DNN-BASED SPEECH RECOGNITION FOR GLOBALPHONE LANGUAGES

Martha Yifiru Tachbelie, Ayimunishagu Abulimiti, Solomon Teferra Abate*, Tanja Schultz**

Cognitive Systems Lab,University of Bremen, Germany
marthayifiru,abulimit,abate,tanja.schultz@uni-bremen.de

ABSTRACT

This paper describes new reference benchmark results based on hybrid Hidden Markov Model and Deep Neural Networks (HMM-DNN) for the GlobalPhone (GP) multilingual text and speech database. GP is a multilingual database of high-quality read speech with corresponding transcriptions and pronunciation dictionaries in more than 20 languages. Moreover, we provide new results for five additional languages, namely, Amharic, Oromo, Tigrigna, Wolaytta, and Uyghur. Across the 22 languages considered, the hybrid HMM-DNN models outperform the HMM-GMM based models regardless of the size of the training speech used. Overall, we achieved relative improvements that range from 7.14% to 59.43%.

Index Terms— GlobalPhone, DNN, Ethiopian Languages

1. INTRODUCTION

With more than 7000 languages in the world [1] and the strong demand 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]. 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 [3]. Data sparseness is a critical issue due to the fact that speech technologies heavily rely on statistically based 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. Apart from parameter estimation, the development of speech processing systems requires language expertise and established conventions, as described below. Large-scale data resources for research are available for about 100 languages and the costs for these collections are prohibitive to all but the most widely spoken and economically viable languages. The lack of language conventions concerns a surprisingly large number of languages and dialects. The lack of a standardized writing system (near to 50% of the wolrd's language [4]) for example hinders web harvesting of large text corpora and the construction of pronunciation dictionaries. Last but not least, despite the well-defined process of system building, it is cost- and time consuming to handle language-specific peculiarities, and requires substantial language expertise. Unfortunately, it is extremely difficult to find system developers who have both, the necessary technical background and the native expertise of a language in question. As a result, one of the pivotal issues for developing speech processing systems in multiple languages is the challenge of bridging the gap between language and technology expertise [3].

In 2002, we released the multilingual text and speech corpus GP to address the lack of databases which are consistent across languages [5]. By that time the database consisted of 15 languages but since then has been extended to cover more languages, more speakers, more word tokens along with their pronunciations, and more text resources. In addition, GP was adopted as a benchmark database for research and development of multilingual speech processing systems and is available from European Language Resources Association (ELRA) [6]. The latest status of GP and GMM based reference benchmark Automatic Speech Recognition (ASR) system performances of 20 languages was provided for researchers and developers working with this database in 2013 [7]. However, the paper [7] does not reflect current state-of-the-art performances based on recent developments in DNN. The current paper intends to fill this gap, i.e. provide new reference benchmarks for GP based on hybrid HMM-DNN. Furthermore, a collection of Ethiopian data is described, which is very similar to GP in terms of speaking style (read), number of speakers (about 100 speakers per language), and size of speech (about 20 hours per language).

1.1. Artificial Neural Networks

Although the history of Artificial Neural Networks (ANNs) goes back to 1943 [8] where the mathematical model of neurons was introduced, they did not outperform traditional HMM-GMM models for many decades due to the lack of processing power, data resources and efficient algorithms. The reasons behind the poor performance of ANNs include

*The authors would like to thank the Alexander von Humboldt Foundation for research fellowship.

the problem of vanishing gradient, lack of high performance computing, and weak temporal correlation structure. Over the years, the problems have been solved through the availability of high performance computing (such as GPU) and the introduction of different types of neural network architectures: Recurrent Neural Networks (RNNs), Convolutional Neural networks (CNN), Long Short Term Memory (LSTM), Bidirectinal LSTM, and more recently Time Delay Neural Networks (TDNN) and Factored TDNN (TDNNf).

