Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 4152–4156
Marseille, 11–16 May 2020

Analysis of GlobalPhone and Ethiopian Languages Speech Corpora for Multilingual ASR

Martha Yifiru Tachbelie[1,2], Solomon Teferra Abate[1,2], Tanja Schultz[1]
[1]Cognitive Systems Lab (CSL), 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 analysis of GlobalPhone (GP) and speech corpora of Ethiopian languages (Amharic, Tigrigna, Oromo and Wolaytta). The aim of the analysis is to select speech data from GP for the development of multilingual Automatic Speech Recognition (ASR) system for the Ethiopian languages. To this end, phonetic overlaps among GP and Ethiopian languages have been analyzed. The result of our analysis shows that there is much phonetic overlap among Ethiopian languages although they are from three different language families. From GP, Turkish, Uyghur and Croatian are found to have much overlap with the Ethiopian languages. On the other hand, Korean has less phonetic overlap with the rest of the languages. Moreover, morphological complexity of the GP and Ethiopian languages, reflected by type to token ration (TTR) and out of vocabulary (OOV) rate, has been analyzed. Both metrics indicated the morphological complexity of the languages. Korean and Amharic have been identified as extremely morphologically complex compared to the other languages. Tigrigna, Russian, Turkish, Polish, etc. are also among the morphologically complex languages.

Keywords: Language relatedness, Multilingual ASR, GlobalPhone, Ethiopian Languages

1. Introduction

With more than 7000 languages in the world (Ethnologue, 2019) 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 (Schultz, 2004; Schultz and Kirchhoff, 2006). Major bottlenecks are the sparseness of speech and text data with corresponding pronunciation dictionaries, 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, Deep Neural Networks (DNN) for acoustic modeling and 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. On the other hand, large-scale data resources for research are available for less than 100 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.

In cross-lingual ASR, resources of a language (source/donor language) are used to develop ASR system for another (target language) with or without little adaptation data from the target language. Multilingual ASR system is described as a system that is able to recognize multiple languages which are presented during training (Schultz and Waibel, 2001). (Vu et al., 2014) described multilingual ASR as system in which at least one of the components (feature extraction, acoustic model, pronunciation dictionary, or language model) is developed using data from many different languages. Although multilingual 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 (Schultz and Waibel, 2001). Furthermore, they provide an appealing solution for multilingual, multi-Ethnic, and economically disadvantaged countries, such as Ethiopia.

Ethiopia is a multilingual and multi-ethnic country where over 80 languages are spoken by the citizens. When it comes to language resources required for the development of speech and language processing tools, almost all Ethiopian languages are under-resourced. 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 and language processing tools. Accordingly, we are currently investigating the development of multilingual ASR system for Ethiopian languages. For this purpose, we will use GlobalPhone (Schultz et al., 2013), a multilingual database of high-quality read speech with corresponding transcriptions and pronunciation dictionaries in more than 20 languages. To use this resource, it is mandatory to identify which languages are closely related with Ethiopian languages and therefore will be useful in the development of multilingual ASR.

In this paper, we present the analysis of GlobalPhone (GP) and speech corpora of Ethiopian languages (Amharic, Tigrigna, Oromo and Wolaytta) which are recently developed (Abate et al., 2020). The aim is to select speech data from GP and related Ethiopian Languages for the development of multilingual ASR system for the Ethiopian languages. We have analysed the phonetic overlaps among GP and Ethiopian languages based on International Phonetic Association (IPA) sound representation. Moreover, morphological complexity of the GP and Ethiopian languages has been analyzed based on type to token ration (TTR) and Out of Vocabulary (OOV) rate calculated on the basis of training transcriptions.

The next section describes the available resources: GlobalPhone and the Ethiopian languages speech corpora. Section 3. provides details of the analysis we made on the available language resources. Conclusions and future directions are presented in Section 4..

2. Available Resources

2.1. GlobalPhone

GlobalPhone (GP) is a multilingual data corpus that comprises (1) audio/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 language models. 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 Language Models (GP-LM). GP-ST is distributed under a research or commercial license by two authorized distributors, the European Language Resources Association (ELRA) (ELRA, 2012) and Appen Butler Hill Pty Ltd. (Ltd, 2012). GP-Dict is distributed by ELRA, while the GP-LMs are freely available for download from our website (LM-BM, 2015).

The entire GP corpus provides a multilingual database of word-level transcribed high-quality 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 pronunciation dictionaries). Thus, GP supplies an excellent basis for research in the areas of (1) multilingual ASR, (2) rapid deployment of speech processing systems to yet unsupported languages, (3) language identification tasks, (4) speaker recognition in multiple languages, (5) multilingual speech synthesis, as well as (6) monolingual ASR.

