INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

One size does not fit all in resource-constrained ASR

Ethan Morris[1]*, Robbie Jimerson*[1]*, and Emily Prud'hommeaux*[2,1]

[1]Rochester Institute of Technology, Rochester NY, USA
[2]Boston College, Chestnut Hill MA, USA

ejm8371@g.rit.edu, rcj2772@rit.edu, prudhome@bc.edu

Abstract

The application of deep neural networks to the task of acoustic modeling for automatic speech recognition has resulted in dramatic decreases in ASR word error rates, enabling the use of this technology for interacting with smart phones and personal home assistants in high-resource languages. Developing ASR models of this caliber, however, requires hundreds or thousands of hours of transcribed speech recordings, which presents challenges for the vast majority of the world's languages. In this paper, we investigate the utility of three distinct architectures that have previously been used for ASR in languages with limited training resources. We train and test these systems on publicly available ASR datasets for several typologically and orthographically diverse languages, which were produced under a variety of conditions using different speech collection strategies, practices, and equipment. Although these corpora are comparable in size, we find that no single ASR architecture outperforms all others. In addition, word error rates vary significantly, in some cases within the range of those typically reported for high-resource languages. Our results point to the importance of considering language-specific and corpus-specific factors and experimenting with multiple approaches when developing ASR systems for languages with limited training resources.

Index Terms: automatic speech recognition, low-resource ASR, under-resourced languages

1. Introduction

Automatic speech recognition (ASR) technology is widely used in many modern technologies, particularly in smartphones and personal assistants. Although an active area of research for over 60 years, ASR has only recently achieved accuracy levels high enough to enable these technologies and to change the way speakers of high-resource languages interact with their devices. These improvements in accuracy are largely due to the application of deep neural architectures trained on large volumes of labeled audio and textual training data. These new ASR pipelines require hundreds, if not thousands, of hours of data [1, 2, 3, 4]. Unfortunately, corpora of this size simply do not exist for the vast majority of the world's 7,000 languages, effectively denying access to these technologies and devices for millions of people around the world.

Existing ASR architectures can be modified to require less data. In addition, large models trained on high-resource languages can be applied to a low-resource language through transfer learning, and existing data can be augmented in a variety of ways in order to create additional synthetic data on a scale appropriate for existing architectures. It is not well recognized, however, that each of these approaches requires careful tuning and adaptation to the target language and to the characteristics of the available corpora; it is not the case that every approach works equally well for every language or corpus. One confounding factor is the great variability in how speech corpora are collected. Speech can be spontaneous or read; recordings can be made professionally in a studio or with a smartphone in a café; the number of speakers and their demographics can be restricted or diverse. In addition, languages themselves vary on a large number of typological, linguistic, and orthographic dimensions that can respond in different ways under different training conditions.

In this paper, we explore the impact of this variability on the accuracy of three ASR architectures previously reported to perform well on small speech datasets. We apply these models to corpora for five languages representing a diversity of morphological properties, sound systems, writing systems, speech collection strategies, speaker pools, and recording settings. We find that one of the three architectures typically outperforms the others though not always and sometimes by only very small margins. Furthermore, we note that there is substantial variability in word error rates across languages, even when corpus size, recording quality, and number of speakers is comparable. Together these results suggest that there may not be a single optimal architecture to use for all low-resource settings, but it remains to be determined what factors contribute to the variability in performance. These findings can help inform future work on designing an appropriate ASR architecture for a target speech corpus given its individual recording and linguistic characteristics, particularly for languages that are both low-resource and endangered, where additional data is difficult or impossible to obtain.

2. Prior work

In their work on Senegalese Wolof, Gauthier et al. [5] use Kaldi [6] to train two ASR models: a subspace Gaussian mixture model (SGMM) with maximum mutual information (MMI) and a DNN with state-level minimum Bayes risk (sMBR) criterion. The SGMM+MMI combination [7] allows for the formation of numerous generative models with efficient training of the substates. The inclusion of sMBR into the DNN approach was found to be most effective [8] at determining sequence-discriminative criteria. Both approaches use the feature-space maximum likelihood regression speaker adaptation method [9] as a means of producing speaker-independent results. In a subsequent paper, the authors found that modeling vowel length contrasts improved word error rate [10]. Previous work on ASR Iban [11, 12] focused on data augmentation via leveraging similarities between Iban and Malay, a closely related language with more abundant data. Using Kaldi with feature enhancements similar to those used for Wolof above, the authors trained GMM, SGMM, and DNN ASR systems, yielding the lowest error rates with the former two architectures.

