to help with accented speech [9]. However, efforts to develop ASR for African languages have so far mostly focused on data collection and model training for single languages or a small number of related or similar languages, including Amharic [10] [11]; Dinka [11]; South African languages [12] [13] [14]; Nigerian English [15]; Hausa [16] [17]; Yoruba [18] [19] [20] [21]; Swahili [22] [23]; Wolof [24]; Somali [25]; Igbo & Fon [26]; Bemba [27] and Akan [28].

Alongside the data collections initiated for these projects, there are also a few open source projects which aim to improve African language data coverage. Mozilla Common Voice (MCV) [29] provides a web-based platform for donating audio data [29], and has released data sets for Hausa, Kinyarwanda, Kabyle and Luganda, with Igbo and Swahili coming soon. Notably, and in large part thanks to the efforts of Digital Uganda, an artificial intelligence company based in Kigali, MCV provides over 2,000 hours of supervised data for Kinyarwanda, the second largest resource for a language on their platform after English. The South African Centre for Digital Language Resources (SADiLaR) also provides access to various resources, including the NCHLT speech corpus of the South African languages [30]. There are other efforts under the umbrella of the Lacuna Fund to create even more data sets, but there is still lots to be done in this space.

In general, technology companies have not yet invested significantly in the development of ASR systems for African languages. For a number of years, Google has supported voice search and voice typing for four African languages: Afrikaans, Amharic, Swahili and Zulu, but many more languages are spoken in Africa. In this paper, we report on our work to develop or improve ASR systems for 15 African languages: Akan, Hausa, Igbo, Ndebele, Northern Sotho, Kinyarwanda, Swati, Southern Sotho, Swahili, Tsswana, Tsonga, Venda, Xhosa, Yoruba and Zulu. This group of languages was selected for various reasons. Firstly, they are major languages in key regions and countries. Secondly, they share many features in common: they all use the Latin alphabet as their primary writing system, with the exception of Hausa they are all members of the Atlantic-Congo language family, and with the exception of Swahili they are all tonal languages. Finally, some kind of data is available for all of these languages, primarily from open source repositories or through data collections conducted by Google.

Using the available open source and collected data, we trained ASR classic hybrid models, multilingual E2E models, and SSL models. The classic ASR models consist of dedicated connectionist temporal classification (CTC) acoustic models [31], pronunciation models and FST-based n-gram language models for each language. We also trained a multilingual Conformer model: a type of LSTM-based recurrent neural network transducer (RNN-T) [32] in which the recurrent encoder is replaced with convolution-augmented transformer layers [33]. For our first experiment with SSL, we pre-trained a model with unsupervised data using wav2vec 2.0 [34] and fine-tuned it for each of the languages, following [35]. We also experimented with pre-training on a larger high-resource unsupervised data set.

Our results show that pooling the available data in multilingual models, and pre-training on unsupervised data show improvements for African languages compared with classic ASR models. However, while lots of efforts have been spent on creating open source data sets, we have found that the lack of high

https://commonvoice.mozilla.org/
quality data, which is typically available for high-resource languages, has limited our ability to train models which meet the quality bar for launch in end-user-facing applications.

2. Languages

The 15 languages which are the focus of this paper are spoken across West, East and Southern Africa by an estimated 242 million people [36]. With the exception of Hausa which is an Afroasiatic language, all the languages form part of the large Atlantic-Congo language family. 11 of the 15 are also more closely related in the Bantu group. See Table 1 for the families and population estimates for each language.

Table 1: Language families and population counts

| ISO | Family                  | Population |
|-----|-------------------------|------------|
| ak  | Central Tano            | 13M        |
| ha  | Chadic (Afroasiatic)    | 70M        |
| ig  | Igbo                    | 18M        |
| nr  | Nguni (Bantu)           | 1M         |
| nso | Sotho-Tswana (Bantu)    | 14M        |
| rw  | Rwanda-Rundi (Bantu)    | 11M        |
| ss  | Nguni (Bantu)           | 3M         |
| st  | Sotho-Tswana (Bantu)    | 14M        |
| sw  | Sabaki (Bantu)          | 24M        |
| tn  | Sotho-Tswana (Bantu)    | 8M         |
| ts  | Tswana-Ronga (Bantu)    | 13M        |
| ve  | Southern Bantu          | 2M         |
| xh  | Nguni (Bantu)           | 19M        |
| yo  | Yoruboid                | 21M        |
| zu  | Nguni (Bantu)           | 11M        |

We include Hausa in our group despite its different lineage because it shares some common features with its Atlantic-Congo neighbours, most notably lexical and grammatical tone. In grammatical tone systems, each word has an inherent tonal pattern, and these patterns can also change to indicate grammatical features, for example tense and aspect on verbs [37]. All languages in the group except for Swahili exhibit lexical and/or grammatical tone. Another distinctive phonological feature common among several of the languages is the use of click consonants. These are especially prevalent in the Southern Bantu languages, and likely came into these languages through contact with neighbouring Khoisan languages [38]. Xhosa has a particularly large set of 18 click consonants.