The GP corpus covers 20 languages, i.e. Arabic (modern standard), Bulgarian, Chinese (Mandarin and Shanghai), Croatian, Czech, French, German, Hausa, Japanese, Korean, Polish, Portuguese, Russian, Spanish, Swedish, Tamil, Thai, Turkish, Ukrainian, and Vietnamese. It comprises wide-spread languages (e.g. Arabic, Chinese, Spanish, Russian), contains economically and politically important languages, and spans wide geographical areas.

In addition to these languages, we have also considered Uyghur for which we have a speech corpus and want to use it in multilingual setting. The Uyghur corpus is a read speech corpus of selected newspaper articles. It contains 12 hours of training speech collected from 41 native speakers with 4k sentences and 1.5 hours of evaluation speech collected from 5 native speakers with 491 utterances.

2.2. Speech Corpora of Ethiopian Languages

Read speech corpora of four Ethiopian Languages (Amharic, Tigrigna, Oromo and Wolaytta) are considered in the analysis. One of the Amharic speech corpora, referred as AM2005, is prepared at the University of Hamburg (Abate et al., 2005). This contains 20 hours of training speech collected from 100 speakers who read a total of 11k sentences, development and test sets read by 20 other speakers (10 each). The other Amharic speech corpus, referred as AM2020, prepared at Addis Ababa University (AAU) together with the preparation of speech corpora of the other three languages.

The corpora of the Ethiopian languages have been collected in Ethiopia under a thematic research funded by AAU (Abate et al., 2020). The Amharic, Tigrigna and Oromo speech corpora consist of speech of 98 readers each. Most of the speakers of the languages, read 121 to 130 sentences. The size of the training speech of these three languages is 26 hours for Amharic and 22 hours for Tigrigna and Oromo. The Wolaytta corpus consists of recordings of 85 speakers where most of them read 140-150 sentences. Considering the difficulty of getting 100 readers for Wolaytta and aiming at collecting not less than 20hrs of speech, 150 utterances were assigned to each speaker. This way it became possible to collect a speech corpus of 29 hours.

3. Analysis Of Globalphone and Ethiopian Languages Corpora

3.1. Language Family

The languages considered in our analysis fall into 10 language families. Austro-Asiatic: Hausa and Vietnamese; Cushitic: Oromo; Indo-European: that includes Germanic (English, German and Swedish), Romance (French, Portuguese and Spanish) and Slavic (Bulgarian, Croatian, Czech, Polish and Russian); Japonic: Japanese; Koreanic: Korean; Kra-Dai: Thai; Omotic: Wolaytta; Semitic: Amharic, Arabic and Tigrigna; Sino-Tibetan: Mandarin; Turkic: Turkish and Uyghur.

3.2. Writing System

The written language contains all types of writing systems, i.e. logographic scripts (Chinese Hanzi and Japanese Kanji), phonographic segmental scripts (Roman, Cyrillic), phonographic consonantal scripts (Arabic), phonographic syllabic scripts (Japanese Kana, Thai), Abugida/Ethiopic (Amharic, Tigrigna), linear nonfeatural (Uyghur) and phonographic featural scripts (Korean Hangul).

3.3. Sound System

Phonetic information is important in multilingual ASR. Considering this fact, we have analysed the sound system of the GP as well as Ethiopian languages. In the analysis, a broad selection of phonetic characteristics have been considered, e.g. tonal sounds (Mandarin, Thai, Vietnamese, Oromo, Wolaytta), consonantal clusters (German), nasals (French, Portuguese), plosive sounds (Amharic, Oromo, Tigrigna, Wolaytta), uvular (Uyghur) and palatized sounds (Amharic, Oromo, Tigrigna, Wolaytta, Russian).

We have analysed and identified language independent phones (polyphones), phones occuring in more than one languages, and language dependent (monophones), phones that occur in only one language (Andersen et al., 1993) as it is done in (Schultz and Kirchhoff, 2006). The phone analysis is done on the basis of the pronounciation dictionaries we have at hand for each of the languages. Table 1 indicates the polyphones that occur in two or more languages.