The prior work on ASR for Amharic was carried out using a corpus originally collected for this purpose a number of years before the development of a working ASR system [13]. The focus of much of the subsequent ASR work was on improving

http://dx.doi.org/10.21437/Interspeech.2021-1970

Table 1: *Characteristics of the train and test sets, including number of male speakers, number of female speakers, duration in hours and minutes, NIST speech signal-to-noise ratio, and WADA signal-to-noise ratio.*

Language	Model	# Male	# Female	duration	NIST SNR	WADA SNR
Seneca	train	7	4	9 h 47 m	24.96 ± 12.85	23.51 ± 21.07
	test	7	4	01 h 40 m	25.19 ± 13.14	25.72 ± 24.02
Wolof	train	8	6	16 h 49 m	35.92 ± 5.59	45.03 ± 34.99
	test	1	1	0 h 55 m	35.39 ± 3.77	38.08 ± 14.59
Bemba	train	5	3	14 h 20 m	32.46 ± 12.01	50.65 ± 33.92
	test	1	1	1 h 18 m	27.47 ± 9.75	36.27 ± 22.52
Iban	train	7	10	6 h 48 m	24.51 ± 8.43	18.54 ± 5.28
	test	2	4	1 h 11 m	22.65 ± 8.79	17.23 ± 4.32
Amharic	train	56	44	20 h 01 m	4.19 ± 3.83	18.80 ± 5.69
	test	14	10	0 h 43 m	4.85 ± 3.37	17.978 ± 7.24

output by exploring the use of acoustic, lexical, and language models units of varying sizes (e.g., syllables rather than phones, morphemes rather than words) [14, 15, 16]. This work relied on traditional HMM/GMM ASR models, primarily within the CMU Sphinx ASR toolkit [17]. Previous work on Bemba [18] used the DeepSpeech architecture [3], which required an initial round of training on a large corpus of English followed by tuning via transfer learning to the small Bemba corpus.

WireNet [19], a novel fully-convolutional ASR architecture distinct from the various Kaldi approaches and from deep neural toolkits designed for high-resource languages (e.g., DeepSpeech [3], wav2letter++ [20]), was developed for Seneca, an under-resourced and endangered language indigenous to the United States and Canada. The WireNet architecture and associated training pipeline produced substantially lower word error rates for a 10-hour Seneca corpus than both DeepSpeech, trained using a transfer learning and data augmentation pipeline, and the most widely used architectures available in Kaldi. The main architectural feature is a stack of Inception [21] and ResNet [22] styled blocks, with wide filter widths to emulate the temporal nature of audio. A multi-staged pipeline was employed with transfer learning from a high-resource language, transitioning into heavily augmented training data, before fine-tuning on the unaugmented data. This learning strategy allows for the neural network's weights to be better initialized as the network can use the larger datasets to converge more quickly, before being refined on the original, smaller dataset.

3. Data

We will be comparing three ASR architectures applied to five corpora for resource-constrained languages: the Native American language, Seneca (50-100 speakers); Senegalese Wolof (10 million speakers); Bemba (5 million speakers); Iban (1.5 million speakers); and Amharic (25 million speakers). Each of these corpora was collected in different settings under different conditions. In order to provide an objective measure of recording quality, we calculated two different signal-to-noise ratios averaged over the full training and test sets of each corpus: the NIST Speech Signal to Noise Ratio (STNR) and the WADA Signal to Noise Ratio (SNR). The results are shown the right-most columns in Table 1. We see that Wolof and Iban have comparable SNR under both methods of calculation, as do Iban and Bemba, while the SNR for Amharic is substantially lower than the others under the NIST method but comprable to Iban under the WADA method. We also observe much less variation in SNR for Iban and Amharic than for the other languages.