All the languages use a form of the Latin script as their primary writing system. In some languages like Xhosa and Zulu, no special characters are used, and phonological features like clicks are represented using Latin letters like <c> and <x>. In other languages, the basic Latin set is supplemented with other characters and diacritics to represent tone and other phonological features. Commonly, high tone is marked with an acute accent, and low tone is marked with a grave accent on tone-bearing vowel and nasal graphemes. Less commonly, other characters like macron, circumflex or caron are used to mark mid tone [39]. Some languages like Hausa do not mark tone in the orthography. A summary of some phonological features and orthographic conventions used for the 15 languages is given in Table 2.

Table 2: Phonological features and orthographic conventions

| Language       | ISO | Tones                          | Clicks | Special characters       | Example word |
|----------------|-----|--------------------------------|--------|--------------------------|--------------|
| Akan           | ak  | high, low                      | no     | open vowels: r, ɔ         | akoko ‘chicken’ |
| Hausa          | ha  | high, low, falling             | no     | implosives: ŋ, ŋ, k, palatalized glottal stop: ’y | faataa ‘spoil’ |
| Igbo           | ig  | high, (mid), low               | no     | ŋ ŋ for tone; ŋ for open vowels | ąkwukwo ‘book’ |
| Ndebele        | nr  | high, low                      | yes    | no                        | umuntu ‘person’ |
| Northern Sotho | nso | high, low                      | no     | ŋ ŋ for ‘sh’ sound; ŋ ŋ for open vowels | moimo ‘work’ |
| Kinyarwanda    | rw  | high, low                      | no     | none                      | urugo ‘home’ |
| Swati          | ss  | high, mid, low                 | yes    | none                      | sandla ‘hand’ |
| Southern Sotho | st  | high, low                      | yes    | none                      | lebese ‘milik’ |
| Swahili        | sw  | none                           | no     | none                      | mto ‘child’ |
| Tswana         | tn  | high, low                      | marginal | ŋ ŋ for ‘sh’ sound; ŋ ŋ for open vowels | šiti ‘cloth’ |
| Tsonga         | ts  | high, low                      | marginal | none                      | xitulu ‘chair’ |
| Venda          | ve  | high, low                      | no     | dental: ŋ, ŋ, ŋ, velar ŋ | tunda ‘support’ |
| Xhosa          | xh  | high, low                      | yes    | none                      | un%xweme ‘shore’ |
| Yoruba         | yo  | high, mid, low                 | no     | ŋ ŋ ŋ for tone; ŋ ŋ ŋ for ‘sh’ sound; ŋ ŋ for open vowels | kojà ‘cross’ |
| Zulu           | zu  | high, low, falling             | yes    | none                      | amanzi ‘water’ |

3. Data

In Tables 3 and 4 the volume of supervised and unsupervised data we used for each language is given. We have used available open source data where possible, for example we have a million utterances and 1,406 hours of training data for Kinyarwanda thanks mostly to the open source repository provided by MCV [7]. In other languages, the basic Latin set is supplemented with other characters and diacritics to represent tone and other phonological features. Commonly, high tone is marked with an acute accent, and low tone is marked with a grave accent on tone-bearing vowel and nasal graphemes. Less commonly, other characters like macron, circumflex or caron are used to mark mid tone [39]. Some languages like Hausa do not mark tone in the orthography. A summary of some phonological features and orthographic conventions used for the 15 languages is given in Table 2.

In the second type of collection, contributors are shown images of common objects and scenes and record them out loud. In the second type of collection, contributors are shown images of common objects and scenes and record themselves describing the image in a few words. These descriptions are then transcribed by other contributors. Supervised long form
data was collected by identifying public videos on YouTube in the target language and transcribing them. The videos were manually identified by linguists at Google and verified as containing content in the target language by the transcribers. We applied the voice activity detection model used in [35] to segment the video. The unsupervised short form data was collected from voice search queries where users had opted in to help Google develop and improve its audio recognition technologies and the Google services that use them.

The unsupervised long form data was collected by identifying public YouTube videos and extracting the audio from the video.