No. of Lang.	No. of Phone	Consonants	Vowels
All	3	m, n, s	-
22	3	k, l, t	-
21	5	b, d, f, p	u
20	1	j	-
19	3	g	i, o
17	2	v, z	-
16	3	h,ʃ	a
15	2	r	e
13	2	w	ɛ
12	4	x,ɲ, ʧ, ʒ	-
11	1	ŋ	-
9	5	ʤ	aː, iː, ə ɨ
8	1	-	ɔ
7	5	ʦ, ʔ	oː, uː, ʌ
6	2	ʧ	eː
5	6	ʎ, ʁ, t^h	y, ai, ø
4	13	c,ç,k^h,kˈ, p^h, pˈ, ɾ, ʂ tˈ, ʧˈ, zˌ	ɪ, uə
3	18	ɗ, ð, ɣ, sˈ, s^j, t^j ʤ	ɛː, œ, ɯ, au, aʊ, eɪ, ja, ju, oʊ, ɔɪ, ua
2	39	b^j, d^j, ɟ, ħ, ʕ, ɦ, m^j, p^j, ɋ, ɕ, ʈ, ʈʂ, ʧʰ, θ, ʤ, v^j, z^j, ʐ	aˌ, ɐ, æ, é, əː, ɛ̃, iˌ, í, oˌ, ɔ̃, øː, uˌ, ʊ, yː ɯː, iə, je, jo, ui, ui, ue

Table 1: Polyphones shared by 2 or more languages

In addition to identification of polyphone and monophone, we analysed the sound overlap among the GlobalPhone and the four Ethiopian languages based on the sound representation of the International Phonetic Association (IPA). That means we considered sounds from different languages similar if they are represented by the same IPA symbol. Otherwise, they are considered as different sounds. Figure 1 indicates the coverage of sounds of a language (values being 0% to 100%), for example Amharic, in the rest of the languages. The dark blue color indicates 100% overlap whereas light yellow indicates low or no overlap.

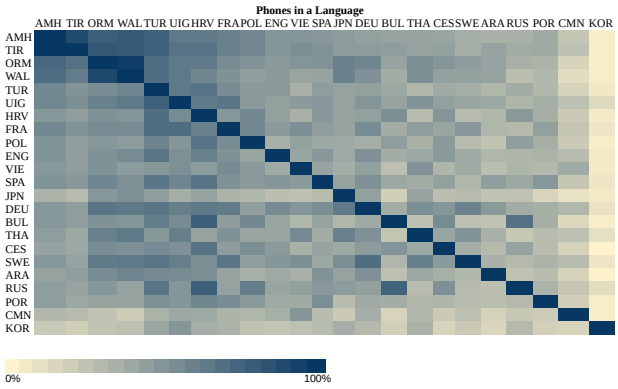

Figure 1: The phonetic sound overlap.

As can be seen from the Figure 1, there is much phone overlap among the Ethiopian languages. Interestingly, all Amharic phones are covered by Tigrigna. We expected high phonetic overlap among the Semitic languages (Amharic, Tigrigna and Arabic), however Arabic has low phone overlap with the rest of the Ethiopian Semitic languages. From GP, Turkish, Uyghur and Croatian cover most phones of the Ethiopian languages. The Figure also shows that there is high phone overlap among three of the Slavic languages: Croatian, Bulgarian and Russian. Although, Polish and Czech also fall under Slavic language family, our analysis does not show high phone overlap of these languages with the rest of the Slavic languages. On the other hand, Korean and Mandarin seem to have less phone overlap with the rest of the GP as well as Ethiopian languages.

3.4. Morphological Property

The morphological complexity of a language affects the quality of a language model and the coverage of decoding vocabulary (pronunciation dictionary). Since language model and pronunciation dictionary are components of an ASR system that affect performance, we have analysed the morphological complexity of the GP and four Ethiopian languages based on the training transcription and the transcription of the evaluation set.

The languages considered in our analysis cover many morphological variations, e.g. agglutinative languages (Turkish, Korean), languages with compounding morphology (German), and non-concatenative root-pattern morphology (Amharic, Tigrigna and Arabic), and also include scripts that completely lack word segmentation (Chinese, Thai).

We have computed type to token ratio (TTR), calculated as vocabulary size divided by text length, based on the training transcriptions. (Kettunen, 2014) showed that TTR can order the languages quite meaningfully in a morphological complexity order or at least groups most of the languages with same kind of morphological complexity and clearly separates the most and least morphologically complex languages. Moreover, (Bentz et al., 2016) showed that TTR is highly correlated with other corpus-based methods (such as word entropy, relative entropy of word structure and word alignment based measure). Since TTR is affected by the length of the text sample, moving average TTR (MATTR) is used (Kettunen, 2014; Covington and McFall, 2010) for coherent texts. However, texts used in our analysis are random sentences selected from different sources, mainly from newspapers. Thus, we computed average TTR (ATTR) for the training transcription based on k disjunct 1000-utterance subsets of the training set instead of MATTR. In each iteration we randomly selected 1000 utterances, computed TTR, removed these utterances from the pool, and then continued with the next iteration. The ATTR is then computed using TTR values of each distinct 1000 utterances. Figure 2 and 3 show the TTR and ATTR for each of the languages considered in our analysis.