Since 2009, ANNs are widely used in ASR and presented dramatic improvement in performance. Numerous studies showed that hybrid HMM-DNN systems outperform the dominant Gaussian Mixture Model (GMM) on the same data [9]. Currently, TDNNs, also called one-dimensional Convolutional Neural Networks, are an efficient and well-performing neural network architectures for ASR [10]. TDNN has the ability to learn long term temporal contexts. Moreover, by using singular value decomposition (SVD) the number of parameters in TDNN models is reduced which makes them inexpensive compared to RNNs. The factored form of TDNNs (TDNNf)[11] has similar structure with TDNN, but is trained from a random start with one of the two factors of each matrix constrained to be semi-orthogonal. TDNNf gives substantial improvement over TDNN and has been shown to be effective in under-resourced scenarios. We have used these state-of-the-art neural network architecture in the development of DNN based ASR systems for 22 languages.

2. THE CORPUS

GlobalPhone is a multilingual data corpus developed in collaboration with the Karlsruhe Institute of Technology (KIT). The complete data corpus 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 latter as GP Language Models (GP-LM). GP-ST is distributed under a research or commercial license by two authorized distributors, ELRA [6] and Appen Butler Hill Pty Ltd. [12]. GP-Dict is distributed by ELRA, while the GP-LMs are freely available for download from our website [13].

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 Amharic corpus used in this experiment is a read speech corpus prepared at the University of Hamburg [14]. It contains 20 hours of training speech collected from 100 speakers who read a total of 11k sentences (29k types), development and test sets read by 20 other speakers. The corpus has 5k and 20k vocabularies development as well as evaluations sets. In this experiment, we have merged the development sets and evaluation sets so as to evaluate the ASR systems with relatively bigger (in size) development and evaluation sets read by 10 speakers for each. Speech corpora of the other three Ethiopian languages have been collected in Ethiopia under a thematic research funded by the Addis Ababa University [15]. The corpora are read speech corpora prepared in the same way as the Amharic corpus, except the use of smart phone for the recording purpose. However, the recording quality is 16kHz. The Uyghur corpus is a read speech corpus of selected newspaper articles collected within the NSF-funded project [16]. It contains 12 hours of training speech collected from 41 native speakers with 4k sentences (63k tokens) and 1.5 hours of evaluation speech collected from 5 native speakers with 491 utterances (10k tokens).

2.1. Language Coverage

The GP corpus covers 20 languages, i.e. Arabic, 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, contains economically and politically important languages, and spans wide geographical areas. Out of the 20, 17 languages are considered in our experiment. In addition to these, four additional African and one Asian languages have been considered. This selection covers a broad variety of language peculiarities relevant for Speech and Language research and development.

In our experiment, 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 palatalized sounds (Amharic, Oromo, Tigrigna, Wolaytta, Russian). 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). The languages cover many morphological variations, e.g. agglutinative languages (Turkish, Korean), compounding languages (German), and non-concatinative root-pattern mor-

phology (Amharic and Tigrigna) also include scripts that completely lack word segmentation (Chinese, Thai).

2.2. Data Acquisition

The data acquisition was performed in countries where the language is officially spoken. In each language about 100 adult native speakers were asked to read about 100 sentences. The first batch of GP data collection was done from May 1996 to November 1997, and a second batch between 2003 and 2012. During the first batch we collected Arabic speech in Tunis, Sfax and Djerba, Tunisia; Mandarin in Beijing, Wuhan and Hekou, China; Shanghai in Shanghai, China; Croatian in Zagreb, Croatia, and parts of Bosnia; Czech in Prague, Czech Republic; French in Grenoble, France; German in Karlsruhe, Germany; Japanese in Tokyo, Japan; Korean in Seoul, Korea; Portuguese in Porto Velho and Sao Paulo, Brazil; Polish in Poland, Russian in Minsk, Belarus; Spanish in Heredia and San Jose, Costa Rica; Swedish in Stockholm and Vaernamo, Sweden; Tamil in India, and Turkish in Istanbul, Turkey. In the second batch we collected Bulgarian in Sofia, Hausa in Cameroon, Thai in Bangkok, Ukrainian in Donezk, and Vietnamese in Hanoi and Ho Chi Minh City. The Amharic, Tigrigna and Oromo are collected from Addis Ababa while the Wolaytta corpus is collected in Wolaytta Sodo, Ethiopia.

