INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Development of Multilingual ASR Using GlobalPhone for Less-Resourced Languages: The Case of Ethiopian Languages

Martha Yifiru Tachbelie[1,2], *Solomon Teferra Abate*[1,2], *Tanja Schultz*[1]

[1]Cognitive Systems Lab, University of Bremen, Germany
[2]School of Information Science, Addis Ababa University, Ethiopia

marthayifiru,abate,tanja.schultz@uni-bremen.de

Abstract

In this paper, we present the cross-lingual and multilingual experiments we have conducted using existing resources of other languages for the development of speech recognition system for less-resourced languages. In our experiments, we used the Globalphone corpus as source and considered four Ethiopian languages namely Amharic, Oromo, Tigrigna and Wolaytta as targets. We have developed multilingual (ML) Automatic Speech Recognition (ASR) systems and decoded speech of the four Ethiopian languages. A multilingual acoustic model (AM) trained with speech data of 22 Globalphone languages but the target languages, achieved a Word Error Rate (WER) of 15.79%. Moreover, including training speech of one closely related language (in terms of phonetic overlap) in ML AM training resulted in a relative WER reduction of 51.41%. Although adaptation of ML systems did not give significant WER reduction over the monolingual ones, it enables us to rapidly adapt existing ML ASR systems to new languages. In sum, our experiments demonstrated that ASR systems can be developed rapidly with a pronunciation dictionary (PD) of low out of vocabulary (OOV) rate and a strong language model (LM).

Index Terms: multilingual speech recognition, Ethiopian languages, GlobalPhone, less-resourced languages

1. Introduction

With more than 7000 languages in the world [1] and the need to support multiple input and output languages, it is one of the most pressing challenge for the speech and language community to develop and deploy speech processing systems in yet unsupported languages rapidly and at reasonable costs [2, 3]. Major bottlenecks are the sparseness of speech and text data with corresponding PDs, the lack of language conventions, and the gap between technology and language expertise. Data sparseness is a critical issue due to the fact that speech technologies heavily rely on statistical modeling schemes, such as Hidden Markov Models (HMM), Deep Neural Networks (DNN) for acoustic modeling as well as n-gram and DNN for language modeling. Although statistical modeling algorithms are mostly language independent and proved to work well for a variety of languages, reliable parameter estimation requires vast amounts of training data. Unfortunately, large-scale data resources for research are available for only a fraction of the worlds' languages and the costs for these collections are prohibitive to all but the most widely spoken and economically viable languages. This calls for the development of cross-lingual and/or multilingual speech processing/recognition systems [4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18].

In cross-lingual (CL) ASR, resources of source (or donor) languages are used to develop an ASR system for a target language with or without little adaptation data from the target language. [14] described multilingual (ML) ASR as system in which at least one of the components (feature extraction, AM, PD, or LM) is developed using data from many different languages. Although ML ASR systems are useful in other contexts, they are particularly interesting for under-resourced languages where training data are sparse or not available at all [4]. Furthermore, they provide an appealing solution for multilingual, multi-Ethnic and economically disadvantaged countries to mitigate the digital divide.

Ethiopia is a multilingual and multi-ethnic country where over 80 languages are spoken. When it comes to language resources required for the development of speech processing tools, all Ethiopian languages (arguably except Amharic) are under-resourced [19]. On the other hand, developing large-scale language resources is not economically viable. Thus, alternative approaches need to be used to make Ethiopians benefit from speech processing tools which are important due to low literacy rate (especially in rural areas) and widespread use of cell-phones. Accordingly, we have investigated the development of CL and ML ASR system for Ethiopian languages using resources of other languages. For this purpose, we have used GlobalPhone [20], a multilingual database of high-quality read speech with corresponding transcriptions and PDs in more than 20 languages. Although almost all Ethiopian languages are under-resourced, we focus in this study on four Ethiopian languages (Amharic and Tigrigna from Semitic, Oromo from Cushitic and Wolaytta from Omotic language families) since we have a medium sized speech corpus for each of them. This enable us to investigate how well the CL and/or ML ASR systems perform and generate proof-of-concepts of using various source/donor languages for rapid adaptation to any Ethiopian target language in question.