Seneca (ISO 639-3 see) is a member of the Iroquoian language family. Highly endangered, it is spoken as a first language by around 50 elders and 100 or more second language learners in parts of what is now western New York State in the U.S. and Ontario, Canada. Seneca's current orthography uses the Roman alphabet with diacritics indicating nasality, with a mostly one-to-one grapheme-to-phoneme mapping. Seneca is described as polysynthetic, combining both agglutinative and fusional elements in its morphology and allowing the incorporation of nouns into the verbal morphology. Unlike most of the other corpora, the Seneca audio data consists of spontaneous speech recorded primarily in casual settings over several years from 11 speakers, 7 male and 4 female. The duration and speaker information for the training and test sets are shown in Table 1. The trigram language model was constructed using a combination of transcripts from the training set and all other available written texts collected by linguists, missionaries, and anthropologists for a total of 49,051 (7,625 unique) words.

Wolof (ISO 639-3 wol) is a member of the Niger-Congo language family spoken by approximately 10 million people in Senegal, the Gambia, and Mauritania. The orthography used in this corpus uses the Roman alphabet with diacritics, for a total of 29 characters with a mostly one-to-one character-to-phone mapping. Since the language has phonemically contrastive vowel length, consonant gemination, and prenasalization, the actual phonetic inventory is somewhat larger than the number of characters used to write the language. Wolof is agglutinative, with a rich inflectional and derivational morphology. Unlike many Niger-Congo languages, Wolof is not tonal [23]. The Wolof audio data used here was recorded in a controlled environment using read speech from 18 speakers, 10 male and 8 female, and consists of 18,000 utterances between 6 and 12 words long [5]. The duration and speaker information about the training and test sets are shown in Table 1. No speaker occurs in both the training and the test set. The trigram language model accompanying this corpus was trained on the transcripts of the audio, as well as a combination of physical books and data scraped from the internet, for a total of 601,609 (29,148 unique) words.

Bemba (ISO 693-3 bem) is also a member of the Niger-Congo language family spoken by approximately 5 million people primarily in Zambia. The orthography uses 23 characters or character combinations of the Roman alphabet to represent its 24 phonemes in a fairly regular 1-to-1 mapping. It has two tones, one of which is marked with an acute accent on the associated vowel, resulting in an addition 5 characters. Like Wolof, it is agglutinative, with a rich derivational and inflectional morphology. The audio data consists of 14,438 utterances ranging from one to

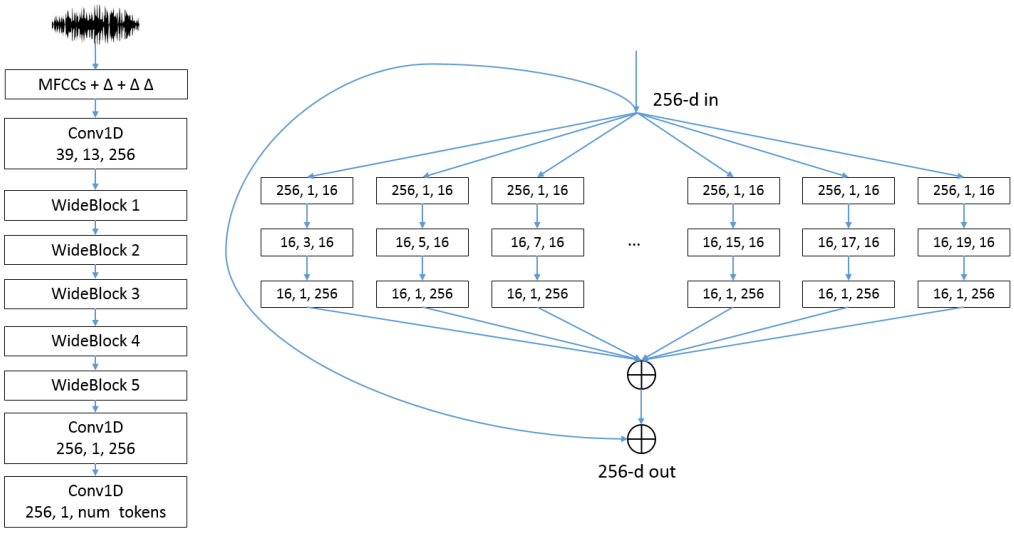

Figure 1: *Left: The overall WireNet architecture. Right: A bottleneck block consisting of 9 paths, each with bottleneck filters centered by filters of different width to capture different temporal dependencies. Each layer shows (# input channels, filter width, # output channels).*

twenty words of read speech recorded in an uncontrolled setting to allow for the presence of background noise and more natural modes of speech [18] from 6 men and 4 women all students in their 20s. The duration and speaker information about the training and test sets are shown in Table 1. No speaker occurs in both the training and the test set. The trigram language model accompanying this corpus was built using the transcripts of the audio, as well as a combination of other web and text sources, for a total of 5.8 million (189,000 unique) words.