The read texts were selected from national newspaper articles available from the web to cover a wide domain with large vocabulary. The articles report national and international political news, as well as economic news, which makes it possible to compare the usage of proper names (Politicians, companies, etc.) across languages. The text for Amharic was also selected from news archives of different medias. Different sources including news articles, bible, books, etc. are used as sources of text for Tigrigna, Oromo and wolaytta texts. For Uyghur the text was also taken from newspaper articles.

The GP, Uyghur and Amharic speech data were recorded with a close-speaking microphone and is available in identical characteristics for all languages: PCM encoding, mono quality, 16bit quantization, and 16kHz sampling rate. Most recordings were done in ordinary rooms or offices, in the majority without background noise. On the other hand, the Tigrigna, Oromo and Wolaytta speech recording was done using smartphones and in different environments and therefore the speech is not free from background noises. However, the speech characteristics is similar to the GP languages.

2.3. Corpus Statistics

The entire GP corpus contains over 400 hours of speech spoken by more than 2000 native adult speakers. The data are organized by languages and speakers and are divided into speaker disjoint sets for training (80%), development (10%), and evaluation (10%). For details of the amount of training, development and evaluation sets speech of the GP languages, we direct the reader to previous publication on GP [7]. Table 1 summarizes the amount of transcribed speech data for the five additional languages.

Table 1. *Statistics for five additional languages*

Language	Training [hrs:min]	Development [hrs:min]	Evaluation [hrs:min]
Amharic	20:00	1:30	1:33
Oromo	22:48	1:11	1:04
Tigrigna	22:06	1:03	1:02
Wolaytta	29:42	1.32	1.43
Uyghur	12:24	-	1:55

Table 2. *Decoding Pronunciation Dictionaries(PD)*

Languages	#Phones	#PD Vocab	OOV	LMToken	PPL
Amharic	40	310k	3.06	4M	41.2
Bulgarian	44	275k	1.07	405M	341.62
Croatian	32	23k	2.09	331M	934.75
Czech	41	277k	4.04	508M	1223.5
French	38	122k	6.028	220M	356.87
German	43	39k	0.059	20M	675.86
Hausa	33	43k	0.32	15M	76.63
Japanese	31	58k	0.18	1600M	89.41
Mandarin	49	73k	0	900M	268.06
Oromo	59	21k	11.73	1.2M	266.17
Portuguese	45	59k	1.09	11M	45.8
Polish	36	49k	0.1	224M	880.83
Russian	47	40k	2.09	334M	1070.74
Spanish	42	43k	4.65	12M	113.44
Swedish	48	25k	0	211M	325.91
Thai	44	23k	0.22	15M	16.64
Tigrigna	44	299k	4.9	4M	211.41
Turkish	31	34k	1.25	7M	55.04
Ukrainian	49	40k	0.0002	94M	105.76
Uyghur	37	40k	13.9	250k	260.59
Vietnamese	59	39k	3.17	39M	1227.01
Wolaytta	57	25k	9.34	226k	254.9

3. PRONUNCIATION DICTIONARIES

Phone-based pronunciation dictionaries are available for each GP language. The dictionaries 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. After this automatic creation process the dictionaries were manually post-processed word-by-word by native speakers, correcting errors in the automatic pronunciation generation and introducing pronunciation variants. To enable the development of multilingual speech processing, the phone names are consistent across languages, leveraging the International Phonetic Alphabet (IPA) [17]. The pronunciation dictionary of Uyghur is prepared in similar fashion. For Amharic and Tigrigna, the pronunciation dictionaries are prepared automatically taking the Consonant-Vowel syllabary feature of the writing system, Abugida. Almost all characters in the writing system translate into a consonant and vowel phones. For Oromo and Wolaytta, pronunciation dictionaries are prepared automatically considering the pronunciation rules of the language. Table 2 gives an overview of the size of the phone sets, the out-of-vocabulary (OOV) with regard to the evaluation set, and amount of vocabulary words cov-

ered in the pronunciation dictionaries. Further details on the GlobalPhone dictionaries are given in [18].