In this paper, we present CL and ML language adaptation methods for the development of ML ASR systems in four Ethiopian languages.

2. Speech corpora

2.1. Globalphone

GlobalPhone (GP) is a multilingual corpus that comprises (1) speech data, i.e. high-quality recordings of spoken utterances read by native speakers, (2) corresponding transcriptions, (3) pronunciation dictionaries covering the vocabulary of the transcripts, and (4) baseline n-gram LMs. The first two are referred to as GP Speech and Text Database (GP-ST), the third as GP Dictionaries (GP-Dict), and the fourth as GP LMs (GP-LM). GP-ST is distributed under a research or commercial license by two authorized distributors, the European Language Resources Association (ELRA) [21] and Appen Butler Hill Pty Ltd. [22]. GP-Dict is distributed by ELRA, while the GP-LMs are freely available for download from our website [23].

http://dx.doi.org/10.21437/Interspeech.2020-2827

The entire GP corpus provides a ML database of word-level transcribed speech for the development and evaluation of large vocabulary speech processing systems in the most widespread languages of the world. GP is designed to be uniform across languages with respect to the amount of data per language, the audio quality (microphone, noise, channel), the collection scenario (task, setup, speaking style), as well as the transcription and phone set conventions (IPA-based naming of phones in all PDs). Thus, GP supplies an excellent basis for research in the areas of (1) ML ASR, (2) rapid deployment of speech processing systems to yet unsupported languages, (3) language identification tasks, (4) speaker recognition in multiple languages, (5) ML speech synthesis, as well as (6) monolingual ASR.

Currently, the GP corpus covers 22 languages, i.e. Arabic (modern standard), Bulgarian, Chinese (Mandarin and Shanghai), Croatian, Czech, French, German, Hausa, Japanese, Korean, Polish, Portuguese, Russian, Spanish, Swahili, Swedish, Tamil, Thai, Turkish, Ukrainian, and Vietnamese. Detailed description of GP can be found [20].

2.2. Ethiopian languages corpora

Read speech corpora of four Ethiopian Languages (Amharic, Tigrigna, Oromo and Wolaytta) are used in our experiments. For Amharic, we have two types of speech data. The Amharic speech corpus [24], referred as AMH2005, contains 20 hours of training speech (11k utterances), development and test sets read by 20 other speakers (10 each). The domain of this corpus is broadcast news and the recording was done in a noise free environment, resulting in very clean speech. Moreover, the maximum length of the utterances is limited to 20 words. The second Amharic speech corpus, referred as AMH2020, is prepared together with the preparation of speech corpora of the other three languages as described below.

The corpora of the Ethiopian languages have been collected in Ethiopia [25]. The Amharic (AMH2020), Tigrigna and Oromo speech corpora consist of speech from 98 speakers per language. Most of the speakers read around 125 sentences. The Wolaytta corpus consists of recordings of 85 speakers where most of them read 140-150 sentences. The domain of these four corpora is mixed, it includes utterances from the news domain, the bible, other religious books, etc. The recordings were done using smartphones in different environments and as a result, they are not as clean as the AMH2005 corpus. In addition, there is no limit with the maximum length of the utterances when sentences are selected. Development and evaluation sets (speech of 4 speakers per set) are held out from the total recordings of each of the corpora. The size (in terms of hours), number of speakers, and utterances for training, development and evaluation sets of the corpora are given in Table 1.

Table 1: *Details of the corpora size in Hours(No. of speakers, No. of utterances)*

Corpus	Train Set	Dev. Set	Eval. Set
AMH2005	20(100, 10875)	1.5(10, 760)	1.5(10, 760)
AMH2020	24(90,11274)	1.2(4, 507)	1.3(4, 508)
ORM	22.8(90, 11297)	1.2(4, 505)	1.1(4, 501)
TIR	22.1(90, 11305)	1.1(4, 511)	1.0(4,507)
WAL	29.7(77, 10939)	1.5(4, 553)	1.7(4, 578)