### 4. Experiments

We conducted our experiments with four settings. The details are as follows:

1. **Classic ASR models.** We developed grapheme-to-phoneme (G2P) to generate pronunciation lexicons, gathered text data and developed text normalization grammars to train n-gram language models, and trained custom CTC acoustic models for seven of the languages. We did not train classic models for Ndebele, Northern Sotho, Swati, Southern Sotho, Tswana, Tsonga, Venda, Xhosa or Yoruba, because at the time the NCHLT corpus was not available to us, and in the case of Yoruba because we had issues with data quality.

2. **Multilingual Conformer.** We trained a multilingual Conformer model — this is a modified LSTM-based RNN-T architecture in which the recurrent encoder is replaced with convolution-augmented transformer layers. The encoder is composed of 12 layers, with left 3 frames stacked 128-channel log-mel features for the 512 conformer model dimension. A streaming encoder is used, where each local self-attention layer looks at 23 left context and 0 right context. The model is trained with FastEmit [40] and Hybrid Autoregressive Transducer (HAT) factorization [41], where the latter technique enables better integration with an external language model.

3. **Multilingual Pre-Training (MPT):** Following Zhang et al.’s work on BigSSL [35], the unsupervised data listed in Table 4 are segmented for pre-training with the wav2vec 2.0 objective [42] and are finetuned on each of the 15 languages using the standard RNN-T loss as the downstream task. We used the 600M-parameter Conformer model, the same architecture reported in [35] for both pre-training and finetuning.

4. **Larger Dataset Multilingual Pre-Training (LDMPT):** This setting is similar to (3). The difference is that before we pre-trained using unsupervised data from the 15 languages, we first applied the same pre-training objective on a much larger high-resource language unsupervised data set. More specifically, we used 900,000 hours of segmented, unlabeled YT-U audio data from the set reported in [35]. The underlying assumption is that this would make the model generalize better, as the model sees more data with different phonetic distribution, even though there is a language mismatch between the high-resource languages and the African languages.

### 5. Evaluation

We evaluated the different model types on test sets composed of supervised short form utterances, which were split off from the acquired or collected data sets (see Section 3). Table 5 shows word error rates (WERs) for each type of model and setting.

### 6. Discussion

In general, the single-language classic ASR models have higher WERs compared with other techniques. This is likely due to the comparatively small amounts of training data available for each individual language. Issues with transcription quality, auto-generated G2P accuracy and scarce text data may also contribute to the higher WERs for classic models. Swahili and Zulu do not exhibit the same disparities as the other languages, likely because these languages have more training data and their lex-
represent accurately in writing, and there is greater potential for variation, which can be attributed to several factors: spontaneous speech contains more hesitations and other features which are hard to transcribe and transcriptions than we see with prompt-based corpora, conversely, where our only source of short-form training data comes from image description tasks, as is the case for Akan, Hausa, Igbo and Yoruba, the Conformer model outperforms the pre-trained models, except in the case of Hausa. In cases where we have both an image description set and another train set, as is the case for Kinyarwanda and Southern Sotho, the pre-trained models show improvements over the Conformer model.

Comparing the two pre-training settings, we note that 7 languages show improvements if we first pre-train the model with high-resource-language data rather than training solely on unsupervised data. This might be due to the fact that the unsupervised data is of inferior quality: many of the videos have background music and might also suffer from audio clipping issues. Potential solutions might be to filter out the speech-music mixed segments and apply some declipping preprocessing algorithms.

In all settings, WERs are higher for languages with only image description data. This suggests two things: (1) relying on this type of data alone is more likely to result in poorer quality models, and (2) the multilingual Conformer model seems to be more resilient to less reliable data.

Looking into the issues with the image description data sets further, we see two types of problems. The first is that some of the utterances are one word in length, and this may present an issue when these utterances are seen in training. The second issue is that there are more discrepancies between the audio and transcriptions than we see with prompt-based corpora, which can be attributed to several factors: spontaneous speech contains more hesitations and other features which are hard to represent accurately in writing, and there is greater potential for transcribers to mishear or misunderstand the audio content.