4. LANGUAGE MODELS

In this experiment, we used the GP language models described in [7]. For the additional five languages, we developed trigram language models using SRILM [19] and different sizes of text corpus obtained from the web. However, as there are no text resources on the web for Wolaytta and we could not get any additional text corpus, only the training transcription has been used to train a trigram language model. Table 2 shows the perplexity (PPL) of the language models on the evaluation set and the amount of training data used to train the language models.

5. SPEECH RECOGNITION SYSTEMS

In this section, we present the large vocabulary ASR systems trained and evaluated on GP languages and the five additional languages. For training, development, and evaluation, we used the audio data as described in Section 2.3, the dictionaries and trigram language models shown in Table 2. All recognition systems were build in the same fashion using Kaldi ASR toolkit [20]. First we built context dependent HMM-GMM based acoustic model for each language using 39 dimensional mel-frequency cepstral coeffcinents (MFCCs) to each of which Cepstral Mean and Variance Normalization (CMVN) is applied. The acoustic model uses a fully-continuous 3-state left-to-right HMM. Linear Discriminant Analysis (LDA) and Maximum Likelihood Linear Transform (MLLT) feature transformation are performed for each of the models. Finaly, Speaker Adaptive Training (SAT) has been done using an affine transform, feature space Maximum Likelihood Linear Regression (fMLLR). The best model of each language, which is mostly the fMMLR, is used to obtain alignments for DNN training. Results are reported in word error rate (WER) for the majority of the languages, except Vietnamese, Mandarin and Thai for which syllable and character error rates are used, respectively. Figure 1 shows the different WER obtained on each of the languages.

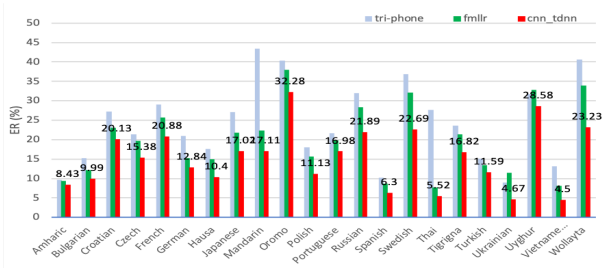

Fig. 1. Word/Syllable/Character Error rates

In DNN acoustic modeling, the same speech data used to train HMM-GMM model for each language has been used. However, three-fold data augmentation [21] was applied prior to the extraction of 40-dimensional MFCCs without derivatives, 3-dimensional pitch features and 100-dimensional i-vectors for speaker adaptation. The neural network architecture used to develop the acoustic models is Factored Time Delay Neural Networks with additional Convolutional layers (CNN-TDNNf) according to the standard Kaldi WSJ recipe [22]. 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 the TDNNf layer immediately following the CNN layers which has 256 bottleneck units. The default hyperparameters (15 hidden layers, initial and final learning rate of 0.0005 and 0.00005, and minibatch size of 128 and 64) of the standard recipe were used except for the number of epochs for which we used 7 instead of the defualt 8. As can be seen from Figure 1, Error rate (ER) reductions have been obtained for all languages regardless of the amount of training speech used. The numbers on the bars are error rates (word/syllable/character) of the hybrid HMM-DNN system.

Table 3 shows the relative error rates (word/syllable /character) reductions for all languages. As can be seen from Table 3, we have obtained relative error rate reductions that range from 7.14% to 59.43%. The highest error rate reduction has been obtained for Ukrainian followed by Vietnamese. The smallest error rate reduction is obtained for Hausa for which only 6.36 hours of training speech has been used to train DNN which has 15 hidden layers. The differences in relative error rate reductions observed among the languages require further analysis and investigation in the future.