3. Multilingual ASR experiments

3.1. Acoustic models

All AM have been built in a similar fashion using Kaldi ASR toolkit [26]. We have built ML (referred as MLnn(lang), where nn indicates the number of languages used in the training and (lang) indicates the language whose training speech is used together with the 22 GP languages' data) context dependent HMM-GMM based AMs using 39 dimensional mel-frequency cepstral coefficients (MFCCs) to each of which cepstral mean and variance normalization (CMVN) is applied. The AMs use a fully-continuous 3-state left-to-right HMM. Then we did Linear Discriminant Analysis (LDA) and Maximum Likelihood Linear Transform (MLLT) feature transformation for each of the models, followed by Speaker Adaptive Training (SAT) using an affine transform, feature space Maximum Likelihood Linear Regression (fMLLR). We performed parameter tuning to find the best number of states and Gaussians for the ML AMs.

To train the DNN-based AMs, we have used the best HMM-GMM models to get alignments and the same training speech used to train HMM-GMM models. We have applied a three-fold data augmentation [27] prior to the extraction of 40-dimensional MFCCs without derivatives, 3-dimensional pitch features and 100-dimensional i-vectors for speaker adaptation. We experimented using two recipe (SWBD and WSJ) and the results are better with WSJ recipe with the following DNN configuration. The neural network architecture we used is Factored Time Delay Neural Networks with additional Convolutional layers (CNN-TDNNf). The Neural network has 15 hidden layers (6 CNN followed by 9 TDNNf) and a rank reduction layer. The number of units in the TDDNf consists of 1024 and 128 bottleneck units except for the TDNNf layer immediately following the CNN layers which has 256 bottleneck units.

3.2. Pronunciation and language modeling

Phone-based PDs are available for each GP language. The GP-Dicts cover the words which appear in the training transcriptions. The majority of the dictionaries were constructed in a rule-based manner using language specific phone sets, and then manually cross-checked by native experts. To enable the development of ML speech processing, the phone names are made consistent across languages, leveraging the International Phonetic Alphabet (IPA) [28]. For the ML AM training, PDs of the source languages are combined.

The PDs and LMs of the four Ethiopian target languages are used to decode speech of the languages. For Amharic and Tigrigna, the PDs are prepared automatically taking the Consonant-Vowel syllabary feature of the writing system. For Oromo and Wolaytta, PDs are prepared automatically based on their writing systems that indicate gemminated and non-gemminated consonants as well as long and short vowels.

Depending on the availability of text data, we have used different sizes of PDs for the target languages. For Amharic (AMH2005 and AMH2020) and Tigrigna, we have used PDs consisting of 310k, 323k and 299k vocabularies, respectively. The 323k vocabulary PD of AMH2020 corpus is the result of a combination of the 310k dictionary of AMH2005 and the words taken from the training transcription of the AMH2020 corpus. However, for Oromo and Wolaytta, the lists of word entries extracted from the training speech transcriptions have been used. The size of the decoding PDs and the Out of Vocabulary (OOV) rates with regard to the evaluation set is given in Table 2. The phones of the target languages that are not covered in the

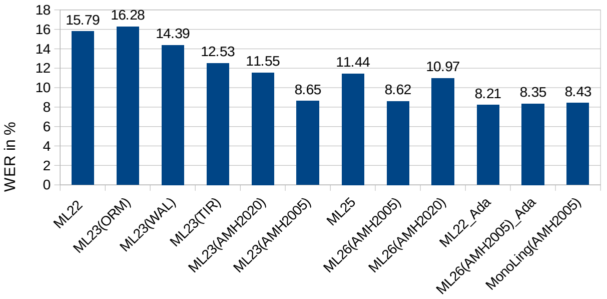

Figure 1: *WER of Amharic*

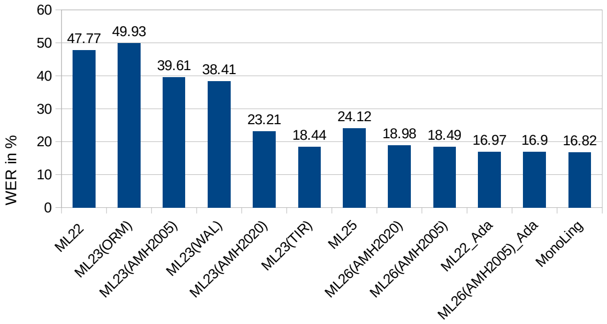

Figure 2: *WER of Tigrigna*

ML22 AM are presented in Table 3. These phones are manually mapped to phonetically nearest phones (long consonants to their short form, plosive ejective to plosive voiceless, rounded vowels to their un-rounded form, etc.) found in the training PDs.