Iban (ISO 693-3 iba) is a member of the Malayo-Polynesian branch of the Austronesian language family spoken by approximately 1.5 million people primarily in Borneo. The orthography used in this corpus consists of 27 characters from the Roman alphabet with a regular 1-to-1 character-to-phoneme mapping. It is an agglutinative language with a fairly rich morphology. The audio data consists of both read news and spontaneous speech from Malaysian television and radio [11]. This is the smallest of the five corpora with just under 7 hours of speech consisting of more than 3,000 utterances. The duration and speaker information about the training and test sets are shown in Table 1. No speaker occurs in both the training and the test set. The trigram language model accompanying this corpus was building using the transcripts of the audio, as well as web news articles, for a total of 2,082,452 (36,310 unique) words.

Amharic (ISO 639-3 amh) is a member of the Semitic branch of the Afro-Asiatic language family spoken by around 25 million people, primarily in Ethiopia. In the Amharic writing system, each of the 240 characters represents a CV syllable with a reliable one-to-one character-to-syllable mapping. The consonant inventory is somewhat large, with a notable three-way stop contrast (voiced, voiceless, ejective). Amharic combines morphological affixation with a root-pattern morphology, in which vowels are inserted or changed within a stable consonant template. The Amharic audio data was recorded in a controlled environment using read speech from 124 speakers, 70 male and 54 female, for a total of 10,850 sentences [13]. The duration and speaker information for the training and test sets are shown in Table 1. No speaker occurs in both partitions. The trigram language model was constructed using the transcriptions of the audio in addition to other text corpora resulting in 120,262 sentences and 2.5 million words.

4. Methodology

4.1. Speech feature extraction

The acoustic models were built utilizing the 13 Mel-frequency cepstral coefficients (MFCCs) and their delta- and delta-delta features using a 25 ms window with 10 ms stride. These features allow for the modeling of the time-dependent audio signal through the inclusion of the derivative approximations. In the Kaldi environment, linear discriminant analysis (LDA) transformation, maximum likelihood transform (MLLT), and feature-space maximum likelihood linear regression (fMLLR) are applied for dimensionality reduction, better robustness when applying to test data, and speaker normalization, respectively.

4.2. Architectures

The original WireNet architecture, Figure 1, included 5 bottleneck blocks, each with 9 paths of varying filter width, stacked upon one another with skip connections between each. A linear sweep was performed to explore the potential optimisations that may occur with varying the number of stacked blocks and their path depths. Through this, we determined that 4 blocks with a width of 21 produced lower word error rate (WER) results than the original 5+9 combination. Additionally, measurements were taken after each stage of the learning pipelines (transfer, augmentation) described in Thai et al. [19] in order to determine the effectiveness of each stage. We found that after applying the transfer learning from the 960-hour LibriSpeech English corpus, the initial epochs of the next stage were spent un-learning these initialized weights and that merely introducing the augmented data as the first stage was sufficient. Decoding was carried out using CTC decoding with a trigram language model.

We experimented with several neural and non-neural architectures available within the Kaldi toolkit. We report here on the best-performing of each: the SGMM [7], which was used previously for Wolof [5] and Iban [11, 24]; and a simple, fully

Table 2: *Word error rate (WER) for each language under the three architectures and two train/test split settings: the original, in which no speaker has utterances in both the train and test sets (Disjoint), and a new split in which each speaker's utterances are split between the training and test sets (Overlap).*

Language	Model	WER	
		Disjoint	Overlap
Seneca	SGMM	-	33.9
	Kaldi DNN+sMBR	-	30.6
	WireNet	-	**24.3**
Wolof	SGMM	25.1	28.3
	Kaldi DNN+sMBR	**24.9**	27.4
	Modified WireNet	29.8	**14.3**
Amharic	SGMM	8.4	4.8
	Kaldi DNN+sMBR	**7.5**	**4.1**
	Modified WireNet	17.4	15.0
Iban	SGMM	16.4	13.0
	Kaldi DNN+sMBR	**15.1**	**12.9**
	Modified WireNet	34.3	19.8
Bemba	SGMM	58.4	13.9
	Kaldi DNN+sMBR	**53.4**	**12.3**
	Modified WireNet	64.4	48.8

connected DNN model with 6 hidden layers, each with 1024 units. The weights were initialized using Restricted Boltzmann Machines, whose small size with quick training methodologies [25] allow for faster convergence. Sequence training was performed using sMBR criterion and a per-utterance Stochastic Gradient Descent weight update. Decoding was again carried out using CTC decoding with a trigram language model.