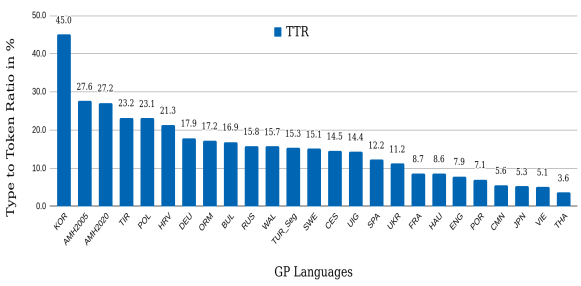

Figure 2: Type to Token Ratio.

As shown in both Figures, TTR and ATTR reflect the morphological complexity of the languages. The languages are arranged from morphologically more complex to less com-

plex. Korean is indicated as the most morphological complex language, followed by Amharic (for both Amharic corpora) and Tigrigna. On the other hand, English, Hausa, Japanese, Mandarin and Thai are identified as less morphological complex languages. However, we used segmented at word (multi-character or multi-syllable) transcription for Japanese, Mandarin and Thai and therefore, the Figure may not reflect the true morphological property of these languages. But, in general, the ATTR seems to reflect the morphological complexity of the languages.

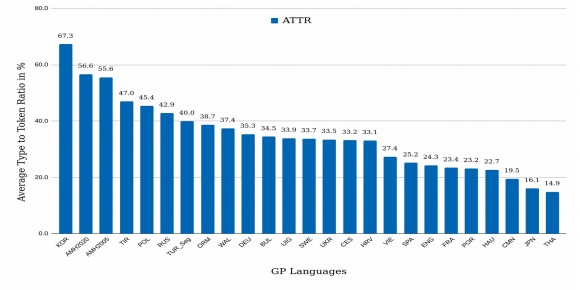

Figure 3: Average Type to Token Ratio.

In the speech and language processing community, Out of Vocabulary (OOV) rate is commonly used as an indication of morphological complexity (Cotterell et al., 2019). In ASR, one OOV word acounts to one or one and half wrongly recognized word/s. Mostly high OOV means high word recognition error rate. We have calculated OOV rates of different vocabulary sizes extracted from the training transcriptions available for each language. Figure 4 shows the OOV rates of different sizes of vocabularies against the complete vocabulary of the training transcription. That means, for example, we take the most frequent 1k word types from the training vocabulary (all unique words of the training transcription) and compute OOV of this list against the training vocabulary.

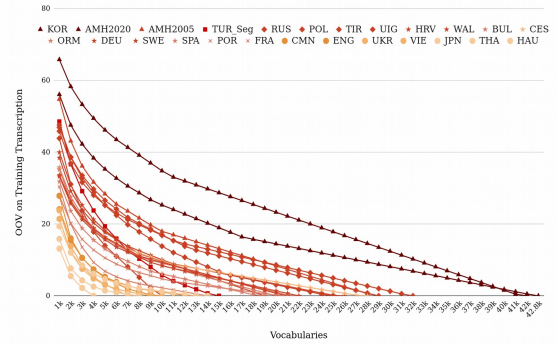

Figure 4: OOV of Different Vocab. Sizes of the Training Set.

As can be seen from Figure 4, Korean and Amharic are characterized by extremely high OOV rate, compared to the other languages considered in our analysis. Moreover, Turkish, Russian, Tigrigna, Polish, Uyghur, Croatian, Wolaytta, Bulgarian, German, Swedish are also characterized by high OOV rates. On the other hand, Mandarin, Thai, Hausa, Japanese, Vietnamese and English are characterized by low OOV rate.

We have also calculated the OOV rate of the evaluation set of each of the languages against the different vocabulary sizes taken from the training transcriptions. Figure 5 shows the OOV on evaluation set. The pattern shown in Figure 5 is similar with that of Figure 4. Our analysis of the morphological complexity of the GP and Ethiopian languages, both using TTR and OOV, helps to know which languages are more challenging with respect to the two components of the ASR system: vocabulary and language models. As a solution to the morphological complexity problem, morphemes (instead of words) have been used as units in these models. The other alternative, i.e. the use of large vocabularies and language models are limited by the (very) small amounts of data available for under-resource languages. Depending on the availability of resources, we will study the impact of these approaches in multilingual as well as monolingual ASR of morphologically complex languages.