Table 2: *Language models of the target languages*

Corpus	PD Vocab	OOV%	LMToken	PPL
AMH2005	310k	3.1	3.97M	41.2
AMH2020	323k	6.2	4M	241.3
ORM	21k	11.7	1.2M	266.2
TIR	299k	4.9	4M	211.4
WAL	25k	9.3	226k	254.9

Table 3: *Phones not covered by GP-Dicts used for training*

Languages	Phones
Amharic	p’ t’ tʃ’ ue uɨ
Tigrigna	x’ p’ t’ tʃ’ ue uɨ
Oromo	bː ɗː fː gː jː kː p’ p’ː wː tʃː pː k’ː ʤː tʃ’ tʃ’ː t’ t’ː ɲː
Wolaytta	bː ɗː gː jː kː p’ p’ː tʃː pː k’ː ʤː tʃ’ tʃ’ː t’ t’ː zː

For all the target languages, we have developed open vocabulary trigram LMs using the SRILM toolkit [29] and different sizes of text corpus obtained from the web, except for Wolaytta. Since we found no text resources on the web for Wolaytta, only the training transcription has been used to train a trigram LM. The LMs are smoothed with unmodified Kneser-Ney smoothing techniques [30]. For Amharic, we have developed two versions of LMs: one using the text from which the 310k vocabularies are extracted and another by adding the training transcripts of the AMH2020 corpus to this text. Table 2 shows the amount of words in the data used for LM training (LMToken) and the perplexity (PPL) of the best LMs on the evaluation set.

3.3. Experimental Results

To develop the ML AMs we have used ML mix approach [3] leveraging the feature of GP-Dicts and sharing data across the GP languages for those phones which are represented by the same IPA symbol. The first experiment we have done is CL experiment using a ML AM (ML22) developed from training speech of 22 languages of GP. Test speech of the Ethiopian languages (Amharic, Oromo, Tigrigna and Wolaytta) have been decoded using ML22 AM and language specific PDs as well as LMs.

From our previous work [31], we have learned that high phonetic overlap exists among the four Ethiopian languages. Thus, we wanted to see the effect of adding speech data of the Ethiopian languages in ML AM training. Therefore, we developed ML23 AMs by adding training speech of one of the Ethiopian languages at a time to the 22 GP languages’ speech data and decoded test speech of the three Ethiopian languages that are not involved in the AMs training.

Since we have got improvements in performance, by just adding the training speech of one Ethiopian language in ML AM training, we further investigated the use of more data from related languages in ML AM. Thus, we have trained ML25 using speech data from 25 languages by including training speech of three of the Ethiopian languages to the 22 GP languages’ speech data, and decoded the test speech of the remaining Ethiopian language.

Finally, we used the training speech of all (26) languages and trained ML26 using the ML mix approach. Since we have two Amharic corpora, we have developed two versions of ML26 AM (by adding one of the Amharic corpus at a time). Test speech of each of the four Ethiopian languages have been decoded. This experiment enables us to see how data of other languages hurt the ML AM in a ML mix approach where speech data as well as PDs of all the languages are mixed together and language specific contexts become loosely represented.

We have also used transfer learning method to adapt the ML AMs to the target languages. In this case, weights of the hidden layers of ML22 and ML26 models are transferred and new output layers are added. The transferred layers are retrained with the target language training data using smaller learning rate and the output layer is trained with higher learning rate. These models are referred as ML22_Ada and ML26_Ada. The WER of all the models for each of the four Ethiopian languages are depicted in Figures 1, 2, 3 and 4.