4.3. Data re-partitioning

The Seneca corpus was deliberately partitioned in order to have each of the speakers represented in both the training and test sets, and the composition of the speakers makes it nearly impossible to have disjoint partitions while maintaining an adequately sized training corpus. An a critically endangered language, Seneca has a very limited and finite set of speakers, which makes the ability to customize models to specific speakers desirable. This kind of partitioning, however, is rare in most ASR corpora where speaker independence is preferred [20, 26, 27]. In order to compare the three architectures for all five corpora under different train-test split strategies, experiments were conducted on the other four corpora with two datasets: one with the original partitioning containing disjoint speakers, and one where the train and test set were combined, shuffled, and then split into train and test sets of sizes equivalent to the original partitioning but where each speaker is represented in both training and testing.

5. Results

Table 2 shows the word error rates (or in the case of Amharic, morpheme error rates) for the three architectures (Kaldi SGMM, Kaldi DNN, and WireNet) under the two train-test splits for each of the four corpora for which the two splits are available and under the available overlapping split for Seneca. We note that the results reported here are either on par with or superior to the prior results for these languages (Section 2).

For all four corpora with available disjoint train-test splits, both Kaldi models yield a lower WER than WireNet with the DNN outperforming the SGMM. In three of the four languages, the difference between the neural and non-neural Kaldi models is modest, while in Bemba this difference is quite large. We note that the difference in WER between Kaldi and WireNet is much smaller for Wolof and Bemba, whose WER overall are relatively high.

However, when trained and tested on overlapping train-test splits, in which each speaker is represented in both the testing and the training data, results were more variable. In particular, WireNet outperforms the Kaldi DNN+sMBR architecture for both Seneca and Wolof. In fact, the DNN+sMBR model for Wolof saw an 11.85% *increase* in WER when trained on overlapping data, while the modified WireNet saw a 52.02% decrease in WER. WER for the overlapping train-test splits for the remaining three languages saw reductions in WER for all three architectures, with both Kaldi models continuing to outperform the WireNet model, sometimes substantially (Bemba) and other times by a smaller margin (Iban).

Overall, there is a very wide variance in WER across the corpora. Even after excluding Amharic, which is technically evaluated at the morpheme level rather than the word level, we see WER as low as 15.1 and as high as 64.4. This variability does not appear to be tied directly to the size of the acoustic training corpus, the number of speakers in the corpus, or any specific linguistic or typological feature. The great disparity in the overlap setting between Bemba and Wolof is particularly puzzling, given the similar characteristics of the two languages, the speaker distribution, and the corpora in terms of their size and recording quality.

6. Discussion

The experiments described here are insufficient to determine which features of the languages, the corpora, the speakers, or the recordings themselves, might account for the differences in performance between WireNet and the two Kaldi architectures for these languages in the two train-test partitionings. In our future work, we will continue to experiment with these architectures and others (e.g., Kaldi's TDNN, which outperformed the SGMM and DNN, though not WireNet, for Seneca) with additional freely available small corpora, with the goal of identifying the source of the variability observed here.

We note that four of these five languages, while having few acoustic training resources, are widely spoken and have established written traditions adequate to support the training of large language models and the collection of additional data. Any ASR system built for these languages should be robust to speaker variation. A substantial majority of the world's 7000 languages, however, are endangered and in need of documentation and preservation, much like Seneca. Architectures designed specifically to be speaker independent may not be the best option for the documentation of languages with very few speakers.

7. Acknowledgments

We are grateful for the cooperation and support of the Seneca Nation of Indians. This material is based upon work supported by the National Science Foundation under Grant No. 1761562. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

8. References

[1] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, pp. 2613–2617, 2019.

[2] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina *et al.*, "State-of-the-art speech recognition with sequence-to-sequence models," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4774–4778.

[3] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International Conference on Machine Learning*, 2016, pp. 173–182.

[4] K. Audhkhasi, B. Kingsbury, B. Ramabhadran, G. Saon, and M. Picheny, "Building competitive direct acoustics-to-word models for english conversational speech recognition," in *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2018, pp. 4759–4763.