Table 3. *Size of Training Speech (hours:min) and Relative Improvement(%)*

Languages	Size	Rel. imp.	Languages	Size	Rel. imp.
Amharic	20:00	10.03	Portuguese	18:06	14.50
Bulgarian	16:48	17.91	Russian	21:00	22.62
Croatian	11:48	12.48	Spanish	17:30	27.59
Czech	26:42	22.01	Swedish	17:42	29.47
French	21:54	18.72	Thai	11:36	28.50
German	14:54	16.02	Tigrigna	22:06	21.07
Hausa	6:36	7.14	Turkish	13:12	13.76
Japanese	30:46	21.82	Ukrainian	10:42	59.43
Mandarin	26:42	23.24	Uyghur	12:24	12.92
Oromo	22:48	15.10	Vietnamese	20:48	44.44
Polish	19:18	29.02	Wollayta	29:42	31.45

6. SUMMARY

In this paper we presented reference benchmark ASR system performances based on hybrid HMM-DNN for the GP languages and five additional languages (Amharic, Oromo, Tigrigna, Wolaytta and Uyghur). Our findings show that error rate reduction can be obtained using DNN even for systems developed with small amount of speech data (only 6 hours of speech) and using the same neural network architecture for all the languages regardless of the training speech size. Overall, we achieved relative error rate improvements between 7.14% and 59.43%, averaging to 22.69% across 22 different languages.

7. REFERENCES

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

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

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

[4] Omniglot, "The online encyclopedia of writing systems and languages," Retrieved February 12 2020 https://www.omniglot.com/writing/stats.htm, 2020.

[5] Tanja Schultz, "Globalphone: A multilingual speech and text database developed at karlsruhe university," in *Proceedings of the ICSLP*, 2002, pp. 345–348.

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

[7] Tanja Schultz, Ngoc Thang Vu, and Tim Schlippe, "Globalphone: A multilingual text and speech database in 20 languages," in *ICASSP*, 2013.

[8] Warren S McCulloch and Walter Pitts, "A logical calculus of the ideas immanent in nervous activity," *The bulletin of mathematical biophysics*, vol. 5, no. 4, pp. 115–133, 1943.

[9] Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Brian Kingsbury, et al., "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal processing magazine*, vol. 29, 2012.

[10] Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[11] Daniel Povey, Gaofeng Cheng, Yiming Wang, Ke Li, Hainan Xu, Mahsa Yarmohammadi, and Sanjeev Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks.," in *Interspeech*, 2018, pp. 3743–3747.

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

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

[14] Solomon Teferra Abate, Wolfgang Menzel, and Bahiru Tafila, "An amharic speech corpus for large vocabulary continuous speech recognition," in *INTERSPEECH*, 2005.

[15] Solomon Teferra Abate, Martha Yifiru Tachbelie, Michael Melese, Hafte Abera, Tewodros Abebe, Wondwossen Mulugeta, Yaregal Assabie, Million Meshesha, Solomon Atinafu, and Binyam Ephrem, "Large vocabulary read speech corpora for four ethiopian languages: Amharic, tigrigna, oromo and wolaytta," in *LREC2020*, 2020.

[16] NSF-Funded Project, "Eager: Automatic speech recognition for uyghur, award number 1519164, 2015-2016," 2016.

[17] IPA, *The principles of the International Phonetic Association*, University College of London, London, UK, 2 edition, 1982.

[18] Tanja Schultz and Tim Schlippe, "Globalphone: Pronunciation dictionaries in 20 languages," in *LREC*, 2014.

[19] A. Stolcke, "Srilm - an extensible language modeling toolkit," in *Intl. Conf. Spoken Language Processing (ICSLP)*, 2002.

[20] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Kumar Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely, "The kaldi speech recognition toolkit," in *ASRU*, 2011.

[21] Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "Audio augmentation for speech recognition," in *INTERSPEECH*, 2015.

[22] Kaldi, "Kaldi download website," Retrieved Febraury 13 2020 https://github.com/kaldi-asr/kaldi/tree/master/egs/wsj/s5/local/chain, 2020.