As shown in Figure 1, a CL recognition of Amharic test speech using ML22 AM resulted in a WER of 15.79%, without using any training speech from the target language, Amharic. This result is achieved due to the use of a strong LM (with perplexity of 41.2) and a large decoding vocabulary (310k) with OOV rate of 3.06%. Moreover, the test speech, the LM training text and dictionary are also from the same domain, which is news domain. For comparison, we have decoded a test speech of AMH2020 corpus, which is from different domains (news, bible, religious books, etc.), with a LM that has a perplexity of 241.26 and a 323k dictionary with OOV rate of 6.21% and achieved a WER 53.2%. Moreover, the number of phones that are not covered by the training dictionary is small compared

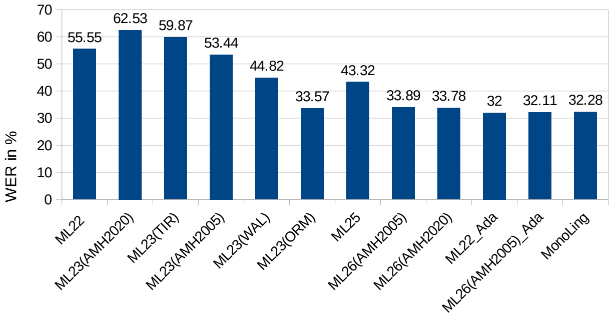

Figure 3: *WER of Oromo*

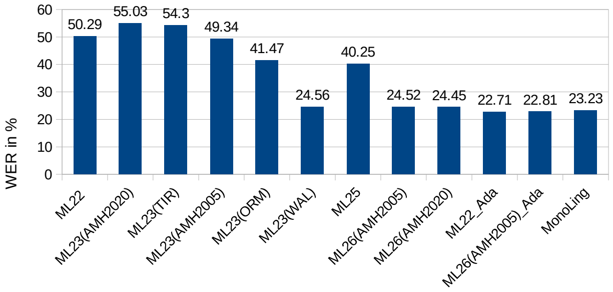

Figure 4: *WER of Wolaytta*

to the other three target languages. This result indicates that a speech recognition system of reasonable performance can be built using training speech of other languages provided that we have a good LM and dictionary with low OOV rate for the target language. One can also observe that adding training speech of a language closely related to Amharic (Tigrigna in this case) in the training, ML23(TIR), resulted in a relative WER reduction of 20.65%, which might be due to the very high phonetic overlap between Amharic and Tigrigna, all Amharic phones (100%) are covered by Tigrigna [31]. However, adding Oromo training speech negatively affected the performance although 78% of Amharic phones are covered by Oromo. Improvement in recognition accuracy (a relative WER reduction of 27.55%) was obtained by adding the training speech of the other three Ethiopian languages (ML25).

CL recognition result of Tigrigna (see Figure 2) using the ML22 AM is worse than the result obtained for Amharic and better than the results of Oromo and Wolaytta. This is attributed to the strength of the LM and the coverage of the dictionary. Compared to Oromo and Wolaytta, better LM (with low perplexity) and dictionary (with low OOV rate) have been used for Tigrigna. Significant WER reduction has been obtained when Amharic speech is added in the training, ML23(AMH2005) and ML23(AMH2020). Although the addition of both AMH2005 and AMH2020 corpora led to WER reductions due to a high phonetic overlap (91% of Tigrigna phones are covered by Amharic), the use of AMH2020 corpus led to a relative WER reduction of 51.41%. This is due to the similar nature of the Tigrigna and AMH2020 corpora, they are from the same domain and recorded in the same setup. The addition of the training speech from the other three Ethiopian language (ML25) did not give improvement over the ML23 system developed using training speech of AMH2020.

Figures 3 and 4 show the WERs for Orromo and Wolaytta, respectively. These languages are closely related although they are from different language families. However, text data for OOV reduction and Language modeling could not be found on the web. Thus, the results of these languages show the performance of ML system in really resource constrained situation. WERs of 55.55% and 50.29% have been obtained in CL recognition experiments using ML22 system for Oromo and Wolaytta, respectively. This is due to the limited quality of the LMs, the limited coverage of the dictionaries, and the relatively high number (compared to Amharic and Tigrigna) of phones that are not covered by the training dictionary of the ML22 system. Since these two languages are closely related, the addition of either of the two languages led to significant WER reductions, see ML23(WAL) in Figure 3 and ML23(ORM) in Figure 4. The addition of Wolaytta training speech brought a relative WER reduction of 19.32% while the addition of Oromo training speech led to a relative WER reduction of 17.54%. When all the training speech, but the target languages, is used in training the ML AM, relative WER reduction of 22.02% and 19.96% have been achieved for Oromo and Wolaytta, respectively.