[5] E. Gauthier, L. Besacier, S. Voisin, M. Melese, and U. P. Elingui, "Collecting resources in Sub-Saharan African languages for automatic speech recognition: A case study of Wolof," in *Proceedings of the Language Resources and Evaluation Conference*, 2016, pp. 3863–3867.

[6] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi Speech Recognition Toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 2011, IEEE Catalog No.: CFP11SRW-USB.

[7] D. Povey, L. Burget, M. Agarwal, P. Akyazi, F. Kai, A. Ghoshal, O. Glembek, N. Goel, M. Karafiát, A. Rastrow, R. C. Rose, P. Schwarz, and S. Thomas, "The subspace Gaussian mixture model: A structured model for speech recognition," *Computer Speech & Language*, vol. 25, no. 2, pp. 404–439, 2011.

[8] K. Veselý, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, pp. 2345–2349, 01 2013.

[9] M. Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," *Computer Speech & Language*, vol. 12, no. 2, pp. 75–98, 1998.

[10] E. Gauthier, L. Besacier, and S. Voisin, "Automatic speech recognition for african languages with vowel length contrast," *Procedia Computer Science*, vol. 81, pp. 136–143, 2016.

[11] S. F. S. Juan, L. Besacier, and S. Rossato, "Semi-supervised G2P bootstrapping and its application to asr for a very under-resourced language: Iban," in *Proceedings of the International Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU)*, 2014, pp. 66–72.

[12] S. S. Juan, L. Besacier, B. Lecouteux, and M. Dyab, "Using resources from a closely-related language to develop ASR for a very under-resourced language: A case study for Iban," in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, 2015.

[13] S. Abate, W. Menzel, and B. Tafila, "An Amharic speech corpus for large vocabulary continuous speech recognition," in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, 2005, pp. 1601–1604.

[14] M. Y. Tachbelie, S. T. Abate, and W. Menzel, "Morpheme-based and factored language modeling for Amharic speech recognition," in *Proceedings of the Fourth Conference on Human language technology: Challenges for computer science and linguistics*, 2009, pp. 82–93.

[15] ——, "Morpheme-based automatic speech recognition for a morphologically rich language - Amharic," in *Proceedings of the Workshop on Spoken Languages Technologies for Under-Resourced Languages (SLTU)*, 2010.

[16] M. Tachbelie, S. T. Abate, and L. Besacier, "Using different acoustic, lexical and language modeling units for ASR of an under-resourced language - Amharic," *Speech Communication*, vol. 56, 2014.

[17] P. Lamere, P. Kwok, E. Gouvea, B. Raj, R. Singh, W. Walker, M. Warmuth, and P. Wolf, "The CMU SPHINX-4 speech recognition system," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, 2003, pp. 2–5.

[18] C. Sikasote and A. Anastasopoulos, "BembaSpeech: A speech recognition corpus for the Bemba language," 2021.

[19] B. Thai, R. Jimerson, R. Ptucha, and E. Prud'hommeaux, "Fully convolutional ASR for less-resourced endangered languages," in *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, 2020, pp. 126–130.

[20] V. Pratap, A. Hannun, Q. Xu, J. Cai, J. Kahn, G. Synnaeve, V. Liptchinsky, and R. Collobert, "qav2letter++: A fast open-source speech recognition system," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6460–6464.

[21] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[23] C. M. B. Dione, "A morphological analyzer for Wolof using finite-state techniques," in *Colloquium on African Languages and Linguistics*, 2011.

[24] S. F. S. Juan, L. Besacier, and S. Rossato, "Semi-supervised G2P bootstrapping and its application to ASR for a very under-resourced language: Iban," in *Proceedings of the Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU)*, 2014.

[25] G. E. Hinton, *A Practical Guide to Training Restricted Boltzmann Machines*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 599–619.

[26] H. Sailor, A. Patil, and H. Patil, "Advances in Low Resource ASR: A Deep Learning Perspective," in *Proceedings of the Sixth International Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU)*, 2018, pp. 15–19.

[27] H.-T. Luong and H.-Q. Vu, "A non-expert Kaldi recipe for Vietnamese speech recognition system," in *Proceedings of the Third International Workshop on Worldwide Language Service Infrastructure and Second Workshop on Open Infrastructures and Analysis Frameworks for Human Language Technologies*, 2016, pp. 51–55.