### Table 5: Word error rates on different training settings

| ISO | Classic | Conformer | MPT | LDMPT |
|-----|---------|-----------|-----|-------|
| ak  | 39.0    | **26.5**  | 44.9| 44.9  |
| ha  | 37.9    | 32.6      | **29.7** | 29.9 |
| ig  | 48.4    | **35.0**  | 63.4| 62.3  |
| nr  | –       | 13.3      | 7.4 | **6.7** |
| nso | –       | **8.5**   | 10.3| 9.8   |
| rw  | 37.2    | 13.8      | 10.4| **9.8** |
| ss  | –       | 7.0       | 4.4 | **3.8** |
| st  | –       | 9.1       | **4.7** | 5.4 |
| sw  | 23.7    | 23.7      | 6.9 | **6.4** |
| tn  | –       | 4.8       | 1.9 | **1.4** |
| ts  | –       | 7.4       | **2.3** | 2.3 |
| ve  | –       | 13.8      | 11.0| 9.9   |
| xh  | –       | 14.2      | 11.1| **7.1** |
| yo  | –       | 45.6      | 62.4| 60.1  |
| zu  | 23.8    | 23.8      | 15.0| **13.1** |

Issues with transcription quality are not necessarily limited to the image-based data sets. In all the data sets, there are issues with spelling variation and variation in the use of diacritics and special characters outlined in Table 2. We see the following kinds of variation: (1) alternative spellings for words; (2) alternative symbols instead of the standard ones, typically because the standard characters are not easily accessible on many keyboards or other input methods; (3) use of alternative diacritics, for example carons instead of circumflexes or vice versa; (4) variation in use of tone-marking diacritics: in some languages, while the standard orthography requires marking tone on every vowel and nasal, in practice tone marking is only used in cases where the meaning would otherwise be ambiguous, or tone is not marked at all. While the first three types of variation can be comparatively easily standardized with text normalization, deficient or non-existent tone marking is much harder to restore, as the tones associated with words can vary based on grammatical features, meaning that a contextual model is required for tone restoration, e.g. [45]. For languages where tone is marked inconsistently, this may play a significant role in the higher WERs we see for some of the 15 languages which are the subject of this paper.

### 7. Conclusions

We have shown that multilingual RNN-T models and self-supervised pre-training techniques can improve ASR quality for African languages. These are just two techniques among many that have been shown to be useful in the low-resource scenario. Other novel ASR modeling techniques which could help include federated learning and personalization, zero-shot learning, data augmentation using synthesized speech (though text-to-speech is generally not available for these languages either), and adding external LMs which would also enable techniques like second pass rescoring. We hope that this work stimulates further research on ASR for these and other languages of Africa.

### 8. Acknowledgements

We would like to thank Parisa Hagani, Manasa Prasad, Isabel Leal, Neeraj Gaur, Brian Farris, Yun Zhu, Alena Aksenova, Pierric Sans, Landis Baker, Mandy Jordan, Eoin Mahon, Clara Rivera and Wei Han for support and feedback on the work presented in this paper, and François Beaufays and Pedro Moreno for executive support.

### 9. References

[1] S. Toshniwal, T. N. Sainath, R. J. Weiss, B. Li, P. Moreno, E. Weinstein, and K. Rao, “Multilingual speech recognition with a single end-to-end model,” in ICASSP 2018. IEEE, 2018, pp. 4904–4908.

[2] B. Li, R. Pang, T. N. Sainath, A. Gulati, Y. Zhang, J. Qin, P. Haghani, W. R. Huang, M. Ma, and J. Bai, “Scaling end-to-end models for large-scale multilingual ASR,” arXiv preprint arXiv:2104.14830, 2021.

[3] D. Hwang, A. Misra, Z. Huo, N. Siddhartha, S. Garg, D. Qiu, K. C. Sim, T. Stromhan, F. Beaufays, and Y. He, “Large-scale ASR Domain Adaptation using Self-and Semi-supervised Learning,” arXiv preprint arXiv:2110.00165, 2021.

[4] A. Misra, D. Hwang, Z. Huo, S. Garg, N. Siddhartha, A. Narayanan, and K. C. Sim, “A comparison of supervised and unsupervised pre-training of end-to-end models,” in INTERSPEECH, vol. 2021, 2021, pp. 731–735.

---

One counter-example in this trend is Northern Sotho (nso), which shows better results in the Conformer model.
[42] T. N. Sainath, Y. R. He, A. Narayanan, R. Botros, R. Pang, D. J. Rybach, C. Allauzen, E. Variani, J. Qin, A. Graenestein et al., “An efficient streaming non-recurrent on-device end-to-end model with improvements to rare-word modeling,” in INTERSPEECH, 2021, pp. 1777–1781.

[43] M. J. Harvilla and R. M. Stern, “Least squares signal declipping for robust speech recognition,” in INTERSPEECH, 2014.

[44] A. Zupon, E. Crew, and S. Ritchie, “Text normalization for low-resource languages of Africa,” arXiv preprint arXiv:2103.15845, 2021.

[45] F. O. Asahiah, O. A. Odejobi, and E. R. Adagunodo, “Restoring tone-marks in standard Yorùbá electronic text: improved model,” Computer Science, vol. 18, no. 3, 2017.