The ML26 system did not lead to performance improvements over the monolingual systems for any target language. The weight transfer/adaptation of the ML systems (ML22 and ML26) to each of the target languages gave slight performance improvements (though not statistically significant), except for Tigrigna, over the monolingual systems. However, the adapted systems excel in reduction of training time and energy consumption (see Table 4). We are convinced that time and energy resources will be among the most important issues to be addressed in future works on the adaptation of ASR systems to new languages.

Table 4: *Minimum and Maximum Time [hh:mm:ss] and energy consumption [KJ]*

	Train		**Adapt.**	
Lang.	**Time**	**Energy**	**Time**	**Energy**
Amharic	13:20:06	1047.0	03:01:24	307.1
Oromo	15:03:54	1174.1	04:05:48	378.5
Tigrigna	14:57:24	1162.9	03:27:16	335.7
Wolaytta	18:00:09	1470.7	04:35:26	434.8

4. Conclusions

In this work, we presented the results of our investigation towards the use of existing language resources of other languages for the development of speech recognition systems for less-resourced languages in a multilingual setup. Our experimental results show that a recognition system of reasonable performance can be built without using any training speech from the target language as long as we have a strong LM and a dictionary with low OOV rate. In addition, the phonetic overlap between source and target languages has a great impact on the performance of CL and ML ASR. ASR performance on par with the monolingual one can be achieved, if training speech of related languages is used. Although weight transfer or adaptation of ML system to the target language gives no significant benefit in terms of WER reduction, it enable us to adapt a ML system to new unseen language rapidly and energy-efficiently.

5. Acknowledgements

We would like to express our gratitude to the Alexander von Humboldt Foundation for funding the research stay at the Cognitive Systems Lab (CSL) of the University of Bremen.

6. References

[1] Ethnologue, "Languages of the world," Retrieved October 21, 2019, from https://www.ethnologue.com/, 2019.

[2] T. Schultz, "Towards rapid language portability of speech processing systems," in *Conference on Speech and Language Systems for Human Communication (SPLASH)*, vol. 1, Delhi, India, November 2004.

[3] T. Schultz and K. Kirchhoff, *Multilingual Speech Processing*. Elsevier Academic Press, 2006.

[4] T. Schultz and A. Waibel, "Language-independent and language-adaptive acoustic modeling for speech recognition," *Speech Commun.*, vol. 35, no. 1-2, pp. 31–51, Aug. 2001. [Online]. Available: http://dx.doi.org/10.1016/S0167-6393(00)00094-7

[5] F. Weng, H. Bratt, L. Neumeyer, and A. Stolcke, "A study of multilingual speech recognition," in *EUROSPEECH*, 1997.

[6] T. Schultz and A. Waibel, "Multilingual and crosslingual speech recognition," in *Proc. DARPA Workshop on Broadcast News Transcription and Understanding*, 1998, pp. 259–262.

[7] T. Schultz, "Globalphone: a multilingual speech and text database developed at karlsruhe university." in *INTERSPEECH*, J. H. L. Hansen and B. L. Pellom, Eds. ISCA, 2002.

[8] S. Kanthak and H. Ney, "Multilingual acoustic modeling using graphemes," in *IN PROCEEDINGS OF EUROPEAN CONFERENCE ON SPEECH COMMUNICATION AND TECHNOLOGY*, 2003, pp. 1145–1148.

[9] J. Cui, B. Kingsbury, B. Ramabhadran, A. Sethy, K. Audhkhasi, X. Cui, E. Kislal, L. Mangu, M. Nußbaum-Thom, M. Picheny, Z. Tüske, P. Golik, R. Schlüter, H. Ney, M. J. F. Gales, K. M. Knill, A. Ragni, H. Wang, and P. C. Woodland, "Multilingual representations for low resource speech recognition and keyword search," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2015, Scottsdale, AZ, USA, December 13-17, 2015*, 2015, pp. 259–266.