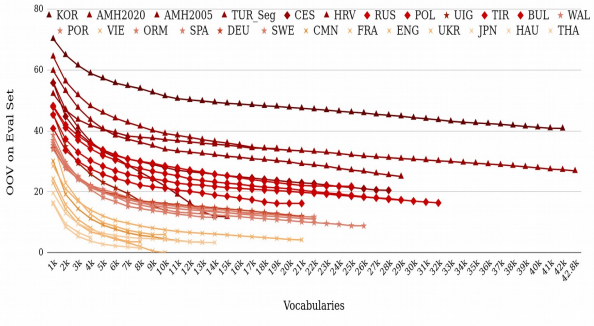

Figure 5: OOV of the Evaluation Set.

4. Conclusion and Future Direction

In this paper we have presented analysis of GlobalPhone and four Ethiopian languages speech corpora that we intend to use for the development of multilingual ASR system for Ethiopian languages. The purpose of the analysis is to select viable speech corpora from the available resources (GP and Ethiopian languages) for multilingual ASR system development. The phonetic analysis shows that there is high phone overlap among Ethiopian languages although the languages are from three different language families. Moreover, Turkish, Uyghur and Croatian are found to have slightly high phone overlap with the Ethiopian languages. Our analysis also shows that, the Ethiopian languages', considered in the analysis, morphology is not simple. Amharic has extremely high morphological complexity, next to Korean, which has low phone overlap with the rest of the languages.

Our next step is conducting multilingual ASR experiments using GP and the Ethiopian languages corpora. In addition, in our analysis we have cosidered two sounds as similar if they are represented with the same IPA symbol. One future work will be using data-driven methods (Le et al., 2006) for identifing phonetic similarity and investigate whether it leads to diferrent results.

5. Acknowledgment

Thanks to the Alexander von Humboldt Foundation for funding the research stay at CSL, University of Bremen.

6. Bibliographical References

Abate, S. T., Menzel, W., and Tafila, B. (2005). An amharic speech corpus for large vocabulary continuous speech recognition. In *INTERSPEECH*.

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

Andersen, O., Dalsgaard, P., and Barry, W. J. (1993). Data-driven identification of poly- and mono-phonemes for four european languages. In *EUROSPEECH*.

Bentz, C., Ruzsics, T., Koplenig, A., and Samardžić, T. (2016). A comparison between morphological complexity measures: Typological data vs. language corpora. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 142–153.

Cotterell, R., Kirov, C., Hulden, M., and Eisner, J. (2019). On the complexity and typology of inflectional morphological systems. *Transactions of the Association for Computational Linguistics*, 7:327–342.

Covington, M. A. and McFall, J. D. (2010). Cutting the gordian knot: The moving-average typeâtoken ratio (mattr). *Journal of Quantitative Linguistics*, 17(2):94–100.

ELRA. (2012). European language resources association elra. ELRA catalogue. Retrieved November 30, 2012, from http://catalog.elra.info.

Ethnologue. (2019). Languages of the world. Retrieved October 21, 2019, from https://www.ethnologue.com/.

Kettunen, K. (2014). Can type-token ratio be used to show morphological complexity of languages? *Journal of Quantitative Linguistics*, 21(3):223–245.

Le, V. B., L., B., and Schultz, T. (2006). Acoustic-phonetic unit similarities for context dependent acoustic model portability. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 1.

LM-BM. (2015). Benchmark globalphone language models. Retrieved October 21, 2019, from https://www.csl.uni-bremen.de/GlobalPhone/.

Ltd, A. B. H. P. (2012). Speech and language resources 2012. Appen Butler Hill Speech and Language Resources 2012 - Product Catalogue.

Schultz, T. and Kirchhoff, K. (2006). *Multilingual Speech Processing*. Elsevier Academic Press.

Schultz, T. and Waibel, A. (2001). Language-independent and language-adaptive acoustic modeling for speech recognition. *Speech Commun.*, 35(1-2):31–51, August.

Schultz, T., Vu, N. T., and Schlippe, T. (2013). Globalphone: A multilingual text and speech database in 20 languages. In *ICASSP*.

Schultz, T. (2004). Towards rapid language portability of speech processing systems. In *Conference on Speech and Language Systems for Human Communication (SPLASH)*, volume 1, Delhi, India, November.

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