[10] Z. Tüske, D. Nolden, R. Schlüter, and H. Ney, "Multilingual MRASTA features for low-resource keyword search and speech recognition systems," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4-9, 2014*, 2014, pp. 7854–7858.

[11] Z. Tüske, P. Golik, D. Nolden, R. Schlüter, and H. Ney, "Data augmentation, feature combination, and multilingual neural networks to improve ASR and KWS performance for low-resource languages," in *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014*, H. Li, H. M. Meng, B. Ma, E. Chng, and L. Xie, Eds., 2014, pp. 1420–1424.

[12] A. Datta, B. Ramabhadran, J. Emond, A. Kannan, and B. Roark, "Language-agnostic multilingual modeling," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 8239–8243.

[13] O. Adams, M. Wiesner, S. Watanabe, and D. Yarowsky, "Massively multilingual adversarial speech recognition," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 96–108.

[14] N. T. Vu, D. Imseng, D. Povey, P. Motlícek, T. Schultz, and H. Bourlard, "Multilingual deep neural network based acoustic modeling for rapid language adaptation," *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7639–7643, 2014.

[15] M. Müller and A. H. Waibel, "Using language adaptive deep neural networks for improved multilingual speech recognition," 2015.

[16] E. Chuangsuwanich, "Multilingual techniques for low resource automatic speech recognition," Ph.D. dissertation, 2016.

[17] L. Besacier, E. Gauthier, M. Mangeot, P. Bretier, P. C. Bagshaw, O. Rosec, T. Moudenc, F. Pellegrino, S. Voisin, E. Marsico, and P. Nocera, "Speech technologies for african languages: example of a multilingual calculator for education," in *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, 2015, pp. 1886–1887.

[18] S. T. Abate, M. Y. Tachbelie, and T. Schultz, "Deep neural networks based automatic speech recognition for four ethiopian languages," in *ICASSP 2020*, 2020.

[19] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, "Automatic speech recognition for under-resourced languages: A survey," *Speech Communication*, vol. 56, pp. 85–100, 2014. [Online]. Available: https://doi.org/10.1016/j.specom.2013.07.008

[20] T. Schultz, N. T. Vu, and T. Schlippe, "Globalphone: A multilingual text and speech database in 20 languages," in *ICASSP*, 2013.

[21] ELRA, "European language resources association elra," ELRA catalogue. Retrieved November 30, 2012, from http://catalog.elra.info, 2012.

[22] Appen Buttler Hill Pty Ltd, "Speech and language resources 2012," Appen Butler Hill Speech and Language Resources 2012 - Product Catalogue, 2012.

[23] LM-BM, "Benchmark globalphone language models," Retrieved October 21, 2019, from https://www.csl.uni-bremen.de/GlobalPhone/, 2012.

[24] S. T. Abate, W. Menzel, and B. Tafila, "An amharic speech corpus for large vocabulary continuous speech recognition," in *INTERSPEECH*, 2005.

[25] S. T. Abate, M. Y. Tachbelie, M. Melese, H. Abera, T. Abebe, W. Mulugeta, Y. Assabie, M. Meshesha, S. Atinafu, and B. Ephrem, "Large vocabulary read speech corpora for four ethiopian languages: Amharic, tigrigna, oromo and wolaytta," in *LREC2020*, 2020.

[26] 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*. IEEE Signal Processing Society, 2011.

[27] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *INTERSPEECH*, 2015.

[28] IPA, *The principles of the International Phonetic Association*, 2nd ed. London, UK: University College of London, 1982.

[29] A. Stolcke, "Srilm – an extensible language modeling toolkit," in *IN PROCEEDINGS OF THE 7TH INTERNATIONAL CONFERENCE ON SPOKEN LANGUAGE PROCESSING (ICSLP 2002*, 2002, pp. 901–904.

[30] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," in *34th Annual Meeting of the Association for Computational Linguistics*. Santa Cruz, California, USA: Association for Computational Linguistics, 1996, pp. 310–318. [Online]. Available: https://www.aclweb.org/anthology/P96-1041

[31] M. Y. Tachbelie, S. T. Abate, and T. Schultz, "Analysis of globalphone and ethiopian languages speech corpora for multilingual asr," in *LREC 2020*, 2020.